Predicting Parkinson's disease Progression with a Sustained Vowel

1st Marie-Philippe Gill
Département de Génie Logiciel et des TI
École de Technologie Supérieure
Montréal, Canada
marie-philippe.gill.1@etsmtl.net

2nd Félix Blier
Département de génie des systèmes
École de Technologie Supérieure
Montréal, Canada
felix.blier.1@etsmtl.net

***Abstract*—Although the Unified Parkinson's Disease Rating Scale (UPDRS) is the de-facto gold standard to evaluate the progression of Parkinson's disease (PD) in patients, it is constrained to just providing a snapshot for how the patient feels on a given day. In this work, we propose alternative methods to help clinicians and patients gauge this progression and make decisions. Our method relies on using 16 vocal (or speech) features recorded on at-home devices over six months and worn by 42 subjects to predict UPDRS scores. The subjects enrolled were newly diagnosed and not taking any medication. The objective is to understand and analyze the course of PD, through techniques such as observing the central tendency measures, box plots, etc. To study the feasibility of using speech features from sustained vowels, we use a Support Vector Regressor (SVR) and a multilayer perceptron to predict the motor and total UPDRS scores. Preliminary results show that this method of using sustained vowels can provide promising results to assess the severity of PD symptoms. We observe that the condition of 15 subjects improve over the time of the 6-month study, suggesting that the activities performed daily with the device served as therapy. Furthermore, we find that four different groups of patients naturally emerge on clustering the speech features after dimensionality reduction using linear discriminant analysis (LDA). Further analysis could be interesting to obtain insights about the relation between the different groups.**

***Index Terms*—Parkinson's Disease, Vocal features, SVR, Multilayer Perceptron**

I. Introduction

Parkinson's disease (PD) affects more than 10 million people around the world. It is the second most prevalent degenerative disease, after Alzheimer's. The disease's course is currently evaluated using the Unified Parkinson's Disease Rating Scale (UPDRS) and is considered the gold standard.

The UPDRS [4] test only provides a picture of the PD symptoms of 30 minutes every six months. It does not provide an evaluation of the symptoms on a day to day basis. Therefore, this work aims to analyze a dataset collected by the subject using an at-home device that could help clinicians have a better idea of the disease. It could also make better healthcare decisions tailored to what the subject needs and experience in their daily lives.

Section II presents some information about PD and explains how the work that gathered the dataset can solve a significant problem for subjects and clinicians. Section III present the main objectives of the present work and is being followed in section IV with the methods used. Then, the results are presented in section V and are then discussed in VI.

TABLE I
Common PD symptoms [5]

Most common symptoms	Other possible symptoms
Resting tremor	Fatigue
Slowness (bradykinesia) and stiffness	Soft Speech
Impaired balance	Problems with handwriting
Rigidity of the muscles	Stooped posture
	Constipation
	Sleep disturbance

II. Context

Movement is controlled by dopamine in the brain. PD symptoms start to appear when cells that usually produce dopamine in the brain dies. Subjects can experience many symptoms, and a non-exhaustive list is in table I.

Resting tremor is one of the most common symptoms that is often associated with PD. It is a rhythmic muscle contraction that results in shaking movements in one or more parts of the body, particularly on the hands. However, with activity, this symptom can be suppressed.

There is currently no cure for the disease. However, treatment is available. Levodopa is the most commonly prescribed medications. In the brain, it is converted to dopamine. However, there is a considerable side-effect: long-term use of the medicine can induce motor complications like dyskinesia.

Dyskinesia is an uncontrolled, involuntary movement, but to the contrary of tremor, it is not specifically around one joint. It can involve the entire body, and it is not suppressible. It occurs when the subject takes medication, so when the other PD symptoms are under control, one can suffer from dyskinesia.

As observed on subjects in a study completed by Hanson, Gerratt, and Ward [8], PD can affect the voice by a "bowing of the vocal folds", a "tremor of the supraglottic structures", a "tremulous movement of the arytenoids" and a "supraglottic constriction of the larynx".

PD significantly affects the quality of life of subjects. From swallowing, eating, getting out of bed, every aspect of daily life can be affected. A shocking 50% of subjects with the disease can suffer from depression and anxiety after their diagnosis. The disease progresses at a different pace for every subject, and when symptoms progress, the medication needs to be adjusted accordingly.

The disease's progression is currently evaluated with the UPDRS test, which was first proposed in 1987 [6] and was called version 3.0. The questionnaire consisted of 4 parts : (I) Mentation, Behavior, and Mood (II) Activities of daily living (III) Motor (IV) Complications. It was designed to capture multiple aspects of PD, from motor disability, impairment, mental health, and disease complications.

However, the UPDRS test is performed every six months on average. Considering that PD symptoms fluctuate every day and are highly influenced by factors such as sleep, caffeine, stress, and more, a 30-minute assessment every six months does not provide a good picture of the subject's symptoms on a day to day basis. Clinicians use the UPDRS to make treatment decisions for the subjects, therefore it would be life-changing for subjects to evaluate their symptoms at home for an extended period. This could help make better medical decisions to control the course of the disease.

The dataset was created by [1] with the aim to test the feasibility of using an At-Home Testing Device (AHTD) as the Objective Measures of Motor Impairment test. The dataset was later used and shared by [2], which demonstrated that noninvasive speech tests were able to replicate the UPDRS assessment with about 7.5 UPDRS points different from the clinicians' estimates.

The speech test consisted of subjects telling the vowel ”ah” until the sound's amplitude dropped down a threshold or for 30 seconds.

III. Objectives

The first objective of this work is to understand the domain of the dataset and is presented in section IV. Therefore, research and readings on Parkinson's disease are necessary, as well as to learn more about the speech features that are provided in the dataset, and how the data was collected.

- What is the UPDRS test?
- What are the 16 vocal features provided with the dataset?

The second objective is to perform some analysis of the data, and outcomes will be presented in the results, section V.

- Is there a correlation between the age of the subjects and their motor UPDRS score or total score?
- Is there a correlation between the sex of the subject and their motor UPDRS score or total score?
- Can we identify some clusters between voice features and UPDRS scores?

The third objective is to perform some predictions using different machine learning models and is outside of the scope of the current work. It will be presented in the next version of this paper.

- Does it makes sense to predict UPDRS score from a sustained vowel?
- Can we predict the motor UPDRS (or total) score from 16 vocal features?

IV. Methods

A. Data

In the original study, fifty-two subjects were enrolled [1]. They were all newly diagnosed with the disease and did not take any medication. Two subjects have dropped out of the study before the end of the six months, and two others had started taking medication. Eight others were discarded because they did not have enough recordings. There are 5,875 voice recordings in the six months. Each subject has executed the 4 minutes speech test weekly. Each session consists of 4 execution at an average intensity and two execution at a loud intensity (twice the normal without yelling). The voice is recording by the AHTD device using a high-quality microphone headset.

In the data table, each row corresponds to a test session. The subject is identified by qualitative attributes: a nominal subject id number, the gender in a binary attribute, and the age in nominal value. Subjects are composed of 28 men and 14 women. The other features are quantitative and continued values.

A column contains the number of days since recruitment when the test session occurs. (test_time).

The column motor_UPDRS contain the result of the third part of the UPDRS test. The column total_UPDRS present the UPDRS score of the four parts.

Jitter and shimmer parameters are explained by [3]. Jitter is the frequency variation, and the shimmer is the amplitude variation from cycle to cycle in sound waves. Jitter parameter is represented by the relative average difference in percentage (%), the absolute average difference in microseconds (Abs), the average disturbance (RAP), the ratio of disturbance within five periods (PPQ5) and the differences between cycles (DDP), divided by the average period. The shimmer parameter contains the absolute value in decibels (dB), the three-point amplitude perturbation quotient (APQ3), the five-point amplitude perturbation quotient (APQ5), the 11-point amplitude perturbation quotient and the difference between the amplitudes of consecutive periods (DDA).

Harmonic and noise correlation are represented by Harmonics-to-noise (NHR) ratio and the Noise-to-harmonics ratio (HNR). It measures the ”relative amount of additive noise in the voice signal” [7].

Speech signal processing methods used by the recording device are RPDE, DFA, and PPE. In order, they represent the ”ability of the vocal folds to sustain simple vibration,” ”the extent of turbulent noise in the speech signal,” and ”the impaired control of stable pitch during sustained phonation” [2].

B. Tools

We used Jupyter notebooks to analyze the data as it is an excellent resource to code in python and makes it very easy to visualize datasets with multiple graphs. In the notebook, we used many libraries. $Pandas$ allows to load the data and manipulate it, then $Matplotlib$ to build graphs. $statsmodels$

TABLE II
DATA SCALING TECHNIQUES USED IN THIS WORK FOR THE SVR AND THE MLP

	Input	Output
None	None	None
Unscaled Inputs	None	Standardization
Normalized Inputs	Normalization	Standardization
Standardized Inputs	Standardization	Standardization

and $pylab$ are necessary to build a Q-Q plot to see if the distribution followed a normal distribution. For modelization, $scikit-learn$ was used for the SVR and $tensorflow$ for the multilayer perceptron.

We chose all of those libraries because the machine learning community vastly adopts them.

C. Pre-Processing

We performed data reduction to efficiently analyze the impact on UPDRS scores that age and sex can have. This reduction consists of grouping data by age and sex value. Each group value is equal to the mean value of all data included in the group. It is used to analyze the UPDRS score by age and by sex.

Latent Semantic Analysis (LSA) is useful to "reduce the dimensionality of the information retrieval problem" [10]. We used this technique to reduce all recording results in two vectors of data. So, it is possible to represent the data distribution on a two dimensions graphic. The other data dimension reducing tool is Latent Dirichlet Allocation (LDA). LDA "is a Bayesian probabilistic model of text documents." [11] The data analyzed is in numerical values and not text. It is possible to manipulate the subject as a document. Also, determined by the University of North Texas study, after performing classification on documents, "accuracy rate were 67% for LSA and 64% for LDA" [12]. As the accuracy rate is better using LDA than LSA, in this work, we used LSA instead of LDA.

We also experimented with different data scaling techniques. If the measures of the features are from different scales, for example, in kilometers and centimeters, then it can affect the performance of the algorithm. The input variables (features) can be scaled, and the target variable, which we call the label.

We tried four different combinations of scaling techniques presented in table II. We used normalization and standardization. The first one means rescaling the data from the original range of values (the min and the max) to make the new values be between 0 and 1.

Standardization means rescaling the values' distribution so that the mean is 0, and the standard deviation becomes 1. It can also be done by subtracting the mean value of the features to center the distribution.

The main difference between these two scaling techniques is that normalization doesn't necessarily change the data's distribution as standardization does.

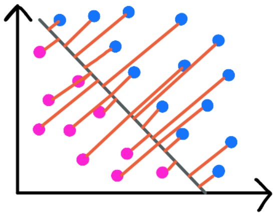


Fig. 1. SVR with a low gamma. The examples far from the decision boundary still have an influence on its position.

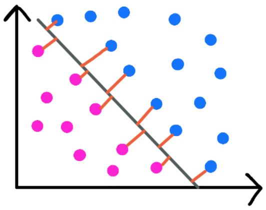


Fig. 2. SVR with a high gamma. Only the examples close to the decision boundary have an influence on it.

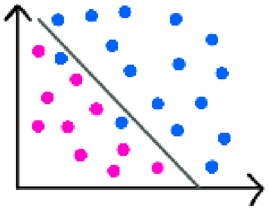


Fig. 3. SVR with a low C.

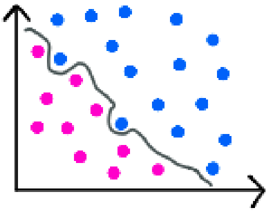


Fig. 4. SVR with a high C.

D. Algorithms used

Support Vector Machine (SVM) is a supervised machine learning algorithm. It can perform either classification or regression tasks. In this work, we used a Support Vector Regressor (SVR) to predict continuous UPDRS scores. The decision surface of an SVM is a hyperplane in an N-dimensional space. The N is equal to the number of features. Data points are used as support to the hyperplane, and the objective is to maximize the margin, which represents the distance between the separation of classes.

SVM can only separate data that is linear. In other words, data that you can separate the classes with a straight line. However, with a kernel trick, it is possible to transform non-linear data into a new dimension, making the new projection linearly separable.

To optimize the SVR, we tuned many parameters. We will focus on the Radial Basis Function (rbf) kernel as it is the one used in the experiments of this paper, for which the parameters are $gamma$ and C. $Gamma$ defines how far points can have an impact on the decision boundary, as explained in figure IV-D. If the value of gamma is too low, it could lead to examples close to the boundary being misclassified. If the value is too high, then the boundary could over-fit by fitting perfectly all of the training data.

The C parameter controls the shape of the decision boundary. It is a tradeoff between the shape of the separation in regards to classifying data points correctly.

$gamma$ and C controls different aspects of the boundary. $gamma$ is really about which points should be taken into consideration to influence the separation, while c controls the shape of the separation.

Kmeans is an iterative clustering algorithm that split data set in K clusters. This algorithm is used to group subjects that have a similar result in the dimensional reduction of data set. Kmeans parameters iterate over a number of time chosen in parameters. The initial position of centroids can be indicated or applied randomly. With the random initialization, the algorithm tries a number of possibilities written in parameter and keep the best inertia result. K-mean algorithm is easy to implement and produce a tighter cluster. The downside of this method is that the number of clusters is subjective, and the initialization position of centroids has a significant impact on the final result [15].

A multilayer perceptron is composed of simple or multiple sigmoid processing elements or neuron. These elements are structured in multiple layer that interact using weighted connection as a collective system. This architecture is based on the brain network neuron system. The advantage to "attempt to mimic the human brain" is that "many problems in pattern recognition are solved more easily by humans than by computer." [13] In this work, a multilayer perceptron is trained to predict the UPDRS score based on the dataset of vocal features recorded.

E. Evaluation metrics

The default score metric of the SVR implemented in $sklearn$ is the coefficient of determination R^2 [14], and it is defined as the proportion of the variance in the dependent variable (the labels) that is predictable from the independent variable (the features). It answers the question, "how much variation in the labels do the features explain?" or "how much does including the features reduce our error in predicting the labels?". Considering $y = [y_1, y_2, ..., y_n]$ is a vector of true labels associated with the dataset, and $\widehat{y} = [\widehat{y}_1, \widehat{y}_2, ..., \widehat{y}_n]$ are the predicted values from the model.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Where SS_{res} is the residual sum of squares with the predicted values, measuring how far the predictions are from the correct labels:

$$SS_{res} = \sum_i (y_i - \widehat{y}_i)^2$$

SS_{tot} is the total sum of squares with the mean of the observed data, and measures how far the data points are from the mean. $\overline{y}$ is the average of the observed data:

$$SS_{tot} = \sum_i (y_i - \overline{y}_i)^2$$

$$\overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

If a model always predicts the mean of the observed data ($\overline{y}$), then $R^2 = 0$, as there will be no variance. It means that the model is not better than a model only using the data mean, which would mean the model is not appropriate. Therefore, the best possible score is 1.0, which means the predicted values of $\widehat{y}$ will be precisely equal to the expected labels y.

TABLE III
CENTRAL TENDENCY MEASURES OF THE DATA

	Age	motor_UPDRS	total_UPDRS
Mean	64.80	21.29	29.01
Standard deviation	8.82	8.12	10.70
Minimum	36	5.03	7
25%	58	15	21.37
50%	65	20.87	27.57
75%	72	27.59	36.39
Max	85	39.51	54.99

Another metric that we used to evaluate the models is the mean squared error (MSE). It measures the average of the squared difference between the predictions and the labels.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

Therefore, if $MSE = 0$, it means there are no errors, and the model is perfectly fit to the data. This is not good behavior, because it shows that the model would be over-fitting and would most probably not generalize well when presented with new data. Similarly, if we observe $MSE = 100$, the model is underfitting and did not learn from the training data. There is no correct value for the MSE, but the lower, the better, and it is always a trade-off.

V. RESULTS

A. Data analysis

We started the analysis of the data by looking at some central tendency of the data, presented in table III.

As shown in the table III, the standard deviation of the age is low compared to the range of values recorded (between 36 and 85). It means that most subjects are aged near the mean (64.80 years old). It is different for the motor_UPDRS and total_UPDRS, which have a high standard deviation in a lower range of values. It means that the UPDRS scores are very scattered.

As we work on medical data with different subjects, the actual labels we are working with might be closely related to one subject and might not generalize well to different subjects. With that in mind, it seemed fitting to know how many recordings per subject are available.

As most subjects already have a decent number of recordings, we determined that it is unnecessary to discard any of them in a pre-processing step. On average, subjects had 139 recordings over the six months. The minimum number of recordings is 101, and the maximum is 168, so it does not vary too much.

The age histogram is presented in figure 5. The age of the subjects seem to be gathered in 3 groups: under 50 years old, between 55 and 68, and over 70.

Then, we want to know if the motor and total UPDRS scores were following a normal distribution. So we have plotted some probability graphs which are very similar to quantile-quantile

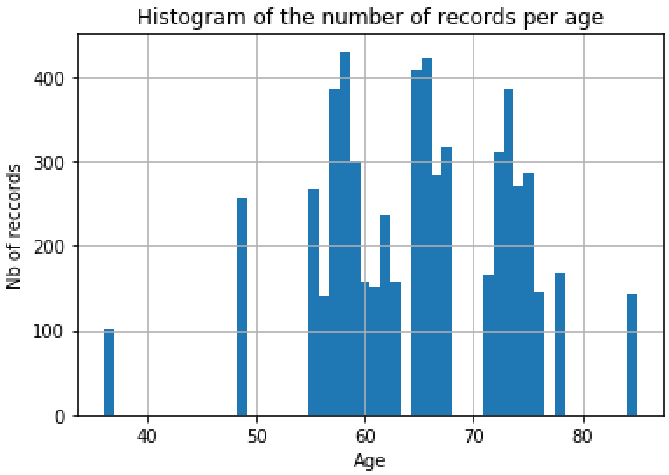


Fig. 5. Number of record by subject age

TABLE IV
UPDRS SCORES BY SEX

	Age	motor_UPDRS	total_UPDRS
Men	65.055	21.469	29.724
Women	64.267	20.924	27.505

plots (Q-Q plots) and are used to assess if a dataset follows a normal distribution. The data are plotted against a distribution that if they were following a normal distribution, then it would form a straight line.

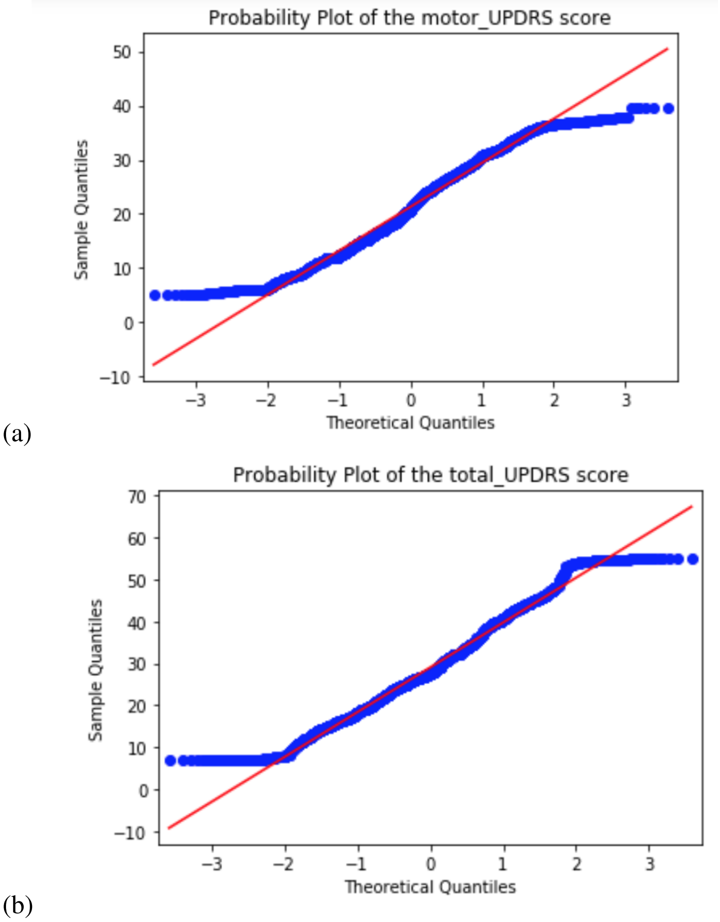
Both the motor and total UPDRS scores are following to a certain extent a normal distribution. On the other hand, we can see at the beginning and the end (figure 6), that it's diverging from a normal distribution. This might mean that some subjects have a much lower or higher score than expected.

To visualize if there are outliers in the data, we have plotted some of the features in boxplots. In figure 7, we can see that there is no outlier for the test time. This was expected as the test time feature is supposed to represent the beginning of the study until the very end, after six months of recording data. However, for the age, there is an outlier. A subject is aged 36 years old when the mean is usually at 64 years old, with half of the subjects between the age of 58 and 72. Therefore, 36 is quite low.

The impact of age on UPDRS score is represented on a scatter plot with a linear regression line figure (8 and 9). Data are very scattered, but the UPDRS score seems to increase with age.

The result of grouping men and women data together is showed in table IV. The total and motor UPDRS score are higher with men than women. The mean age for men and women is very close, so age should not be a factor explaining the higher scores for men.

We also studied the evolution of the motor UPDRS and total UPDRS scores over time for the 42 subjects, presented in annex I and annex II. The expected outcome of this analysis was to see scores that would get worst over time because PD is a degenerative disease. However, for 15 subjects, their motor UPDRS scores improved between the beginning and the end of the study. [2] noted that it might be because of the physical therapy required by using the device every day, enabling subjects to improve their condition.



(a)

(b)

Fig. 6. (a) Probability plot of the motor UPDRS score and (b) the total UPDRS score

B. Clustering

We have tried to apply a different quantity of clusters (k between two and ten) with the K-means algorithm on two vectors resulting from LSA. At each k size, the result was plotted on a graph to find witch better group subject on the two-dimension representation. The best result is obtained with a k size of four. It is presented on figure 11. On the figure, a big gray dot represents the centroid of each cluster, and each small dot is a subject associated with a cluster by his color. The number at the top-left corner of small dots is the id number of each subject.

C. Covariance

We built a covariance matrix to see if features are similar. The result is illustrated in figure 10.

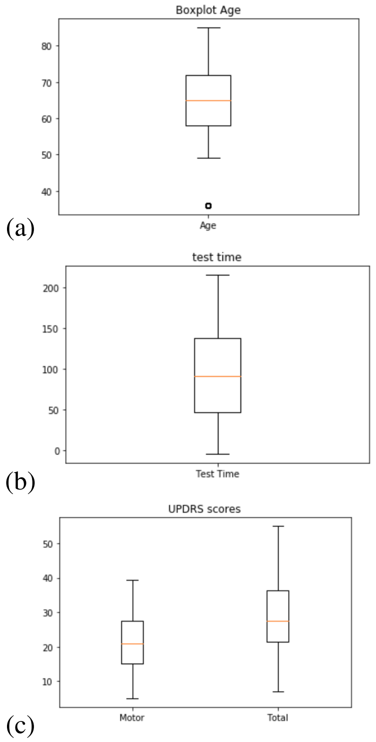


Fig. 7. Boxplot of the (a) age, (b) test time, and (c) UPDRS scores (motor and total)

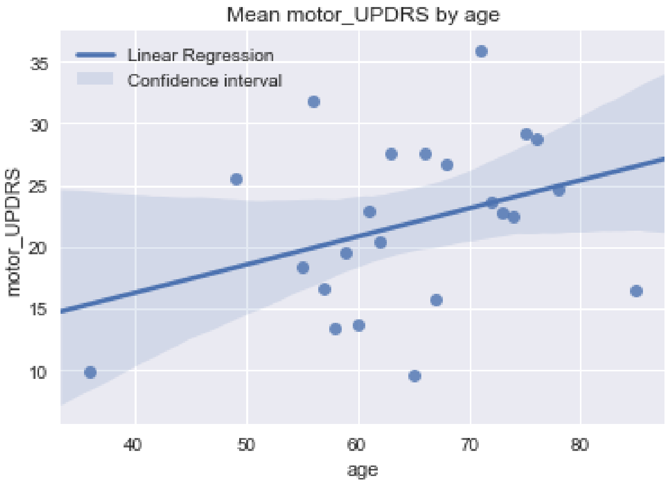


Fig. 8. Correlation with age and motor UPDRS score.

D. Modelisation

The dataset was split into two parts: 80% of the dataset was used for training, and the 20% left was used for testing purposes.

We first used a SVR. We performed a gridsearch over a range of hyperparameters shown in table VII. During the search for the optimal hyperparameters, a cross-validation with 5 folds was performed on the training subset. The best hyperparameters used are shown in bold characters. We tried to predict both the motor and the total UPDRS score. The

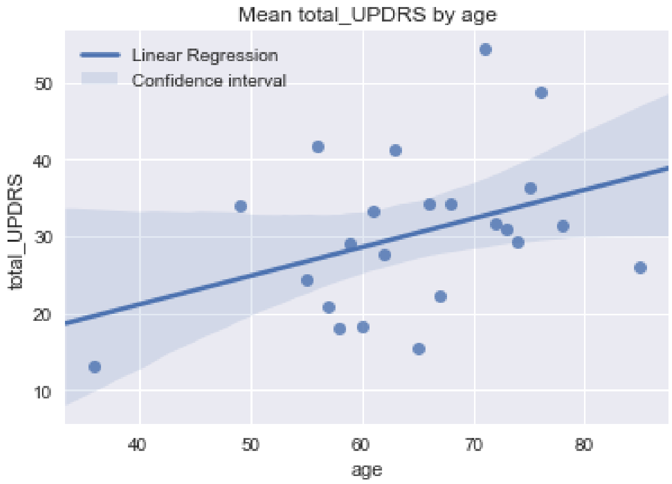


Fig. 9. Correlation with age and total UPDRS score.

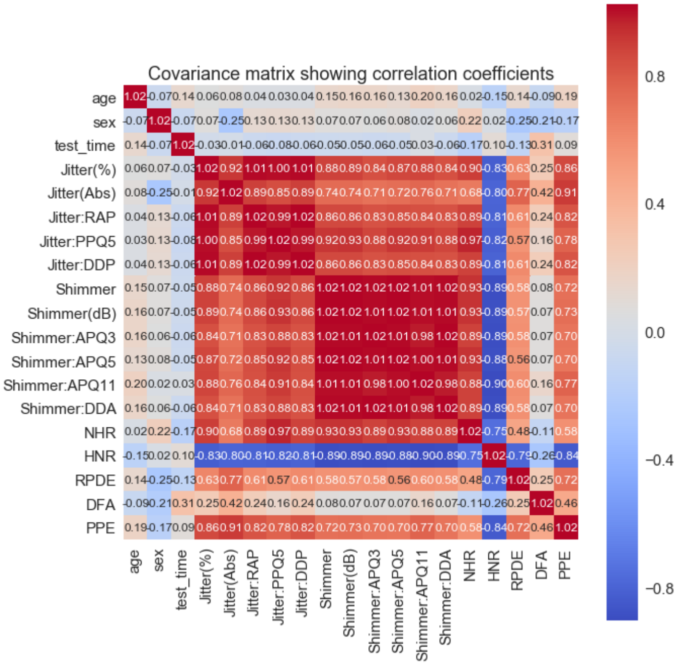


Fig. 10. Covariance matrix of features

results are shown in table V.

The MSE score for no data scaling is much higher than the other data scaling options, from 60.09 to values below 1. We believe this is because we have a feature time $test_time$ where the maximum value recorded is 215. However, some other features like $Jitter(abs)$ have a maximum value of 0.000446. Therefore, we observe a wide range of variation between the features, where one would dominate over the others. This might prevent the SVR from learning well. When we apply some kind of scaling, we can immediately see how helpful that is for the SVR, as we get scores that are close to 0.

Therefore, based on the MSE results, standardized inputs and outputs provide the best results for the SVR. We also

TABLE V
R2 SCORES ON THE TESTING CORPUS FOR THE MOTOR AND TOTAL UPDRS SCORE USING DIFFERENT DATA SCALING TECHNIQUES.

Data Scaling	Motor UPDRS		Total UPDRS	
Including subject_id	Yes	No	Yes	No
None	0.46	0.07	0.36	0.09
Unscaled inputs	**0.52**	0.13	**0.43**	0.14
Normalized inputs	0.19	0.10	0.21	0.14
Standardized inputs	0.25	**0.15**	0.26	**0.15**

TABLE VI
MSE SCORES ON THE TESTING CORPUS FOR THE MOTOR AND TOTAL UPDRS SCORE USING DIFFERENT DATA SCALING TECHNIQUES.

Data Scaling	Motor UPDRS		Total UPDRS	
Including subject_id	Yes	No	Yes	No
None	60.09	63.37	98.70	106.81
Unscaled inputs	0.89	0.95	0.84	0.92
Normalized inputs	**0.86**	0.95	0.84	0.92
Standardized inputs	**0.86**	**0.94**	**0.82**	**0.91**

experimented with keeping the subject id as a feature and removing it. Without surprise, keeping the subject identification yields better results. The intuition behind this might be because the UPDRS score is subjective and dependent on who the subject is, which is why keeping this information helps get better results.

The R2 scores for the SVR are interesting. We would have expected the R2 score to be higher for standardized inputs, following the trend of the MSE. However, it is higher for unscaled inputs, where it is 52%, meaning 52% of the data fits the regression model. We do not know how to explain this behavior, and when an MSE tends to zero, usually, the R2 should tend to 1, confirming the fit of the model. It is not the case here. It seems that the feature has less influence on the label, but even while having explaining less the variance, they provide better results as the MSE is lower then.

Based on K-mean clustering, it seems that subjects can be grouped together in four categories based on vowel pronunciation. If we compare figure 11 and table X in annex III, we find that there is no relation between groups and UPDRS scores. Subjects in the same cluster can have a very high UPDRS score, and the same cluster will contain a low UPDRS score. We suppose that subject in the same cluster have their voice affected by the same manners by the disease and it doesn't mean that their disease reach the same level on the UPDRS scale.

In the covariance matrix, there is a lot of high scores. It supposes that there are many features correlated. By these correlations, we think that some features are not useful. Therefore, we could reduce the dimensionality of the features by eliminating features that do not provide much information. This reduction would be useful to reduce the complexity of the data analysis.

TABLE VII
HYPERPARAMETERS TRIED DURING THE GRIDSEARCH OF THE SVR TO FIND THE MOST OPTIMAL VALUES, WHICH ARE HIGHLIGHTED IN BOLD.

Hyperparameter	**Values**
Kernel	**RBF**, Linear
Gamma	$\mathbf{1 \times 10^{-3}}$, 1×10^-4
Epsilon	0.01, 0.1, 0.5, **1**, 10
C	0.1, 1, 10, **100**

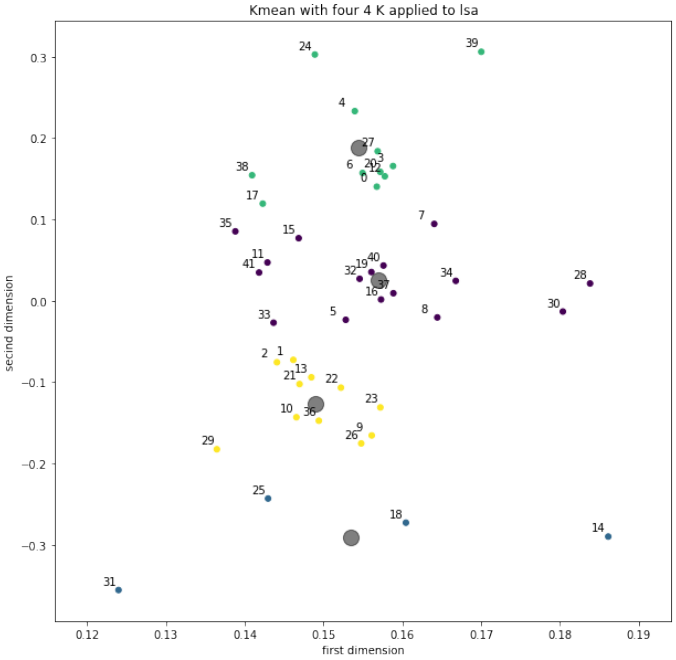


Fig. 11. Clustering on 2d LSA representation

TABLE VIII
MSE SCORES OF THE MLP ON THE TESTING CORPUS FOR THE MOTOR AND TOTAL UPDRS SCORE USING DIFFERENT DATA SCALING TECHNIQUES.

Data Scaling	Motor UPDRS		Total UPDRS	
Including subject_id	Yes	No	Yes	No
None	nan	69.607	121.255	nan
Unscaled inputs	1.067	nan	nan	nan
Normalized inputs	0.152	0.425	0.176	0.429
Standardized inputs	**0.118**	0.396	0.125	0.384

We performed the learning of their multiple layer perception model using the same training and testing dataset as the SVR model. The perceptron model has a layer of 20 units with relu activation followed by a layer of 20 units with linear activation. The training is done over 200 epochs. We tried to predict both the motor and the total UPDRS score executing 30 times the learning and predicting session for different scalers. We thought that the subjects' id could skew data because the same id is used multiple times for a subject. This is why the learning is done with and without the subject id. The results are shown in table VIII. The best result is got predicting MOTOR UPDRS with the subject ID in the learning data set. Nan values mean that the model is unable to learn the problem because of a very large error gradient for weight updates **??**. Figure 12 prove that the leaning is effective because of the constant downward curve.

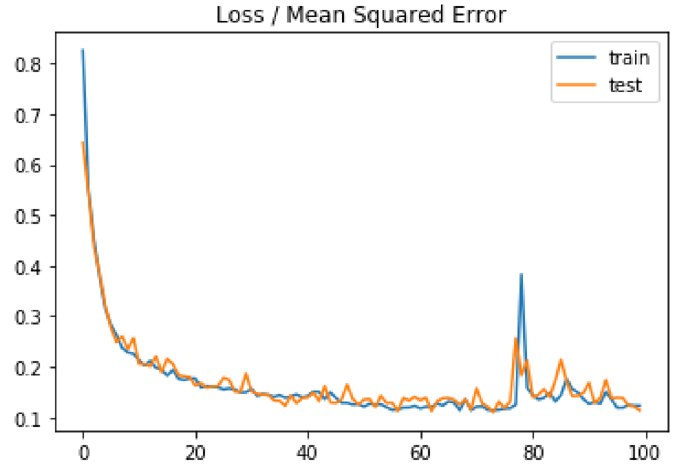


Fig. 12. Loss by mean square error over epochs. The X axis represents the number of epochs, and Y represents the MSE

VI. Discussion

The fact that predicting UPDRS scores is better when the subject id is included, it means that the score is subjective to each subject. This also follows our intuition that the severity and the progression of disease are, of course, closely related to the patient, as each human experiences differently the symptoms of PD.

The MLP network seems to have the best potential to predict the motor and total UPDRS scores. The total UPDRS score includes other features that are not related to motor symptoms, like the mood of the patient, daytime sleepiness, turning in bed, etc. These are all symptoms of PD that, at first, we thought were not possible to predict using only speech. However, we have very good results on the total score. This means that it is actually possible, from the speech, the predict well symptoms that are not related at all to speech.

We used the same database as shared in [2]. The authors of the paper used the models iteratively reweighted least squares (IRLS), least absolute shrinkage and selection operator (LASSO), and CART, which is a decision tree. We compared our results with the results obtained in the paper, and the results are in table IX. We can see that the SVR has a slightly better MAE than Lasso, but IRLS is still better.

Of course, when we scale the data, the MAE is decreasing tremendously because it is a measure that is very sensitive to the range of the target variable. As MLP can't provide results from a data which was not scaled in any way, we can't obtain an MAE.

TABLE IX
Comparison of the testing MAE results we got from our SVR and MLP, and the results of the original paper using the same database. They used an IRLS model, lasso and CART.

Model	Data Scaling	Testing MAE
SVR	**None**	8.14
	Unscaled	0.73
	Normalized	0.76
	Standardized	0.75
MLP	**None**	8.89
	Unscaled	-
	Normalized	-
	Standardized	-
IRLS	**None**	6.7
Lasso	**None**	8.5
Cart	**None**	7.5

VII. Conclusion

The device AHTD is one of the first work to study the feasibility of at-home monitoring for PD subjects. It opens the way for future work to be done in uncontrolled environments. This device is precise with the detection of 16 vocal features adapted to detect PD voice symptoms. Research was done on those 16 features to understand them and explore their values in the dataset. We looked at the data's central tendency measures, made probability plots and boxplots to analyze the dataset we are working with. This exploration was useful to confirm that the database is valid for future modeling approaches and does not contain outliers to remove. We also did some analysis of the correlation between UPDRS scores and age of subjects. The severity of the disease directly correlates with the age of the subject that came to light during this data exploration step.

K-mean on LDA reduction lean to discovering four groups of subjects who seem to have similar voice features. When predicting the motor and total UPDRS scores, the MLP provides better results than the SVR, especially when using standardized inputs. We conclude that vocal features can indeed be used to predict the progression of Parkinson's Disease. Future work could be done to analyze if some features are more useful than others to predict the scores and witch can be removed.

Code Available

The code to reproduce the experiments presented in this paper is available online at https://github.com/Mymoza/DataMining-UPDRS/

Acknowledgment

We want to thank Laureano Moro-Velazquez for his guidance and comments on speech and Parkinson's Disease, Sylvie Ratté for her support in the choice of the database and advice on the objectives, and Desh Raj for his help on scientific writing.

References

[1] Goetz, Christopher G and Stebbins, Glenn T and Wolff, David and DeLeeuw, William and Bronte-Stewart, Helen and Elble, Rodger and Hallett, Mark and Nutt, John and Ramig, Lorraine and Sanger, Terence and others,"Testing objective measures of motor impairment in early Parkinson's disease: Feasibility study of an at-home testing device," *Movement Disorders*, Vol. 24, No. 4, 551–556,(2009). Wiley Online Library.

[2] Tsanas, Athanasios and Little, Max A and McSharry, Patrick E and Ramig, Lorraine O,"Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests," *IEEE transactions on Biomedical Engineering*, Vol. 57, No. 4, 884–893,(2009). IEEE.

[3] Teixeira, João Paulo and Oliveira, Carla and Lopes, Carla,“Vocal acoustic analysis–jitter, shimmer and hnr parameters,” *Procedia Technology*, Vol. 9, 1112–1122,(2013). Elsevier.

[4] Goetz, Christopher G and Tilley, Barbara C and Shaftman, Stephanie R and Stebbins, Glenn T and Fahn, Stanley and Martinez-Martin, Pablo and Poewe, Werner and Sampaio, Cristina and Stern, Matthew B and Dodel, Richard and others,“Movement Disorder Society-sponsored revision of the Unified Parkinson’s Disease Rating Scale (MDS-UPDRS): scale presentation and clinimetric testing results,” *Movement disorders: official journal of the Movement Disorder Society*, Vol. 23, No. 15, 2129–2170,(2008). Wiley Online Library.

[5] ParkinsonCanada,“Understanding Parkinson’s,”, (2020). https://www.parkinson.ca/about-parkinsons/understanding-parkinsons/

[6] Fahn, SRLE,“Unified Parkinson’s disease rating scale,” *Recent development in Parkinson’s disease*,(1987). Macmillan.

[7] Awan, Shaheen N and Frenkel, Michael L,“Improvements in estimating the harmonics-to-noise ratio of the voice,” *Journal of Voice*, Vol. 8, No. 3, 255–262,(1994). Elsevier.

[8] Smith, Marshall E and Ramig, Lorraine Olson and Dromey, Christopher and Perez, Kathe S and Samandari, Ráz,“Intensive voice treatment in Parkinson disease: laryngostroboscopic findings,” *Journal of Voice*, Vol. 9, No. 4, 453–459,(1995). Elsevier.

[9] Brownlee “How to use Data Scaling Improve Deep Learning Model Stability and Performance,”, (2019). https://machinelearningmastery.com/how-to-improve-neural-network-stability-and-modeling-performance-with-data-scaling/

[10] Dumais, Susan T,“Latent semantic analysis,” *Annual review of information science and technology*, Vol. 38, No. 1, 188–230,(2004). Wiley Online Library.

[11] Hoffman, Matthew and Bach, Francis R and Blei, David M,“Online learning for latent dirichlet allocation,” *advances in neural information processing systems*, 856–864,(2010).

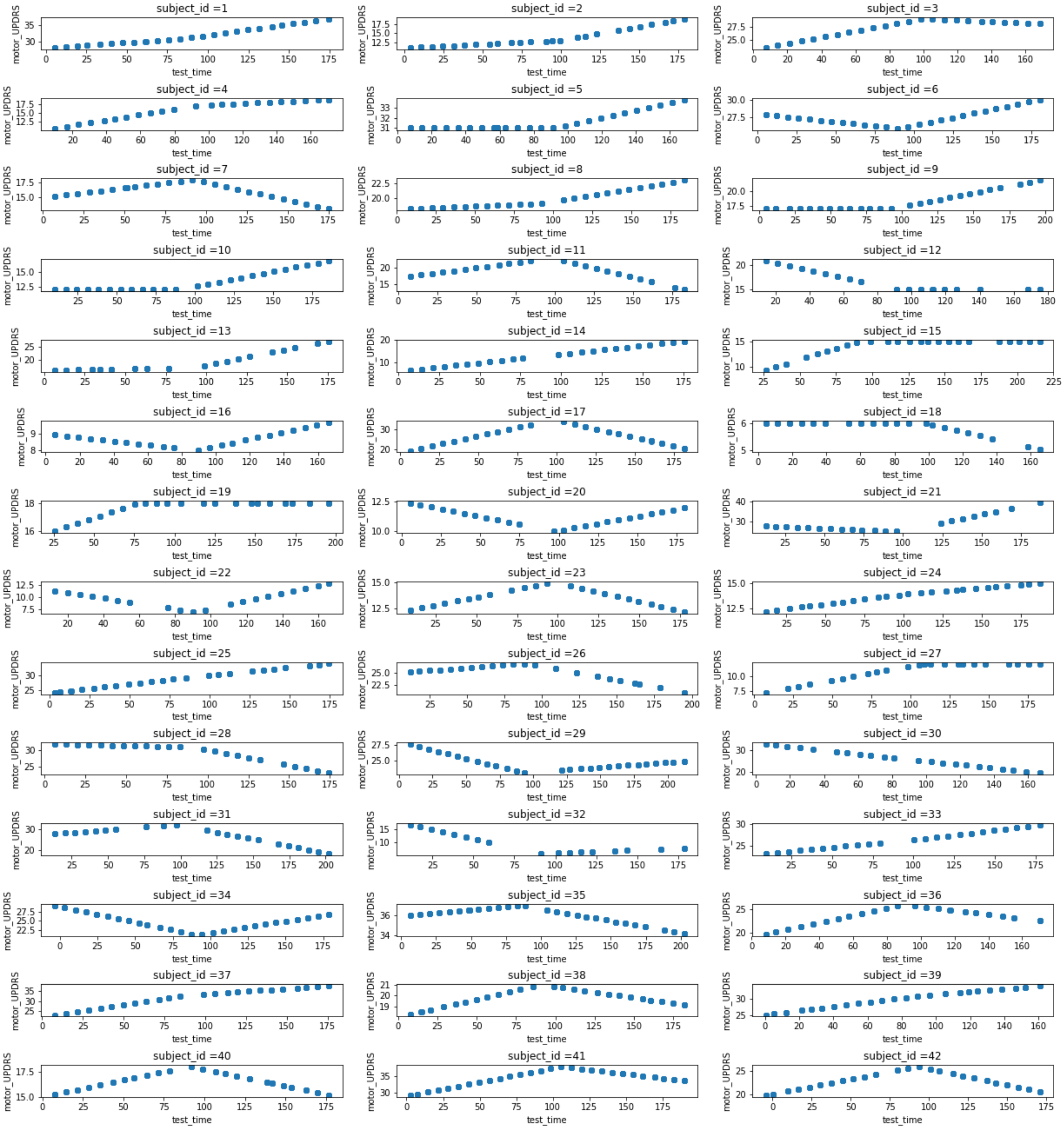
[12] Anaya, Leticia H, *Comparing Latent Dirichlet Allocation and Latent Semantic Analysis as Classifiers.*, ERIC, 2011.

[13] Pal, Sankar K and Mitra, Sushmita,“Multilayer perceptron, fuzzy sets, classifiaction,” , (1992).

[14] Wikipedia “Coefficient of determination,”, (2020). https://en.wikipedia.org/wiki/Coefficient_of_determination

[15] Santini, Marina,“Advantages & disadvantages of k-means and hierarchical clustering (unsupervised learning),” *URL: http://santini.se/teaching/ml/2016/Lect_10/10c_Unsupervise dMethods. pdf (Accesed 17.04. 2019)*,(2016).

VIII. Annex I : Evolution of the motor_UPDRS score over time since the trial started for the 42 subjects



IX. Annex II : Evolution of the total UPDRS score over time since the trial started for the 42 subjects

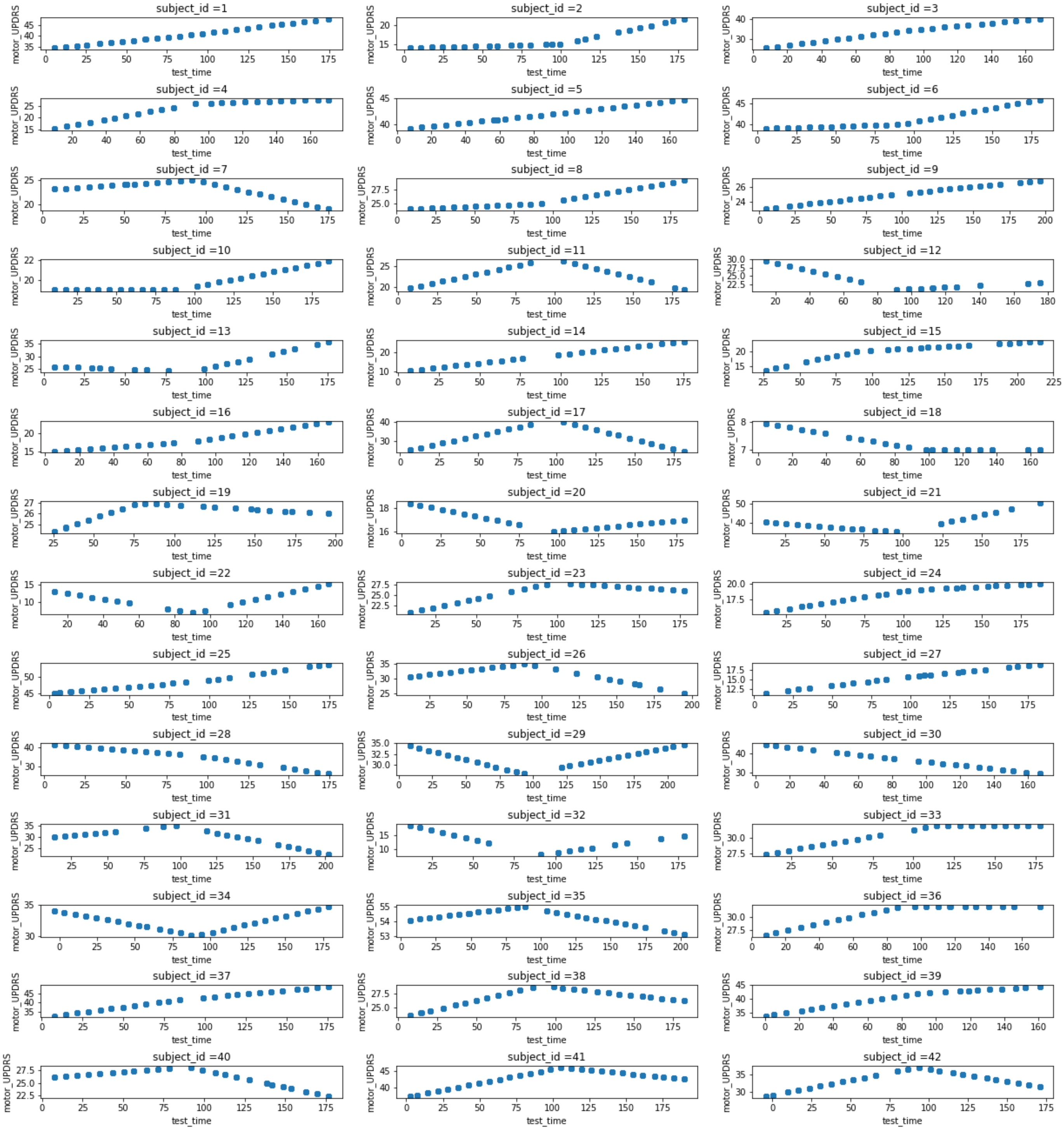


TABLE X
Mean UPDRS score of every subject.

subject_id	motor_UPDRS	total_UPDRS
1	31.90	40.73
2	13.81	16.28
3	27.12	33.36
4	15.79	23.59
5	31.63	41.85
6	27.53	41.34
7	16.05	23.07
8	19.89	25.89
9	18.31	25.08
10	13.42	19.85
11	18.99	22.86
12	16.89	24.14
13	19.52	27.84
14	13.01	18.02
15	13.96	19.88
16	8.71	18.56
17	26.43	31.94
18	5.82	7.30
19	17.61	26.16
20	11.18	16.93
21	29.09	40.14
22	9.80	10.97
23	13.47	25.46
24	13.76	18.28
25	28.73	48.64
26	25.04	31.33
27	10.79	15.43
28	29.17	34.89
29	24.63	31.51
30	25.92	36.90
31	26.40	29.56
32	9.94	13.04
33	26.37	30.46
34	24.68	32.32
35	35.99	54.25
36	23.39	30.42
37	31.86	41.62
38	19.77	26.81
39	29.88	40.10
40	16.51	25.95
41	34.40	42.60
42	22.84	33.24

X. Annex III : Mean UPDRS score of subjects