Performance, Transparency and Time. Feature selection to speed up the diagnosis of Parkinson’s disease

Pierluigi Costanzo
costanzo.p@live.unic.ac.cy
University of Nicosia
Cyprus

Kalia Orphanou
orphanou.k@unic.ac.cy
University of Nicosia
Cyprus

Abstract
Accurate and early prediction of a disease allows to plan and improve a patient’s quality of future life. During pandemic situations, the medical decision becomes a speed challenge in which physicians have to act fast to diagnose and predict the risk of the severity of the disease, moreover this is also of high priority for neurodegenerative diseases like Parkinson’s disease. Machine Learning (ML) models with Features Selection (FS) techniques can be applied to help physicians to quickly diagnose a disease. FS optimally subset features that improve a model performance and help reduce the number of needed tests for a patient and hence speeding up the diagnosis. This study shows the result of three Feature Selection (FS) techniques pre-applied to a classifier algorithm, Logistic Regression, on non-invasive test results data. The three FS are Analysis of Variance (ANOVA) as filter based method, Least Absolute Shrinkage and Selection Operator (LASSO) as embedded method and Sequential Feature Selection (SFS) as wrapper method. The outcome shows that FS technique can help to build an efficient and effective classifier, hence improving the performance of the classifier while reducing the computation time.

Keywords: Parkinson’s disease, Machine Learning, Logistic Regression, Feature Selection

1 Introduction
Parkinson’s disease (PD) is a neurodegenerative disorder that leads to partial or full loss in motor reflexes, difficulties in speech, mental processing, and other vital functions [3]. As a degenerative disorder, it gets worse over time. There is no known cure at the moment but only therapies that can help reduce the effects of the difficulties this disease causes[10], hence it’s crucial to have tools that can diagnose it at an early stage possible. Voice problem is an effect of PD and tests based on speech data analysis are one of the non-invasive diagnostic tools to diagnose the disease, they are also low cost and easy to self-use [5].

Machine Learning (ML) techniques applied to biomedical data[2] have to fulfill critical characteristics of the medical decision process[4]: performance, transparency, time. It is vital that a patient diagnosis is correct and, hence the ML algorithm predicts the diagnosis as accurately as possible.

Trust is a crucial component between the physician and the patient for the physician to give and for the patient to get the best possible care. This is translated in ML as model explainability, so called “glass-box” ML algorithms are models whose output is explainable and they can be used trustworthy in the process of medical decisions. Moreover, collecting tests for a patient diagnosis is expensive and time-consuming, and in critical situations, time might not be an option.

This research study made use of data from non-invasive tests on healthy and PD patients. As the dataset include multiple record for each individual, stratified leave-one-subject-out (LOSO) cross-validation technique was used. Three FS methods were tested with two different metric optimization techniques, accuracy and cross-entropy. Different models were output and the results were compared with each other and with a baseline model without any FS technique. The evaluation was intended as classifying individual by the majority of the multiple records for each of them.

In Section 2 the dataset used for the early diagnosis of Parkinson is described. In Section 3 the methodology to test the different FS techniques is described, which ML classifier has been used, the data preprocessing steps, more details about the three FS, evaluation of the best feature selected and evaluation for comparing the results of the different models. Section 4 is an overview of the results and Conclusions are in Section 5.

2 Dataset
The dataset used is the result of a study made on 40 individuals, 20 People with Parkinsonism (PWP) and 20 healthy individuals, who voluntarily participated in the research at the Department of Neurology in Cerrahpaşa Faculty of Medicine, Istanbul University [8]. All individuals recorded 26 voice samples of sustained vowels, numbers, words, and short sentences. The Table 1 summarizes the features in the dataset.

There are 5 statistics of Pitch of vocal oscillation, which measure how high or low is the sound based on the frequency of vibration of the sound waves produced: median, mean, standard deviation, minimum, maximum. There are 5 Jitter metrics, which measure the variation in the frequency of the
Table 1. Features Summary

| ID | Feature                      | Description                                      | Type   | MEAN   | STD. DEV | MIN    | MAX    |
|----|------------------------------|--------------------------------------------------|--------|--------|----------|--------|--------|
| 1  | Jitter (local)               | Variation in the frequency of the sound          | FLOAT  | 2.68   | 1.78     | .19    | 14.38  |
| 2  | Jitter (local, absolute)     |                                                  | FLOAT  | .0002  | .0001    | .00006 | .0008  |
| 3  | Jitter (rap)                 |                                                  | FLOAT  | 1.25   | .98      | .06    | 8.02   |
| 4  | Jitter (ppq5)                |                                                  | FLOAT  | 1.35   | 1.14     | .08    | 13.54  |
| 5  | Jitter (ddp)                 |                                                  | FLOAT  | 3.74   | 2.94     | .19    | 24.05  |
| 6  | Shimmer (local)              | Variation in the amplitude of the sound          | FLOAT  | 12.92  | 5.45     | 1.19   | 41.14  |
| 7  | Shimmer (local, dB)          |                                                  | FLOAT  | 1.19   | .42      | .10    | 2.72   |
| 8  | Shimmer (apq3)               |                                                  | FLOAT  | 5.70   | 3.02     | .50    | 25.82  |
| 9  | Shimmer (apq5)               |                                                  | FLOAT  | 7.98   | 4.84     | .71    | 72.86  |
| 10 | Shimmer (apq11)              |                                                  | FLOAT  | 12.21  | 6.02     | .52    | 44.76  |
| 11 | Shimmer (dda)                |                                                  | FLOAT  | 17.10  | 9.05     | 1.49   | 77.46  |
| 12 | Autocorrelation              |                                                  | FLOAT  | .85    | .09      | .54    | .99    |
| 13 | Noise-to-Harmonic            | Noise in the sound assessment                     | FLOAT  | .23    | .15      | .002   | .87    |
| 14 | Harmonic-to-Noise            |                                                  | FLOAT  | 9.99   | 4.29     | .70    | 28.42  |
| 15 | Median pitch                 | How high or low is the sound                     | FLOAT  | 163.37 | 56.02    | 81.46  | 468.62 |
| 16 | Mean pitch                   |                                                  | FLOAT  | 168.73 | 55.97    | 82.36  | 470.46 |
| 17 | Standard dev. of pitch       |                                                  | FLOAT  | 27.55  | 36.67    | 0.53   | 293.88 |
| 18 | Minimum pitch                | How high or low is the sound                     | FLOAT  | 134.54 | 47.06    | 67.96  | 452.08 |
| 19 | Maximum pitch                |                                                  | FLOAT  | 234.86 | 121.54   | 85.54  | 597.97 |
| 20 | Number of pulses             | Pulse related metrics                             | INT    | 109.74 | 150.03   | 0      | 1490   |
| 21 | Number of periods            |                                                  | INT    | 105.97 | 149.42   | 0      | 1489   |
| 22 | Mean period                  |                                                  | FLOAT  | .007   | .002     | .002   | .01    |
| 23 | Standard dev. of period      |                                                  | FLOAT  | .001   | .001     | .0001  | .01    |
| 24 | Fraction of unvoiced frames  |                                                  | FLOAT  | 27.68  | 20.98    | 0.00   | 88.16  |
| 25 | Number of voice breaks       | Voice related metrics                             | INT    | 1.13   | 1.61     | 0      | 12     |
| 26 | Degree of voice breaks column|                                                  | FLOAT  | 12.37  | 15.16    | 0.00   | 69.12  |

sound: jitter in %, absolute jitter in microseconds, jitter as relative amplitude perturbation (rap), jitter as 5-point period perturbation quotient (ppq5) and jitter as average absolute difference of differences between jitter cycles (ddp). There are 6 Shimmer metrics, which measure the variation in the amplitude of the sound: shimmer in %, absolute shimmer in decibels (dB), shimmer as 3-point, 5-point and 11-point amplitude perturbation quotient (apq3, apq5, apq11), and shimmer as average absolute differences between the amplitudes of shimmer cycles (dda). There are Noise-to-Harmonic (NHR) and Harmonic-to-Noise (HNR) ratios that assess the noise in the sound, and a correlation of the two. There are 4 pulse related metrics: number of pulses, number of periods mean of period and standard deviation of period. There are 3 voice related metrics: fraction of unvoiced frames, number and degree of voice breaks.

The dataset also includes the binary class information, which is used for training and testing the model, and the Unified Parkinson’s Disease Rating Scale (UPDRS) score which instead has been excluded in this study.

3 Methodology
3.1 Proposed Method
The dataset is composed of 40 individuals and 26 recordings for each individual. A stratified leave-one-subject-out (LOSO)[9] k-fold (with k=4) cross-validation technique was used to divide the training dataset for evaluating the performance of each FS method. As the number of subjects is only 40, k=4 was used to evaluate the model 4 times with 25% of the subjects held out as validation set. The LOSO was used to avoid data leakage, hence all the recordings of a subject were or in the held in training set or in the held out validation set. The stratification was used to balance PWP and healthy individuals in the training and validation set.

At each iteration, a preprocessing step was applied (power transformation and standardization) and three feature selection techniques were tested with two different metrics optimization, maximization of the accuracy and minimization of the cross-entropy loss.

The Logistic Regression model was the algorithm chosen. It is a linear model, a gradient descent based algorithm
used in classification problems where the probabilities of the outcome are modeled with a logistic function. A binary logistic model, as in this study, has a dependent variable with only two possible outcomes. Moreover as stated in Section 1, model explainability is important, hence the choice of this model as Logistic Regression is a “glass-box” ML algorithm and the output is easily explainable.

The performance of the Logistic Regression model trained with all the features was compared with each performance of the model trained with the best subset of features from each FS techniques tested.

3.2 Preprocessing the data
In Machine Learning, preprocessing the data is important for training and testing effectively and efficiently a gradient descent based algorithms classifier, like logistic regression. The difference in feature ranges, variance and skewness, can influence the algorithm and make it difficult and slower to converge toward the minima (error minimization). In this study, a Power Transformation (Yeo-Johnson transformation [11]) and a Standardization step was applied. At each iteration of the k-fold, the transformation steps were fit with the train fold and applied to both train and test fold before training and evaluating the model.

3.3 Feature selection techniques
The aim of feature selection techniques is to discard redundant features from the dataset before training a ML model, improving the performance of the model and reducing the execution time. In this study, three feature selection techniques were tested, Analysis of Variance (ANOVA) as filter based method, Least Absolute Shrinkage and Selection Operator (LASSO) as embedded method and Sequential Feature Selection with forward selection (SFS) as wrapper method.

Filter based FS methods rank the features based on their independence of the class labels [1]. ANOVA and \( \chi^2 \) tests calculate the dependency between two variables, the first is used between numerical and categorical variables, the second for categorical variables only. In this study the ANOVA test was used between all the input features and the class as the dataset contains numerical features. The advantage of using these techniques is that they are light computationally and there is less risk of overfitting, on the other hand the features selected might not be an optimal choice as redundant features are also captured as the test is independently performed between each feature and the class. From the 1st to all features ranked, the technique was evaluated against accuracy and cross-entropy loss scores to select the best performing features.

Wrapper based FS methods are greedy search approaches which perform an evaluation of the best possible subset of feature combination against an evaluation criteria [1]. There are Selection Algorithms and Heuristic Search Algorithms and for this study the SFS was tested, evaluating the subset of features against accuracy and cross-entropy loss scores. These techniques are computationally expensive and can lead to overfitting.

Embedded based FS methods include FS during the training process and they are techniques which might vary from ML algorithm to algorithm [1]. In this study the LASSO (L1 regularization) [7] technique was tested and evaluated against accuracy and cross-entropy loss scores.

The accuracy score measures the percentage of correctly classified class. In this study is intended as correctly classified individuals by the majority of the 26 records classified for each individual. For the FS process, the subset of features was chosen by maximizing the accuracy score.

The cross-entropy loss score, or negative log-likelihood, measure the distance between the predicted class probability and the actual class. In this study, it is intended as the distance between the predicted classified individuals probability, the mean of the predicted class probability of the 26 records classified for each individual, and the actual classified individuals. For the FS process, the subset of features was chosen by minimizing the cross-entropy loss score (see Eq. 1).

\[
H(p, q) = - \sum_i p_i \log q_i
\]

with \( p \) the actual class label, which in binary classification is 0 and 1, and \( q \) the predicted class probability.

3.4 Model + Feature selected performance evaluation
For each FS technique, the best features were selected based on maximization of the accuracy and minimization of the cross-entropy loss. The filter and embedded based FS methods are ranking methods, hence the subsets of features were extracted by ranking them and comparing the accuracy and cross-entropy loss for 1 to 25 features. The wrapper based FS methods are not ranking methods hence the subsets of features were extracted by comparing the accuracy and cross-entropy loss for the best subsets of 1 to 25 features.

Figure 1 and Figure 2 describe the performance of the FS filter technique. For each combination of features, from 1 to 25, the performance of the accuracy and the cross-entropy has been evaluated. Based on the accuracy (figure 1), 1 feature yield the best metric. Based on the cross-entropy (figure 2), 7 features yield the best metric.

Similarly to the FS embedded technique, for each combination of features, ranked, the performance of the accuracy and the cross-entropy has been evaluated. Based on the accuracy (figure 3), 4 features yield the best metric. Based on the cross-entropy (figure 4), 9 features yield the best metric.
Table 2. Results Summary

| FS   | Strategy       | # FS | Accuracy     | Specificity | Sensitivity (Recall) | Precision | F1-Score | MCC  |
|------|----------------|------|--------------|-------------|----------------------|-----------|----------|------|
| -    | -              | 26   | .525 (.150)  | .550 (.191) | .525 (.150)          | .527 (.162) | .514 (.153) | .052 (.311) |
| ANOVA| MAX accuracy   | 1    | .675 (.096)  | .650 (.342) | .675 (.096)          | .736 (.109) | .653 (.116) | .399 (.191) |
|      | MIN cross-entropy | 7    | .600 (.245)  | .600 (.283) | .600 (.245)          | .607 (.255) | .594 (.248) | .207 (.500) |
| SFS  | MAX accuracy   | 1    | .675 (.096)  | .650 (.342) | .675 (.096)          | .736 (.109) | .653 (.116) | .399 (.191) |
|      | MIN cross-entropy | 3    | .700 (.082)  | .650 (.300) | .700 (.082)          | .749 (.107) | .686 (.086) | .445 (.182) |
| LASSO| MAX accuracy   | 4    | .625 (.171)  | .650 (.300) | .625 (.171)          | .672 (.209) | .611 (.165) | .293 (.375) |
|      | MIN cross-entropy | 9    | .575 (.171)  | .600 (.283) | .575 (.171)          | .594 (.197) | .567 (.167) | .168 (.367) |

**Figure 1.** The Accuracy performance of the ranked features with the ANOVA technique. The dashed line indicate the best performance (highest accuracy).

**Figure 2.** The cross-entropy performance of the ranked features with the ANOVA technique. The dashed line indicate the best performance (lowest cross-entropy).

**Figure 3.** The Accuracy performance of the ranked features with the LASSO technique. The dashed line indicate the best performance (highest accuracy).

**Figure 4.** The cross-entropy performance of the ranked features with the LASSO technique. The dashed line indicate the best performance (lowest cross-entropy).
Performance, Transparency and Time. Feature selection to speed up the diagnosis of Parkinson’s disease

Figure 5. The accuracy performance with the accuracy strategy of the subset of 1 to 25 features with the SFS technique. The dashed line indicate the best performance (highest accuracy).

Figure 6. The accuracy performance with the cross-entropy strategy of the subset of 1 to 25 features with the SFS technique. The dashed line indicate the best performance (highest accuracy).

Differently to the FS filter and embedded technique, for the wrapper based FS method, the subsets of features were extracted by comparing the metrics for the best subsets of 1 to 25 features. In other words, the SFS methodology starts from an empty subset of features, and each feature is added to the subset if yield better result in terms of metric strategy. Figure 5 and Figure 6 describe the performance of the strategy of maximization of the accuracy and minimization of the cross-entropy loss respectively, in terms of accuracy performance of the subset selected. Based on the maximization of the accuracy strategy (Figure 5), 1 feature yields the best performance. Based on the minimization of the cross-entropy loss strategy (figure 6), 3 features yield the best performance.

The Logistic Regression model trained with all the features was compared with the model trained with the best subset of feature from the FS techniques tested. The performance comparison was based on accuracy, specificity, sensitivity (recall), precision, F1-score, and Matthews correlation coefficient (MCC) scores[8]. These scores were averaged after the k-fold cross-validation. As for the accuracy previously, in this study the performance metrics are intended as individuals classified by the majority of the 26 records classified for each individual.

The specificity measures the true negative rate out of the actual negative class, the sensitivity (recall) measures the true positive rate out of the actual positive class, the precision measures the true positive rate out of the predicted positive class, the F1-score is the harmonic mean of the precision and recall.

The MCC score[6] measures the quality of a binary classifier, a correlation coefficient between predicted and actual class. It ranges between -1, the worst classifier and +1, the best classifier, and a value of zero indicates a random classifier (see Eq. 2).

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

4 Experimental Results

The results are summarized in the Table 2. The classes in the dataset are balanced, hence there is no difference between Accuracy and Sensitivity. In the biomedical field the Sensitivity is an important metric of how correctly a classifier detect individuals with a condition, hence it’s possible to early diagnosis a disease and increases the chance of therapies treatment that might help the symptoms of a disease from becoming worse. All the FS technique lead to better performance than the model trained with all features.

The SFS technique with MIN cross-entropy loss strategy is the technique that achieve the best performance, with a subset of 3 features, Jitter (ppq5), Shimmer (apq11) and the correlation between NHR and HNR.

The ANOVA and SFS technique both with MAX accuracy strategy achieve the same performance and yield to the same subset of 1 feature, the Jitter (ppq5). The ANOVA technique with MIN cross-entropy loss strategy and LASSO technique with both MAX accuracy strategy and MIN cross-entropy loss strategy achieve the lowest performance and yields a subset of 7, 4 and 9 feature respectively. They all include the features Jitter (ppq5) and Shimmer (apq11).

5 Conclusions

From the test performed with the FS techniques, by reducing the number of feature to train a model, it’s possible to speed
up ML inference and increase at same time the model performance for diagnosis Parkinson’s disease. The approach proposed is only part of all the possible classifiers and feature techniques that can be applied but it gives the idea of the opportunities in this field of research. Regarding the features, Jitter (ppq5), Shimmer (apq11) are common features selected by each FS technique. Next step is to understand why and how these subset of features yield the most information that lead to improving the classifier performance for PD detection.

Acknowledgments
This research is the output of the course COMP-592DL Project in Data Science attended during the Master of Science in Data Science at the University of Nicosia.

Acknowledge and thanks to Dr Ioannis Katakis and Dr Kalia Orphanou, for the help and the support, not only during the writing of this paper but also during the entire course of study.

References
[1] Girish Chandrashekar and Ferat Sahin. 2014. A survey on feature selection methods. Computers & Electrical Engineering 40, 1 (2014), 16–28.
[2] Kenneth R. Foster, Robert Koprowski, and Joseph D. Skufca. 2014. Machine learning, medical diagnosis, and biomedical engineering research-commentary. Biomedical engineering online 13, 1 (2014), 1–9.
[3] Joseph Jankovic. 2008. Parkinson’s disease: clinical features and diagnosis. Journal of neurology, neurosurgery & psychiatry 79, 4 (2008), 368–376. pmid:18344392.
[4] Igor Kononenko, Ivan Bratko, and Matjaž Kukar. 1997. Application of machine learning to medical diagnosis. Machine Learning and Data Mining: Methods and Applications 389 (1997), 408.
[5] Max Little, Patrick McSharry, Eric Hunter, Jennifer Spielman, and Lorraine Ramig. 2008. Suitability of dysphonia measurements for telemonitoring of Parkinson’s disease. Nature Precedings (2008), 1.
[6] Brian W. Matthews. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochimica et Biophysica Acta (BBA)-Protein Structure 405, 2 (1975), 442–451.
[7] Andrew Y. Ng. 2004. Feature selection, L 1 vs. L 2 regularization, and rotational invariance. In Proceedings of the twenty-first international conference on Machine learning. 78.
[8] B. E. Sakar, M. E. Isenkul, C. O. Sakar, A. Sertbas, F. Gurgen, S. Delil, H. Apsydin, and O. Kursun. 2013. Collection and Analysis of a Parkinson Speech Dataset With Multiple Types of Sound Recordings. Journal of medical systems, 828-834 pages. https://doi.org/10.1109/JBHI.2013.2245674 ID: 1.
[9] C. Okan Sakar and Olcay Kursun. 2010. Telediagnosis of Parkinson’s disease using measurements of dysphonia. Journal of medical systems 34, 4 (2010), 591–599.
[10] Neha Singh, Viness Pillay, and Yahya E. Choonara. 2007. Advances in the treatment of Parkinson’s disease. Progress in neurobiology 81, 1 (2007), 29–44.
[11] In-Kwon Yeo and Richard A. Johnson. 2000. A new family of power transformations to improve normality or symmetry. Biometrika 87, 4 (2000), 954–959.