Detection of Parkinson's Disease from handwriting kinematics feature extraction using ML classification algorithms

Hari Priya Vuppu
Computer Science
Georgia State University
Atlanta, GA, United States
hvuppu1@student.gsu.edu

Advisor: Dr. Wei Li
Committee: Dr. Yanqing Zhang

Abstract - **Parkinson's disease is a neurodegenerative disorder that leads to neuro-motor deficits. This paper discusses the formulation of a detection method used to diagnose Parkinson's disease using pictures of spiral tests done on patients from which a set of features are extracted, the resultant data of which will be run through classifier models to compare the accuracy using each model and thereby using the model with the highest accuracy.**

Keywords – Parkinson's Disease, Tremors, Spiral test, Classification Algorithms,

I. INTRODUCTION

When brain cells that produce dopamine, a neurotransmitter that regulates movement, quit producing it or die, Parkinson's disease (PD) develops. PD is referred to be a movement disorder since it can produce tremors, slowness, stiffness, and issues with walking and balance. Since PD affects handwriting, the handwriting style can be utilized as a diagnostic tool for PD. This develops slowly and is difficult to identify in its early stages due to delayed symptoms. A neurologist will often evaluate the patient's medical history and examine the patient's computed tomography scans and magnetic resonance images, or a body movement analyst would examine the patient's motions to identify the disease.

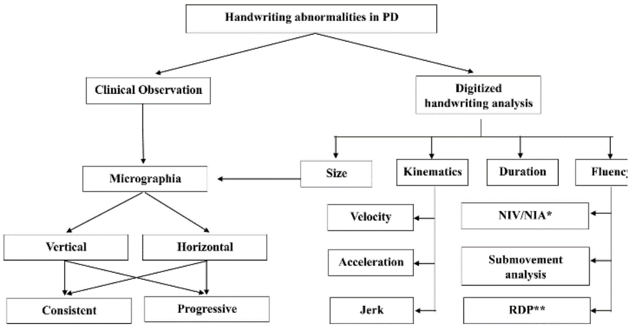
A. Handwriting

We will be analyzing one of these motions of the patient – handwriting. Since PD produces tremors, it impacts hand movements as well. This results in text that is not legible. Besides this, we also take advantage of the fact that movement while penning a text includes both on-surface movements of the hand and in-air trajectories done when the hand moves in the air between each stroke.

In order to analyze the handwriting of PD patients and get insight into the motor disruption brought on by PD, a number of handwriting activities have been developed. The Archimedean spiral is now the most well-liked handwriting activity for tremor testing. Spiral sketching is commonly used to assess motor function in a variety of movement disorders, including Parkinson's disease (PD). It has been demonstrated that both the absolute location of the pen during writing and pen movements above the writing surface is significant for the diagnosis of PD (when the pen does not leave the trajectory). Additionally, the amount of surface pressure applied while writing by hand is quite important.

B. Database

For this project, the ‘Parkinson’s Disease Spiral Drawings Using Digitized Graphics Tablet Data Set’ database from UCI Machine Learning Repository was made use of. This consists of 77 patient records where 62 people have PD and 15 are healthy individuals. To collect this dataset, a tablet connected to a computer is placed in front of the individual performing the task, and the task is performed with a stylus on the tablet.



The pressure from the pen acting orthogonally to the digitizing tablet surface can also be captured in addition to the x/y coordinates of the pen position when writing. Micrographia may serve as a presymptomatic neurobehavioral biomarker of Parkinson's disease, according to studies. "A noticeable reduction in the size of the writer's letters relative to the calligraphy before the onset of the biological lesion generating the change" is the definition of micrographia.

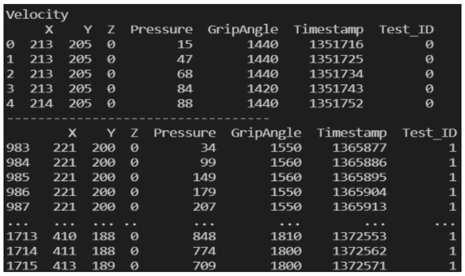
C. Feature Extraction

To get to the conclusion that a person has Parkinson’s disease using this method of analysis, a few factors impacting the writing of an individual will have to be taken into consideration. These factors/features that are extracted from the process of writing are:

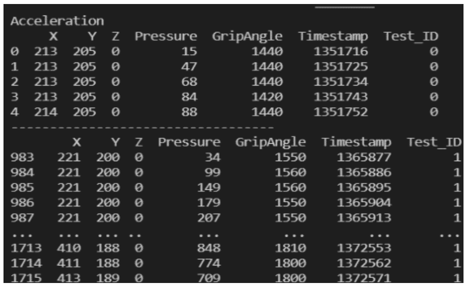
1. No. of strokes
2. Stroke speed - Stroke length divided by stroke duration in mm/s
3. Velocity - Rate at which the position of a pen changes with time in mm/s
4. Acceleration - Rate at which the velocity of a pen changes with time in mm/s^2
5. Jerk - Rate at which the acceleration of a pen changes with time in mm/s^3
6. Horizontal velocity/acceleration/jerk
7. Vertical velocity/acceleration/jerk
8. Number of changes in velocity direction - The mean number of local extrema of velocity.
9. Number of changes in acceleration direction - The mean number of local extrema of acceleration.
10. In-airtime – The time gap between one letter and the next.
11. On-surface time - Time spent on-surface during the writing

The value of pressure recorded by the tablet during the specific task and the pace at which pressure varies in regard to time are the two basic characteristics of pressure.

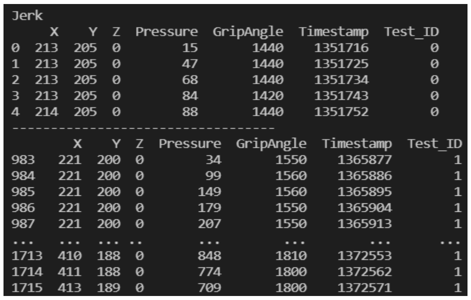
Given are the vectors of the given features which will further be used for the process of detection. This is obtained by running the feature extraction python file. Here we have velocity, acceleration, Jerk, and NCA.



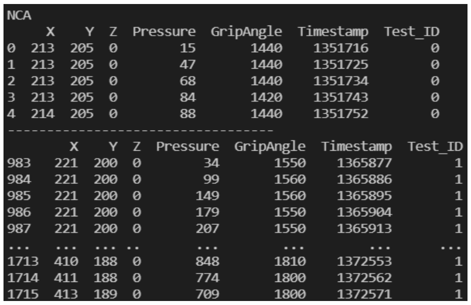
```
Velocity
      X    Y  Z  Pressure  GripAngle  Timestamp  Test_ID
0  213  205  0        15       1440    1351716        0
1  213  205  0        47       1440    1351725        0
2  213  205  0        68       1440    1351734        0
3  213  205  0        84       1420    1351743        0
4  214  205  0        88       1440    1351752        0
----------------------------------
         X    Y  Z  Pressure  GripAngle  Timestamp  Test_ID
983    221  200  0        34       1550    1365877        1
984    221  200  0        99       1560    1365886        1
985    221  200  0       149       1560    1365895        1
986    221  200  0       179       1550    1365904        1
987    221  200  0       207       1550    1365913        1
...    ...  ... ..       ...        ...        ...      ...
1713   410  188  0       848       1810    1372553        1
1714   411  188  0       774       1800    1372562        1
1715   413  189  0       709       1800    1372571        1
```



```
Acceleration
      X    Y  Z  Pressure  GripAngle  Timestamp  Test_ID
0  213  205  0        15       1440    1351716        0
1  213  205  0        47       1440    1351725        0
2  213  205  0        68       1440    1351734        0
3  213  205  0        84       1420    1351743        0
4  214  205  0        88       1440    1351752        0
----------------------------------
         X    Y  Z  Pressure  GripAngle  Timestamp  Test_ID
983    221  200  0        34       1550    1365877        1
984    221  200  0        99       1560    1365886        1
985    221  200  0       149       1560    1365895        1
986    221  200  0       179       1550    1365904        1
987    221  200  0       207       1550    1365913        1
...    ...  ... ..       ...        ...        ...      ...
1713   410  188  0       848       1810    1372553        1
1714   411  188  0       774       1800    1372562        1
1715   413  189  0       709       1800    1372571        1
```

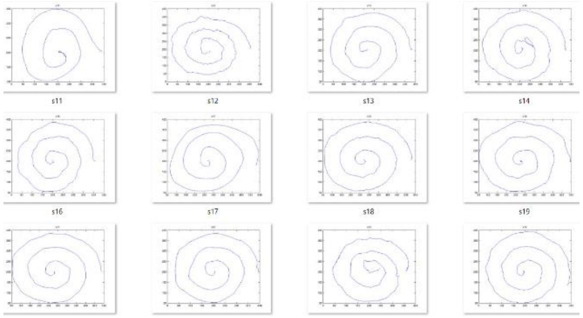


```
Jerk
      X    Y  Z  Pressure  GripAngle  Timestamp  Test_ID
0  213  205  0        15       1440    1351716        0
1  213  205  0        47       1440    1351725        0
2  213  205  0        68       1440    1351734        0
3  213  205  0        84       1420    1351743        0
4  214  205  0        88       1440    1351752        0
----------------------------------
         X    Y  Z  Pressure  GripAngle  Timestamp  Test_ID
983    221  200  0        34       1550    1365877        1
984    221  200  0        99       1560    1365886        1
985    221  200  0       149       1560    1365895        1
986    221  200  0       179       1550    1365904        1
987    221  200  0       207       1550    1365913        1
...    ...  ... ..       ...        ...        ...      ...
1713   410  188  0       848       1810    1372553        1
1714   411  188  0       774       1800    1372562        1
1715   413  189  0       709       1800    1372571        1
```

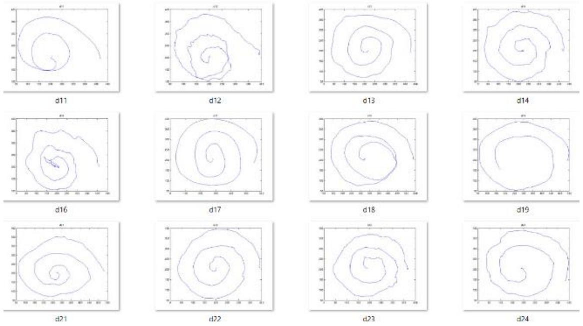


```
NCA
      X    Y  Z  Pressure  GripAngle  Timestamp  Test_ID
0  213  205  0        15       1440    1351716        0
1  213  205  0        47       1440    1351725        0
2  213  205  0        68       1440    1351734        0
3  213  205  0        84       1420    1351743        0
4  214  205  0        88       1440    1351752        0
----------------------------------
         X    Y  Z  Pressure  GripAngle  Timestamp  Test_ID
983    221  200  0        34       1550    1365877        1
984    221  200  0        99       1560    1365886        1
985    221  200  0       149       1560    1365895        1
986    221  200  0       179       1550    1365904        1
987    221  200  0       207       1550    1365913        1
...    ...  ... ..       ...        ...        ...      ...
1713   410  188  0       848       1810    1372553        1
1714   411  188  0       774       1800    1372562        1
1715   413  189  0       709       1800    1372571        1
```

These are the coordinates obtained from the writing exercises of the patients. They were asked to take two types of tests. The first one is the Static Spiral Test (SST) which involves drawing an Archimedean spiral of three coiled turns which will appear on the tablet screen and the patients will be asked to retrace the same. The second one is the Dynamic Spiral Test (DST) in which the patient will be asked to retrace the same spiral. Only this time, the spiral flashes in order to identify the drawing abilities of the patient. The x and y coordinates for both spiral tests are represented in a single image to get the most out of the data recovered from the graphic tablet used. The sample of the type of drawings obtained are given below:

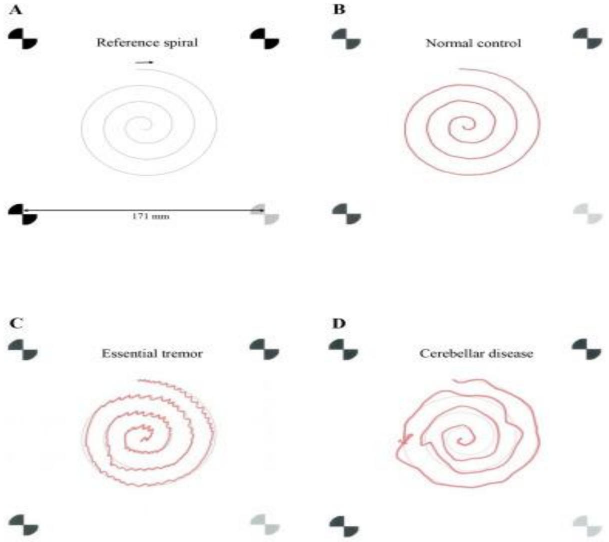


Static Spiral Test drawings



Dynamic Spiral Test drawings

The shape of the spiral is the primary distinction between spirals drawn by PD patients and non-PD patients for both SST and DST.



The number of distorted movements while drawing the spiral, the pressure applied while drawing, the amount of time gap between two placements of the pen, and factors like these will essentially determine the condition of the patient.

II. DETECTION

A. PYTHON LIBRARIES

The libraries used for the code of detection of PD are:

NumPy: A general-purpose array processing package is NumPy. In addition to tools for working with these arrays, it offers a high-performance multidimensional array object. It serves as Python's foundational package for scientific computing.

It has several characteristics, significant ones among them being:

An efficient N-dimensional array object

advanced (broadcasting) features

Software for combining C/C++ and Fortran code

Fourier transform, random number, and useful linear algebra features

Seaborn: Python's Seaborn visualization package allows for the plotting of statistical visuals. It offers lovely default styles and color schemes to enhance the appeal of statistics charts. It is constructed on top of the Matplotlib toolkit and is tightly integrated with the Pandas data structures.

With Seaborn, visualization will be at the heart of data exploration and comprehension. For a better comprehension of the dataset, it offers dataset-oriented APIs that allow us to switch between various visual representations for the same variables.

Plots are mostly used to show how different variables relate to one another. These variables may be entirely numerical or may represent a category, such as a group, class, or division. Seaborn categorizes the plot into the following groups:

Relational plots: This type of graphic is used to see how two variables are related.

Categorical plots: This graphic discusses categorical variables and their visualization of them.

Plots used to examine univariate and bivariate distributions include distribution plots.

Regression plots: The main purpose of the regression plots in Seaborn is to provide a visual aid that highlights patterns in a dataset during exploratory data analysis.

Plots in a matrix an array of scatterplots makes up a matrix plot.

Multi-plot grids: Drawing many instances of the same thing is helpful.

Pandas: Pandas is a Python library for data analysis. Pandas is built on top of two essential Python libraries: NumPy for mathematical operations and Matplotlib for data presentation. By serving as a wrapper for these libraries, Pandas makes it easier to access many of matplotlib's and NumPy's methods.

Some of its features are:

- Data from different file objects can be loaded.
- Easy handling of missing data (represented as NaN) in floating point as well as non-floating-point data.
- Size mutability: columns can be inserted and deleted from DataFrame and higher dimensional objects
- Data set merging and joining.
- Flexible reshaping and pivoting of data sets
- Provides time-series functionality.
- Powerful group by functionality for performing split-apply-combine operations on data sets.

os: Python's OS module offers tools for communicating with the operating system. OS is included in the basic utility modules for Python. A portable method of exploiting operating system-specific functionality is offered by this module. There are numerous functions to deal with the file system in the *os* and *os.path* modules.

math: The Python 3 standard library includes a built-in module called math that offers common mathematical constants and functions. Calculations in the areas of numerics, trigonometry, logarithms, and exponentials can all be done using the math module.

time: The Python time module offers a variety of coding representations for time, including objects, integers, and strings. Along with functions other than time representation, it also allows you to measure the effectiveness of your code and wait while it executes.

pathlib: The pathlib module in Python is used to demonstrate how to work with files and directories.

A Python package called pathlib offers an object API for dealing with files and directories.

B. RESULTS

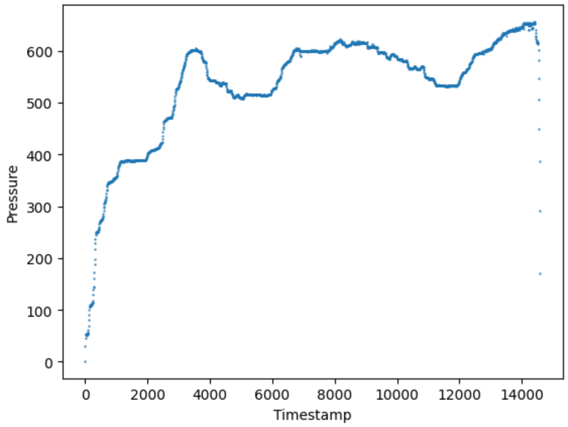
Using the obtained features during the handwriting, we find the velocity, acceleration, jerk, and thereafter NCV and NCA. This is what the feature extraction code consists of. After this, we will be using the sklearn library to run the classifier algorithm with these features.

sklearn: The Sklearn Library offers effective, user-friendly tools for any type of predictive data analysis and is primarily used for data modeling.

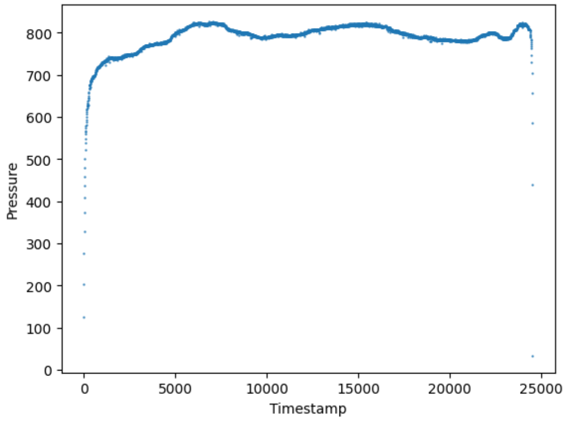
- Preprocessing
- Regression
- Classification
- Clustering
- Model Selection
- Dimensionality Reduction

Firstly, logistic regression is performed on the test set to check the accuracy. This gives us an accuracy of 70 %.

We also obtain the sample attribute difference between the control members and members having Parkinson's disease. This is represented in the form of graphs. This is based on the pressure applied on the tablet surface.



Parkinson's members



Control members

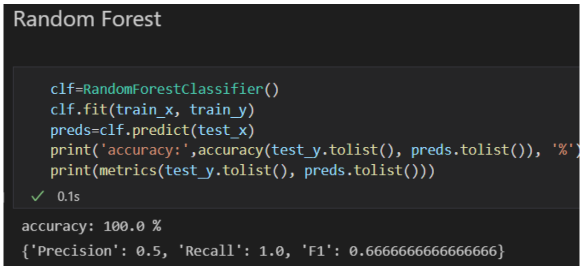
Now, we run three different classifiers for this data we have. The three classifiers used here are:

Support Vector Machine - A supervised machine learning approach called Support Vector Machine (SVM) is used for both classification and regression. Although we also refer to regression concerns, categorization is the most appropriate term. Finding a hyperplane in an N-dimensional space that clearly classifies the data points is the goal of the SVM method. The number of features determines the hyperplane's size.

Random Forest Classifier - A random forest is a meta-estimator that employs averaging to increase predicted accuracy and reduce overfitting after fitting numerous decision tree classifiers to distinct dataset subsamples.

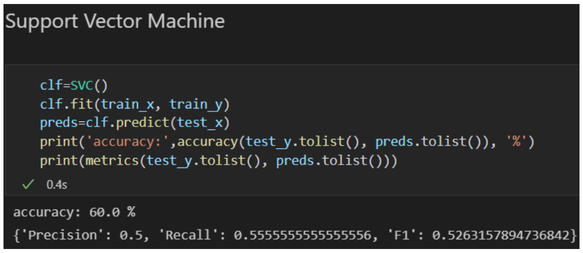
K-Nearest Neighbor - The k-nearest neighbors (KNN) technique calculates the likelihood that a data point will belong to one group, or another based on which group the data points closest to it do.

The results received upon running the three classification algorithms is something as follows:



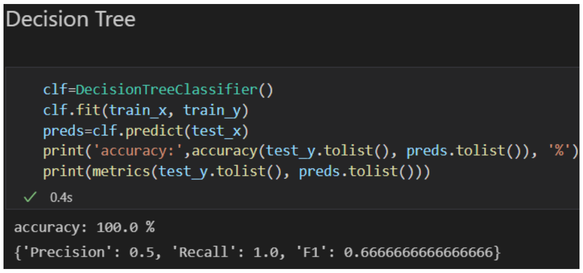
Random Forest

```
    clf=RandomForestClassifier()
    clf.fit(train_x, train_y)
    preds=clf.predict(test_x)
    print('accuracy:',accuracy(test_y.tolist(), preds.tolist()), '%'
    print(metrics(test_y.tolist(), preds.tolist()))
  ✓ 0.1s
accuracy: 100.0 %
{'Precision': 0.5, 'Recall': 1.0, 'F1': 0.6666666666666666}
```



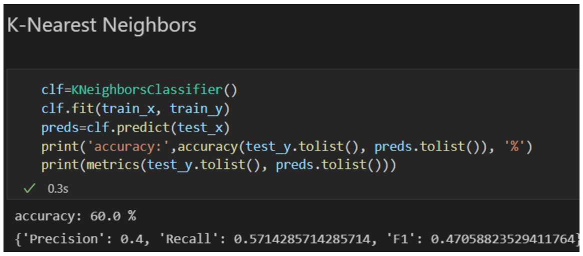
Support Vector Machine

```
    clf=SVC()
    clf.fit(train_x, train_y)
    preds=clf.predict(test_x)
    print('accuracy:',accuracy(test_y.tolist(), preds.tolist()), '%')
    print(metrics(test_y.tolist(), preds.tolist()))
  ✓ 0.4s
accuracy: 60.0 %
{'Precision': 0.5, 'Recall': 0.5555555555555556, 'F1': 0.5263157894736842}
```



Decision Tree

```
    clf=DecisionTreeClassifier()
    clf.fit(train_x, train_y)
    preds=clf.predict(test_x)
    print('accuracy:',accuracy(test_y.tolist(), preds.tolist()), '%')
    print(metrics(test_y.tolist(), preds.tolist()))
  ✓ 0.4s
accuracy: 100.0 %
{'Precision': 0.5, 'Recall': 1.0, 'F1': 0.6666666666666666}
```

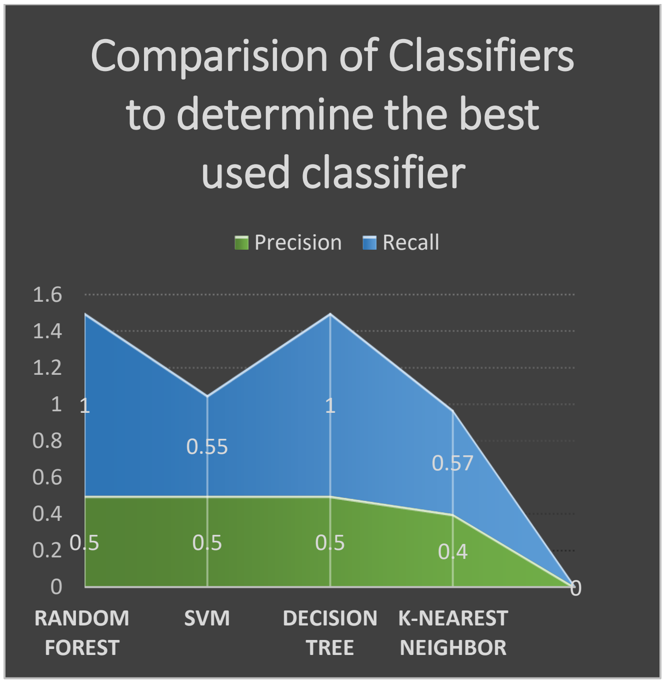


K-Nearest Neighbors

```
    clf=KNeighborsClassifier()
    clf.fit(train_x, train_y)
    preds=clf.predict(test_x)
    print('accuracy:',accuracy(test_y.tolist(), preds.tolist()), '%')
    print(metrics(test_y.tolist(), preds.tolist()))
  ✓ 0.3s
accuracy: 60.0 %
{'Precision': 0.4, 'Recall': 0.5714285714285714, 'F1': 0.47058823529411764}
```

When we verify using the visual representation of the comparison of the results with the values we have, we can conclude that Random Forest and Decision Tree have highest Area under Curve (AUC).

The accuracy of these algorithms is 100%, 60%, 100%, and 60% respectively. But since Random Forest Classifier had higher sensitivity than the decision tree and does not overfit as it does for the decision tree, we choose Random Forest Classifier as our classification model.



III. CONCLUSION

The number of persons who have PD has significantly increased recently. It so ranks among the biggest health issues.

Since there is now no treatment for it, early discovery is crucial to enabling effective care. Additionally, it is critical to routinely track the development of the symptoms. However, this necessitates frequent doctor visits for the patient who must cope with travel, waiting, appointments, etc.

Technological methods like these would be very efficient and very helpful in terms of time and cost for a patient and thereby to society as it should be.

This can also be further developed by scanning text beside the spiral test. This algorithm can be integrated into mobile devices which might help in immediate diagnosis when required. Furthermore, sensors can be placed on the body of the potential patient or in the surrounding environment of the person as a daily usable device to first and foremost diagnose the disease and thereafter regularly monitor the progress of it which would help the doctors take early action.

IV. FUTURE WORKS

We intend to gather more information in order to look into the usage of deep learning in extracting the visual descriptors from SST and DST drawing and its capability to distinguish between PD and non-PD drawing pattern. The amount of the data would allow for efficient deep learning network training.

Examining the processing times reveals a negative impact of the feature selection approaches on the overall processing durations. It has been found that while the genetic algorithm's feature selection procedure takes a long time, the classification time is shortened.

The test times won't be impacted by processing time. Tests will only be conducted on the features we have identified when a new sample is taken into consideration. When it is taken into account that it cuts the classification time even more because the number of characteristics is decreased, it will have a favorable impact.

Using fusion techniques would be another option to improve the outcomes. They might be used with a collection of classifiers, a collection of visual descriptors, or a collection of other data sources. The latter could be accomplished by distinguishing between PD and non-PD patients using both the drawing pattern and the voice pattern.

REFERENCES:

[1] S. N. Heyn, C. P. Davis and M. C. Stöppler, "Parkinson's Disease Symptoms, Causes, Stages, Treatment, and Life Expectancy," MedicineNet, 28 August 2018. [Online]. Available: https://www.medicinenet.com/parkinsons_disease/article.htm.

[2] "Tremor in Parkinson's," American Parkinson Disease Association, [Online]. Available:

https://www.apdaparkinson.org/what-isparkinsons/symptoms/tremor/.

[3] Handwriting Analysis in Parkinson's Disease: https://movementdisorders.onlinelibrary.wiley.com/doi/10.1002/mdc3.12552

[4] Diagnosis of Parkinson's Disease Using Spiral Test Based on Pattern Recognition: https://www.researchgate.net/publication/359710965_Diagnosis_of_Parkinson's_Disease_Using_Spiral_Test_Based_on_Pattern_Recognition

[5] A systematic approach to diagnose Parkinson's disease through kinematic features extracted from handwritten drawings: https://www.researchgate.net/publication/348605648_A_systematic_approach_to_diagnose_Parkinson's_disease_through_kinematic_features_extracted_from_handwritten_drawings

[6] Parkinson's Disease Diagnosis using Spiral Test on Digital Tablets: https://thesai.org/Downloads/Volume11No5/Paper_60-Parkinsons_Disease_Diagnosis

[7] Handwriting as an objective tool for Parkinson's disease diagnosis:

https://link.springer.com/article/10.1007/s00415-013-6996-x

[8] P. Drotár, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, and M. Faundez-Zannuy, "Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease," Artificial intelligence in medicine, vol. 67, no. 1, pp. 39-46, 2016.

[9] K.-C. Lan and W.-Y. Shih, "Early Diagnosis of Parkinson's Disease using a Smartphone," in The 11th International Conference on Mobile Systems and Pervasive Computing (MobiSPC-2014), Tainan, 2014.

[10] Parkinson's News Today, "Parkinson's Disease Statistics," Parkinson's News Today: https://parkinsonsnewstoday.com/parkinsons-disease-statistics/