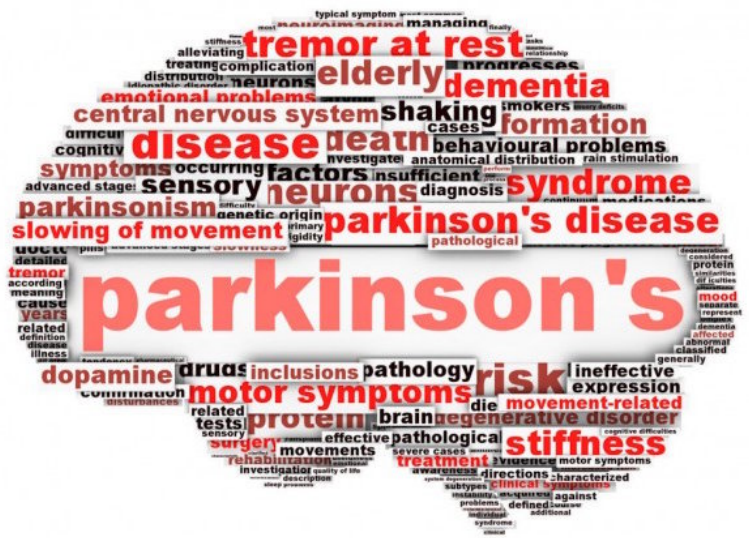
Natural language analysis to detect Parkinson's disease

Paula Andrea Pérez-Toro[1], **Juan Camilo Vásquez-Correa**[1,2], Martin Strauss[2], Juan Rafael Orozco-Arroyave[1,2], and
Elmar Nöth[2]
[1] Faculty of Engineering, University of Antioquia, Medellin, Colombia
[2] Pattern Recognition Lab, Friedrich-Alexander University of Erlangen-Nürnberg
September 30, 2019

Introduction: Parkinson's Disease (PD)

- Second neuro-degenerative disorder worldwide.
- 6.000.000 Parkinson's patients around the world.
- Neurologists evaluated PD according to MDS-UPDRS-III scale (Goetz et al. 2008).



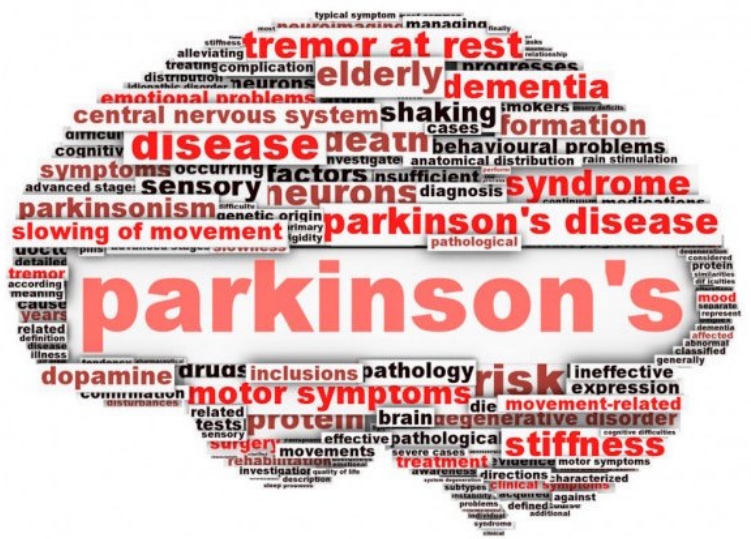
Motor impairments

- Bradykinesia
- Rigidity
- Resting tremor
- Micrographia
- Dysartrhia

Introduction: Parkinson's Disease (PD)

Non-motor symptoms

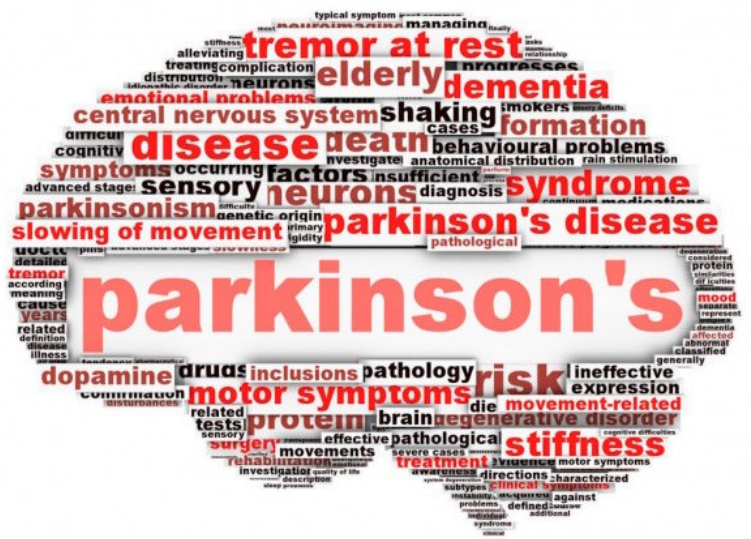
- Sleep disturbances.
- Depression.
- Cognitive impairments.
- Communication disorders.



Introduction: Parkinson's Disease (PD)

Communication and Language impairments

- Deficits in grammar production.
- Less use of action verbs.
- Low information context.
- Simple syntax.
- Differences in sentence length, number of propositions, and grammatical complexity.



Introduction: Hypothesis and Aims

Hyphotesis:

We believe that using NLP methods can also capture the effect of language impairments that affect the communication capabilities in PD, and also to detect the presence of the disease.

Introduction: Hypothesis and Aims

Hyphotesis:

We believe that using NLP methods can capture the effect of language impairments that affect the communication capabilities of PD patients, and detect the presence of the disease.

Aims:

- To model components related to communication deficits in PD using verbal information.
- To analyze the suitability of NLP methods to discriminate PD vs. Healthy Control (HC) subjects.

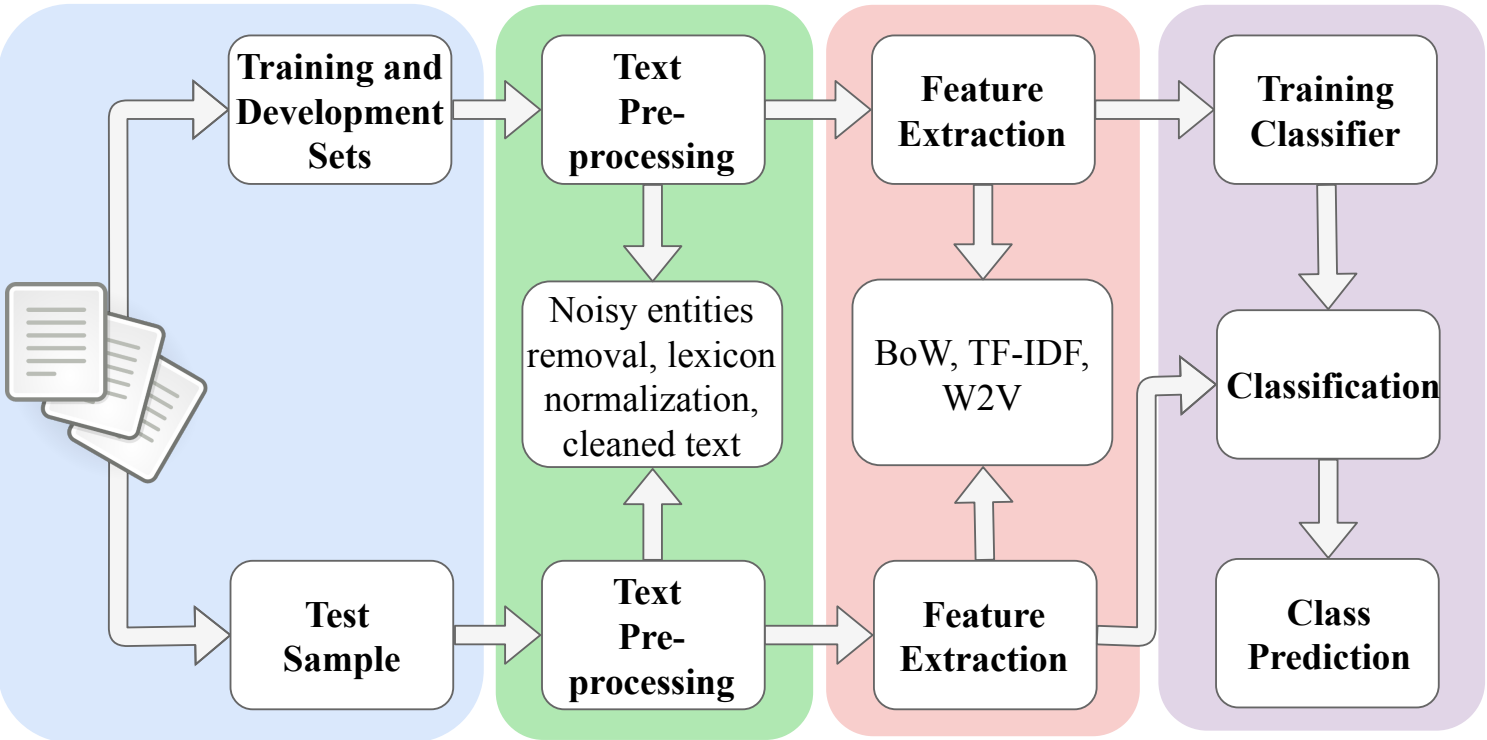
Database

Table: General information of the subjects. Time since diagnosis, age and education are given in years.

	PD patients	HC subjects
Gender [F/M]	25/25	25/25
Age [F/M]	60.7(7.3)/61.3(11.7)	61.4(7.1)/60.5(11.6)
Education [F/M]	11.5(4.1)/10.9(4.5)	11.5(5.2)/10.6(4.4)
Time since diagnosis [F/M]	12.6(11.5)/8.7(5.8)	
MDS–UPDRS–III [F/M]	37.6(14.0)/37.8(22.1)	

- The task consisted on asking the participants to talk about their daily routines
- Average duration of the monologues: 48$\pm$29 seconds for the patients and 45$\pm$24 for the healthy subjects.

Methods: Methodology



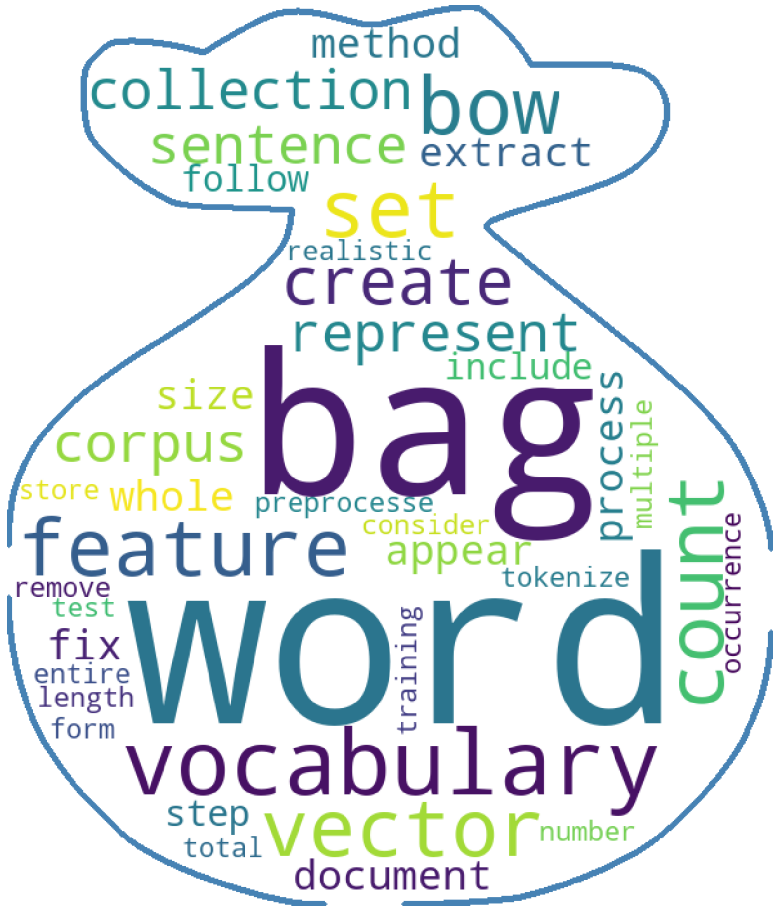
Methods: Pre-processing

The data is cleaned and standardized, making it noise–free and ready for analysis.

Methods: Bag of Words-BoW

Collection of words into a feature vector.

1. The sentences are represented as a collection of words.
2. Vocabulary $\rightarrow$ 1182 words.
3. The words of the transcripts are counted and stored as the feature vector.



Methods: Term Frequency-Inverse Document Frequency–TF-IDF

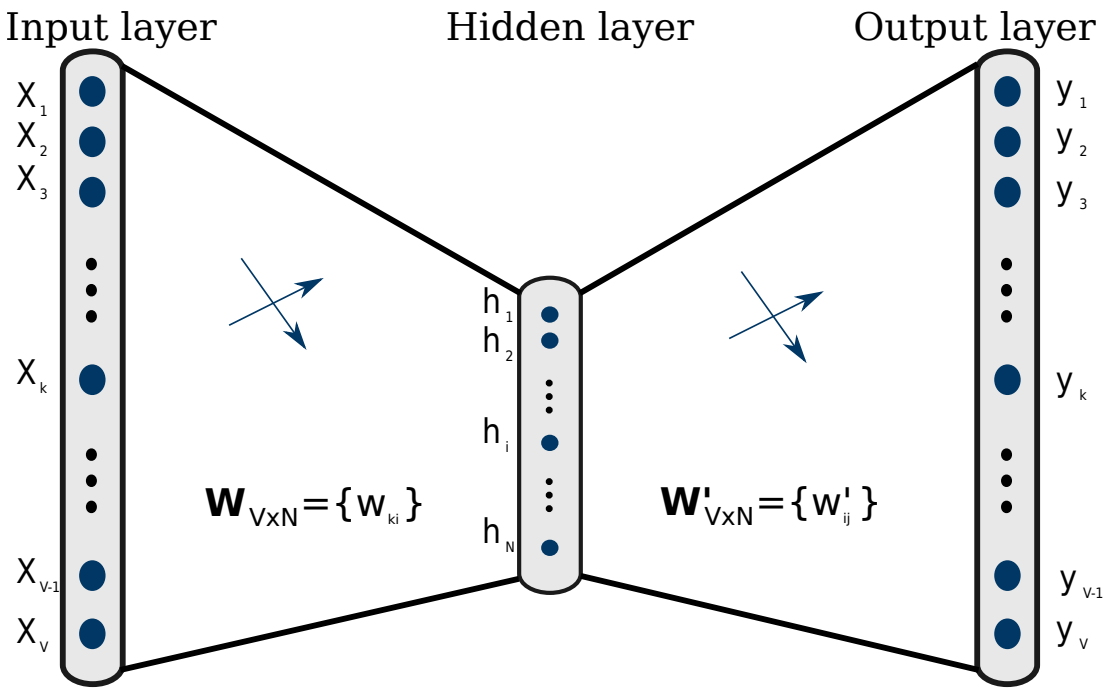
- TF: gives the relative frequency of a specific word.
- IDF: the frequency of occurrence of the word in the collection of documents.
- TF-IDF features aims to model the vocabulary of the patients, and the relevance of the word they use in their transcripts.
- TF-IDF is given for the word $W_{i,j}$ by:

$$W_{i,j} = \mathrm{TF}_{i,j} \log\left(\frac{N}{d_{f_i}}\right)$$

$\mathrm{TF}_{i,j}$: the number of occurrences of the term i in the document j.
d_{f_i}: the number of documents containing i.
N: the total number of documents.

Methods: Word2Vec-W2V

- A Neural Network with one hidden layer.
- Input $\rightarrow$ One-hot-Encoding representation of the words.
- Activations of the hidden layer are the “word vectors".



Methods: Word2Vec-W2V

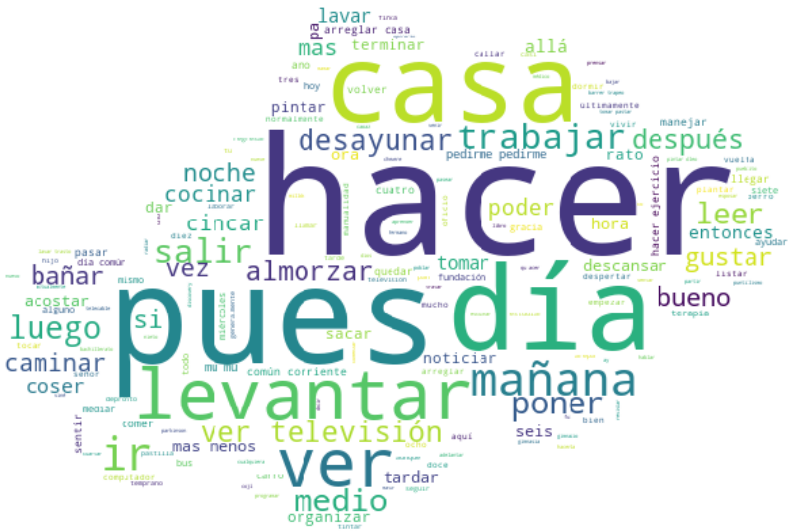
- The model was trained with a continuos bag of words (CBOW) architecture.
- Trained using the Spanish WikiCorpus, which contains 120 millions of words.
- The model considered a window size of 7 words to model the temporal context.
- Dimension of the word vectors was set to 100.
- Statistical functionals were computed for the transcript of each user: average, standard deviation, skewness, and kurtosis.

Methods: Classification

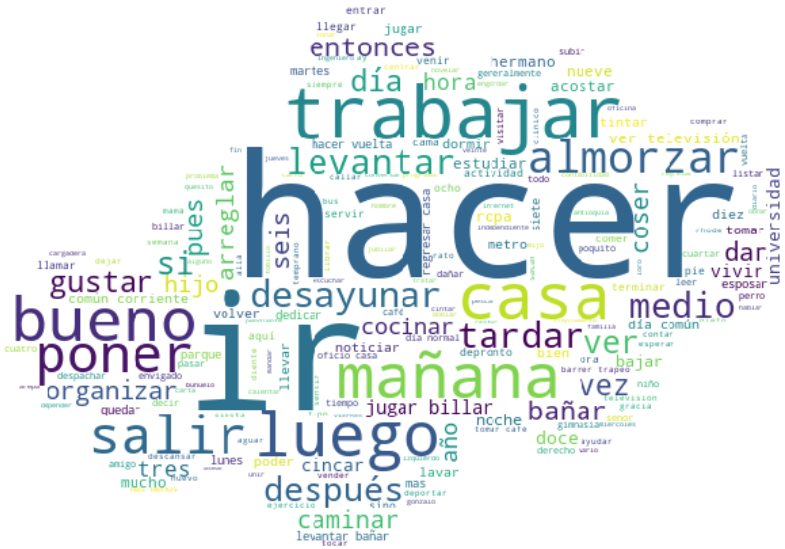
- Two classifiers are considered: A soft margin Support Vector Machine (SVM) with Gaussian kernel, and a Random Forest (RF).
- **Validation:** A ten-fold cross-validation scheme was implemented.
- An early fusion strategy was implemented to combine the different feature sets.

Results

A



B



Word cloud representation: A) PD patient. B) HC subject.

Results

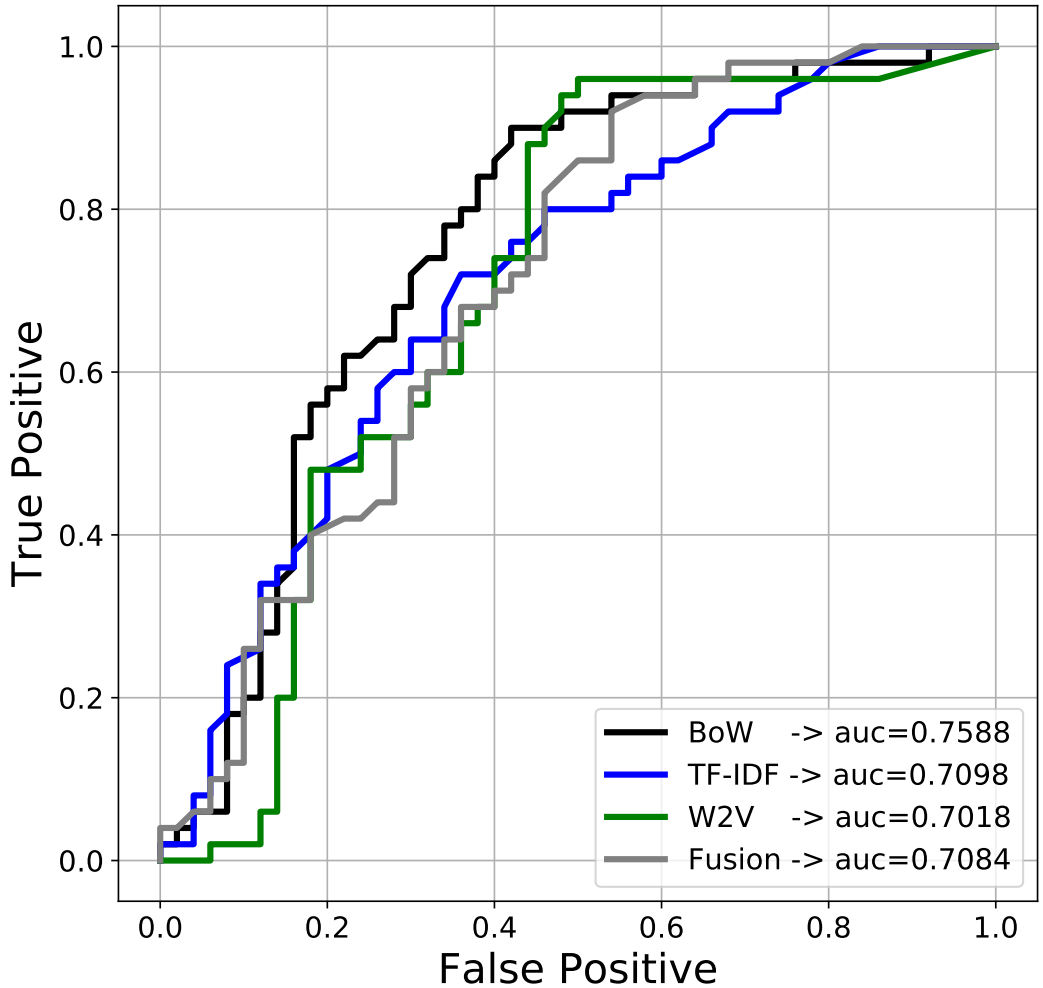
Table: Classification results.

Features	RBF-SVM				RF			
	Acc(%)	Sens(%)	Spe(%)	AUC	Acc(%)	Sens(%)	Spe(%)	AUC
BoW	62.0	70.0	54.0	0.60	**70.0**	**74.0**	**66.0**	**0.76**
TF–IDF	58.0	58.0	56.0	0.60	67.0	68.0	66.0	0.71
W2V	**72.0**	**92.0**	**52.0**	**0.66**	67.0	74.0	60.0	0.71
Fusion	60.0	62.0	58.0	0.62	66.0	68.0	64.0	0.71

Notes: **Acc**: accuracy. **Sens**: sensitivity. **Spe**: specificity. **AUC**: Area under the ROC curve.

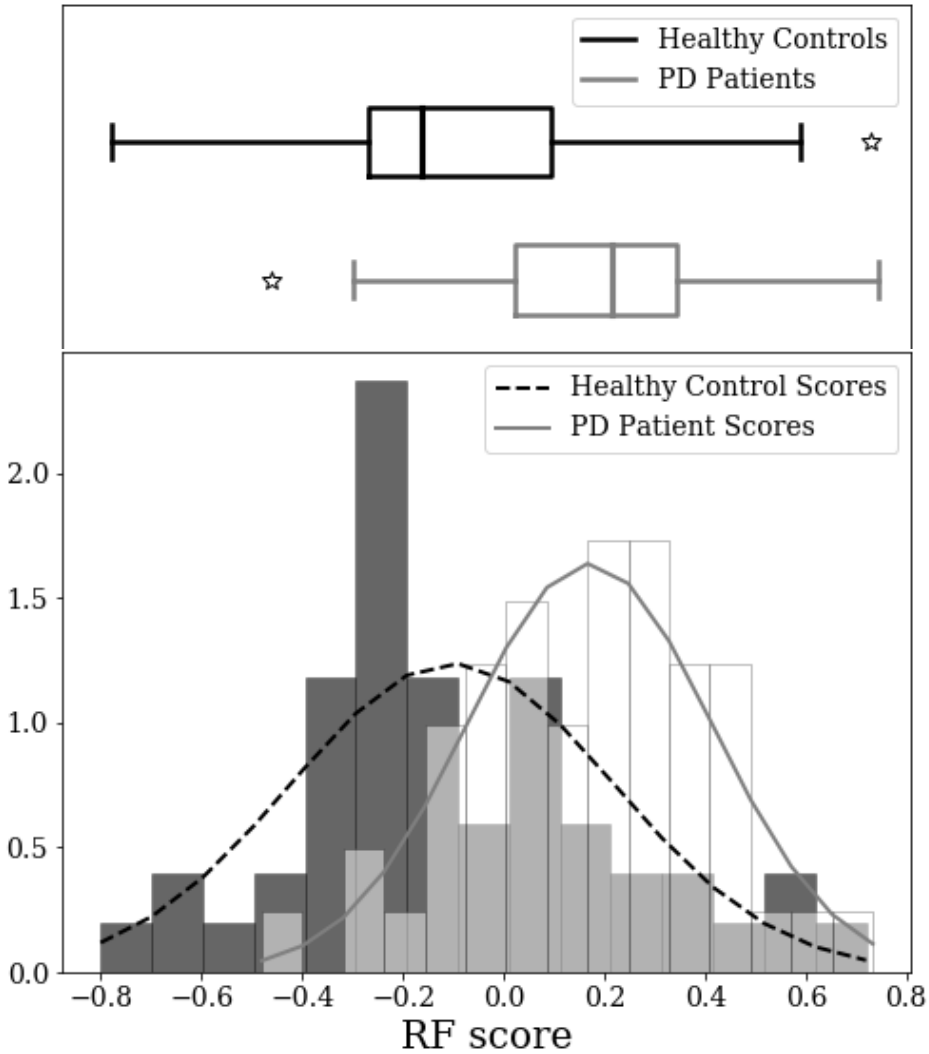
- PD patients are better discriminated in most of the cases.
- The fusion strategy did not improve the results indicating that the considered features are not complementary.
- Further research is required to find an optimal strategy to merge such information.

Results



ROC curves for the different feature sets.

Results



Scores obtained for the BoW feature set.

Conclusion

- Several NLP techniques were considered in this paper to discriminate between HC subjects and PD patients.
- The proposed approach allows the study of different communication disorders that cannot be observed in motor activities.
- PD patients do mainly passive activities like reading, thinking, and taking their medication, while HC subjects do more active activities.
- The results suggest that there is information that reflects language impairments in PD patients.

Conclusion

- **Limitation:** the task performed by the participants might not reflect properly the communication deficits of PD patients, but the difference between the daily routine performed by the patients and the HC subjects.
- Our team is currently collecting more recordings with the aim to evaluate the suitability of other tasks.
- Further experiments will explore more robust word embedding methods such as ELMo or BERT to improve the performance of the system.
- Fusion of acoustic and language information will be implemented.
- Evaluation of specific non-motor impairments of PD patients will be addressed in further experiments: depression, anxiety, among others.

Thank you for your attention. Questions?

Camilo Vasquez
Pattern Recognition Lab, Department of Computer Science,
Friedrich-Alexander University Erlangen-Nurenberg,
Erlangen, Germany
juan.vasquez@fau.de

TAPAS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie grant agreement No 766287. This project was also funded by CODI at UdeA grant # PRG2017-15530.

References I

Goetz, C.G. et al. (2008). “Movement Disorder Society-sponsored revision of the Unified Parkinson’s Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results”. In: *Movement Disorders* 23.15, pp. 2129–2170.

Natural language analysis to detect Parkinson's disease

Paula Andrea Pérez-Toro[1], **Juan Camilo Vásquez-Correa**[1,2], Martin Strauss[2], Juan Rafael Orozco-Arroyave[1,2], and
Elmar Nöth[2]
[1] Faculty of Engineering, University of Antioquia, Medellin, Colombia
[2] Pattern Recognition Lab, Friedrich-Alexander University of Erlangen-Nürnberg
September 30, 2019