Tampere University

Hanna Ylinen

AUTOMATIC ASSESSMENT OF PARKINSON’S DISEASE USING SPONTANEOUS SPEECH

Faculty of Information Technology and Communication Sciences
Bachelor’s thesis
March 2023

TIIVISTELMÄ

Hanna Ylinen: Parkinsonin taudin automaattinen arviointi spontaanista puheesta
Kandidaatin tutkielma
Tampereen yliopisto
Tietotekniikka
Tammikuu 2023

Parkinsonin tauti on neuroneja tuhoava pitkäaikaissairaus, jonka oireiden kirjoon kuuluu heikentynyt puheentuottaminen. Puheenlaadun heikentyminen voidaan havaita digitaalisen signaalinkäsittelyn avulla, sillä puhesignaali sisältää myös kielellisen informaation ulkopuolista eli paralingvistisiä tietoa. Tässä työssä tunnistetaan Parkinsonin tautia puheesta koneoppimismalleilla seuraten tyypillisen paralingvistisen puheenkäsittelyn tutkimuksen vaiheita. Työn tavoitteena on arvioida, miten erilaiset piirreirroitukset sekä koneoppimismallit kykenevät tunnistamaan Parkinsonin tautia spontaanista puheesta.

Työn kirjallisuustutkimusosassa esitellään paralingvistisen puhesignaalin käsittelyn vaiheet ja tarkastellaan vastaavia tehtyjä tutkimuksia. Vastaavat tutkimukset osoittavat, että Parkinsonin tautia sairastavilla on äänisignaaleissa tunnistettavia ominaisuuksia, joita voidaan käyttää taudin arvioimiseen. Tutkimuksissa on hyödynnetty monenlaisia erilaisia piirrejoukkoja sekä koneoppimismalleja.

Tämän työn tutkimuksessa käytetään kahta eri piirrejoukkoa piirreirroitukseen eli Mel-kepstrikertoimia sekä niin sanottuja eGeMAPS-piirteitä, joiden avulla irrotetaan hyödyllistä informaatiota äänisignaaleista koneoppimismalleja varten. Piirteitä käytetään syötteenä kolmeen eri koneoppimismalliin, jotka ovat tukivektorikone, satunnaismetsä sekä konvoluutioneuroverkko. Koneoppimismallien avulla tunnistetaan Parkinsonin tautia PC-GITA-aineiston monologipuhemateriaalista. Käytetyssä aineistossa on noin minuutin pituinen spontaani puhenäyte sadalta eri henkilöltä, joista puolet olivat terveitä ja puolella oli todettu Parkinsonin tauti aineiston keräysvaiheessa.

Työn tulokset laskettiin käyttäen puhujakohtaista ristiinvalidointi-menetelmää, jossa jokainen puhuja toimii kerrallaan koneoppimismallin testidatana ja loput puhujat opetusdatana. Lopullinen mallin tarkkuus saatiin laskemalla puhujien lukumäärän eli sadan yksittäisen mallin tunnistustarkkuuksien keskiarvo. Tämän menetelmän avulla vähennettiin yksittäisen merkittävästi muusta datasta poikkeavan testidatan vaikutusta kokonaistulokseen.

Tämän työn tutkimuksen tulokset osoittavat, että Parkinsonin tautia voidaan tunnistaa puheesta hyödyntäen koneoppimismenetelmiä. Konvoluutioneuroverkko tuotti parhaimman tarkkuuden Mel-kepstri-kertoimilla 67,40 %:n luokittelutarkkuudella, kun tehtävänä oli erotella Parkinsonin tautia sairastavat puhujat terveistä, kun taas satunnaismetsä tuotti eGeMAPS-piirteiden avulla 75,00 %:n tarkkuuden. Tarkkuuksien alhaista lukua selittää spontaanin puheen monimutkaisuus, aineiston pieni koko sekä valitut koneoppimismenetelmät.

Avainsanat: Koneoppiminen, signaalinkäsittely, parakielellinen puheenkäsittely, Parkinsonin tauti, satunnaismetsä, konvoluutioneuroverkko

ABSTRACT

Hanna Ylinen: Automatic assessment of Parkinson’s disease using spontaneous speech
Batchelor’s thesis
Tampere University
Information technology
January 2023

Parkinson's disease is a neurodegenerative disease with a range of symptoms, including speech impairments. These can be detected with digital signal processing, since speech signals carry paralinguistic information, which means information beyond linguistic information. In this work, Parkinson's disease is being recognized from speech signals using machine learning methods while following the steps of a typical research of paralinguistic speech processing. The main goal of this work is to evaluate how different feature extractions and machine learning models are capable of recognizing Parkinson's disease from spontaneous speech.

The literature research part of this work presents the stages of a typical paralinguistic speech processing pipeline and evaluates related studies and research. Based on the related studies, people with Parkinson's disease have recognizable features in their speech signals which can be used to assess the disease. Additionally, multitude of feature sets and classification models have been applied in the studies.

In the research of this work, for feature extraction MFCCs and eGeMAPS features are used to extract useful information from audio signals. The features work as an input to three different machine learning models used in this study: support vector machine, random forest, and convolutional neural network. These machine learning models are used to identify Parkinson's disease from the monologues of PC-GITA corpus. The data from PC-GITA used in this study consists of around a minute long spontaneous speeches from a hundred people of healthy speaker and people with diagnosed Parkinson’s disease.

The results of this work were evaluated with a speaker-independent cross-validation method, in which each speaker acts as test data for the machine learning model and the remaining speakers as the training data. The final accuracy of the model was obtained by calculating the average accuracy of all folds of one hundred speakers.

The results of this work indicate that Parkinson's disease can be recognized from speech using machine learning methods. Convolutional neural network produced the best accuracy for MFCCs features with 67.40% classification accuracy (Parkinson’s patient versus healthy talker), while random forest produced 75.00% accuracy for eGeMAPS features. The low accuracies are explained by the complexity of spontaneous speech and the chosen machine learning methods.

Keywords: Machine learning, signal processing, paralinguistic speech processing, Parkinson’s disease, support vector machine, random forest, convolutional neural network

CONTENTS

LIST OF SYMBOLS AND ABBREVIATIONS

CNN	Convolutional Neural Network
eGeMAPS	The extended Geneva Minimalistic Acoustic Parameter Set
DCT	Discrete Cosine Transform
GeMAPS	Geneva Minimalistic Acoustic Parameter Set
LLD	Low-Lever Descriptors
LOSO	Leave-One-Speaker-Out method
MFCC	Mel-Frequency Cepstral Coefficients
PD	Parkinson’s disease
PSP	Paralinguistic Speech Processing
ReLU	Rectified Linear Unit
RF	Random Forest
STFT	Short-Time Fourier transform
SVM	Support Vector Machine
UPDRS	Unified Parkinson’s Disease Rating Scale

1. INTRODUCTION

Parkinson’s disease (PD) is a neurodegenerative disorder which is common in people over 60 years old [1]. The disease is caused by a loss of dopaminergic neurons and results in number of symptoms which usually worsen with time. The variety of motor and non-motor symptoms of PD include speech impairments which have been a studied subject in the fields of signal processing and machine learning in recent years [2]. A regular aim of these studies has been the detection of the disease from speech. Consequently, early detection of the disease is crucial because then a medical process can be started to ease with the symptoms and improve the quality of life.

PD is often diagnosed and measured with in 1980s published and later in 2008 improved Unified Parkinson’s Disease Rating Scale (UPDRS) [3]. The diagnose is conducted by a series of questions which are performed by professionals and answered by the patients or their caregivers. The scale provides reliable results but in the field of machine learning the interest is in more objective assessment.

PD patients often develop hypokinetic dysarthria which can be seen as monopitch, reduced stress, speech dysfluencies and inappropriate silences during talking [4]. Speech impairments can be due to motor symptoms such as the rigidity of vocal folds, but it’s not entirely understood which speech problems are based on motor and which are from cognitive symptoms [5]. On the whole, speech carries a lot of information, and early-stage symptoms which can be hard to detect by humans can be detected by pattern recognition algorithms [6].

While the earlier work has shown approvable results in automatic assessment of PD from speech, the accuracy of the assessment declines when speaking tasks get more complex. It is easier to detect speech impairments from similar short and repetitive speaking tasks, but not from unscripted monologues. In this work, the aim is to study how PD can be detected and assessed from speech signals of spontaneous speech with machine learning methods. In detail, a typical speech analysis method is followed to evaluate how different feature extraction and machine learning models assess PD from speech signals.

This thesis is structured as follows. Firstly, chapter two introduces paralinguistic speech processing which is the category this study belongs to. Moreover, chapter 2 encompasses typical pipeline of speech analysis and related studies. In the chapter 3, the main methods are described and in this study two feature extractions and three different classifying models are used. The fourth chapter describes in detail the database and setup used in this study and the following chapter shows the results. Lastly, the final chapter summarizes the results and provides a conclusion.

2. BACKGROUND

In this chapter the topic of speech processing is introduced, and its standard characteristics and methods are examined. Furthermore, studies related to the aims of the present study are discussed.

2.1 Paralinguistic speech processing

Machine learning is part of artificial intelligence and holds an essential role in modern world. The idea of machine learning is based on finding solutions, also known as functions, by learning from input data to tune the parameters of an algorithm. In practice machine learning is used to identify patterns in the data and use them to classify or predict desired outputs related to the data. In fact, machine learning is often utilized together with digital signal processing in which the input data can be signals, audio and time series [7].

Paralinguistic speech processing (PSP) is part of the digital signal processing with the aim of detecting different paralinguistic factors of speech. These factors signify different types of information conveyed by speech beyond its linguistic content. Utilization of paralinguistic factors in various speech processing analysis is possible since speech signals contain cognitive and neurophysiological information of the speaker [8]. Furthermore, in PSP the goal is to create automatic systems to detect and analyze these factors. In the context of PD, for instance, fluidity and clarity of speech production could be such factors that are affected by PD [9].

Paralinguistic factors are often divided into two basic categories: traits and states [9]. Traits are long-term and hardly changing characteristics of a speaker, such as biological or cultural characteristics or personality traits. On the contrary, states are short-term factors such as emotions. Both categories are useful in the case of an automatic assessment of neurodegenerative diseases, such as PD. In addition, longer speech recordings reflect the multitude of symptoms and characteristics of speech better than shorter ones. Additionally, involuntary speech changes which are caused by PD are more likely to occur in longer speech recordings [8].

2.2 Pipeline of a paralinguistic speech processing system

Machine learning consists of two distinguished learning types: supervised and unsupervised learning. In supervised learning the input data is marked with labels and the goal

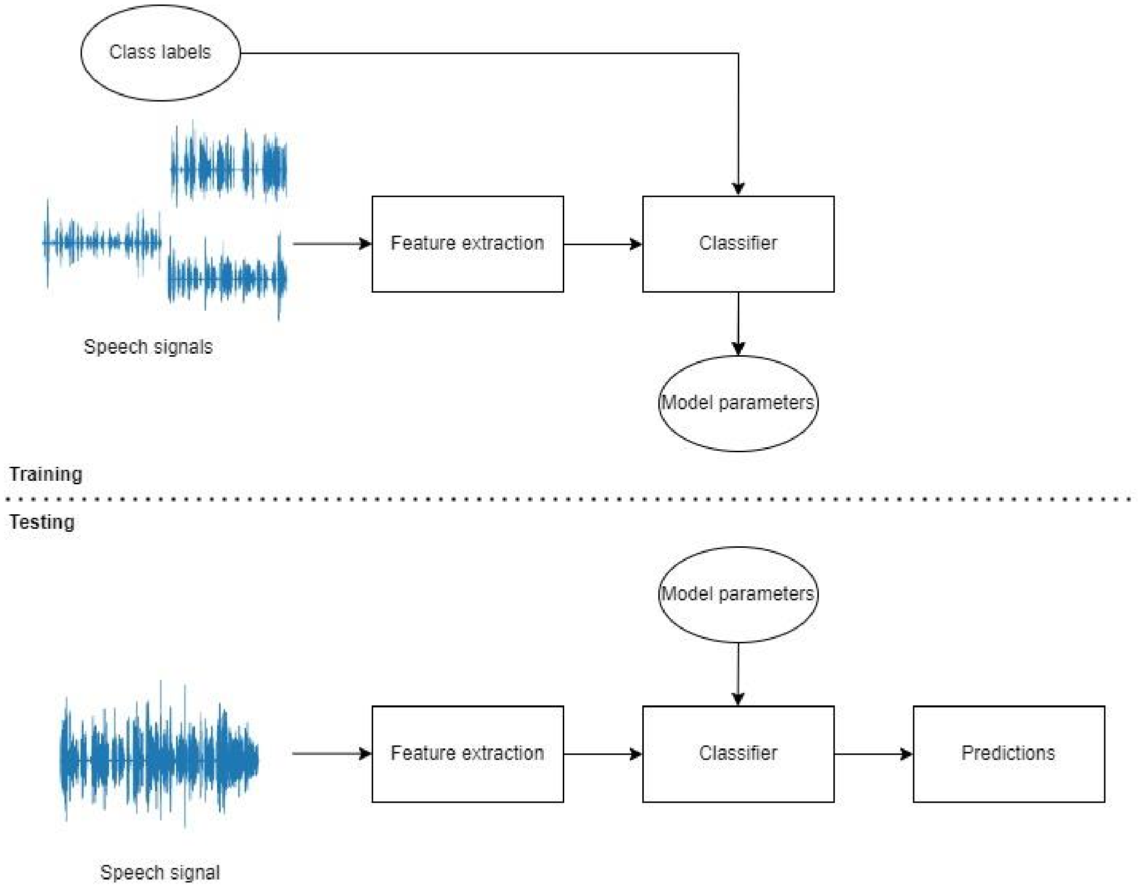

Figure 2.1. *A typical machine learning pipeline with training and testing.*

is often either classification or regression task. In classification, the goal is to identify the labels of previously unseen data samples with a taught model [7]. The model is trained with a labeled dataset and its performance is evaluated with another dataset called test set. Performance of the model is estimated by comparing the predicted labels of the test set with the correct ones.

PSP systems often follows a typical supervised ML pipeline which is visualized in figure 2.1 [8]. As seen on the figure, the training process starts with speech signals as inputs to the pipeline. The speech signals often have a length of multiple second and from them short-time feature vectors are then extracted on the next phase. In fact, in PSP systems, it is important to keep on track which feature vectors belong to each speech signal, so the classification can be done over time. On the third phase, the features with their labels are used to train a classifier. As a result, model parameters are acquired.

Similarly to the training, during the testing process features are extracted from the input as seen in the figure 2.1. In this case the model is tested with only one input signal without labels. Additionally, in contrast to the training process, the classifier now works without labels. The classifier predicts the label for the testing data based on the information carried by the model parameters and properties of the input samples.

A standard requirement for a good performance in machine learning is that the training data should be representative of the data encountered in practical deployment of the system, as represented by the test set in the system development process [9]. Therefore, a typical solution for the training process is to divide the original data into training and testing sets. In the case of small datasets or datasets with unevenly balanced labels, a cross-validation technique is often utilized in the process [10]. Cross-validation can minimize the effect of biased small-scaled test data, since the full dataset can be used for testing by dividing the data multiple times into training and testing sets in different ways.

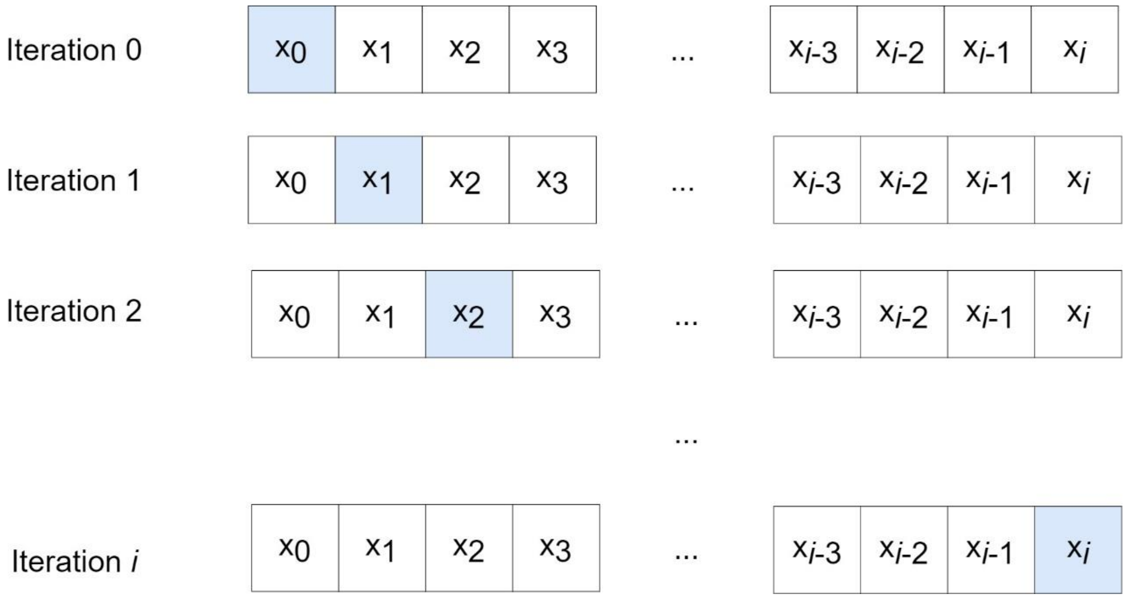
Iteration 0
x_0 x_1 x_2 x_3 ... x_{i-3} x_{i-2} x_{i-1} x_i
Iteration 1
x_0 x_1 x_2 x_3 ... x_{i-3} x_{i-2} x_{i-1} x_i
Iteration 2
x_0 x_1 x_2 x_3 ... x_{i-3} x_{i-2} x_{i-1} x_i
...
Iteration i
x_0 x_1 x_2 x_3 ... x_{i-3} x_{i-2} x_{i-1} x_i

Figure 2.2. *An example of cross-validation with* $\boldsymbol{x}$ *symbolizing the speakers. In this case, there are* $\boldsymbol{i}$ *number of speakers and iterations. In each iteration the speakers who belong to the training data are marked with white and the test speakers are marked with blue.*

To be precise, an application of cross-validation called Leave-One-Subject-Out (LOSO) is often used in speech processing [11]. The method is visualized in figure 2.2 with i as the total number of speakers. The LOSO cross-validation refers to having one speaker as a testing data and other speakers for the training data for a machine learning model [10]. The training and testing process is repeated such that all individual speakers end up being tested with separately trained models. For instance, in the case of 100 speakers there are 100 folds in which each speaker is used only once as the test data. Classification accuracy is calculated from the correctly labelled test samples divided by the total number of them. The final result is obtained as the arithmetic mean of the classification accuracies of all the cross-validations.

2.3 Related work

There have been numerous studies of the usage of machine learning to identify PD and other similar dysarthria or speech impairments from speech with the help of speech databases made for these studies. The scope of these databases varies from short utterances such as vowels to long monologues. Additionally, different combinations of features and machine learning models have proven beneficial.

For shorter utterances, Karan et al. [12] used recordings from PC-GITA corpus [13]. The corpus consists of multiple different speech tasks including short-term utterances, reading a text out loud and long monologues. From the database, sustained vowels and isolated words were chosen in this study. The researchers used non-negative matrix factorization to three different types of features. The features were extracted from the parts of the speech signals which contained discontinuities and changes of speech. For classification, support-vector machine (SVM) was used with speaker-independent k-fold method. One of the three features used, called Mel-Frequency Cepstral Coefficients (MFCC), resulted on average 56% classification accuracy on vowels and on average 60% on words.

Similarly, Syed et al. [14] detected PD using all the speech task recordings in PC-GITA database. They extracted features with the Extended Geneva Acoustic Minimalistic Feature Set (eGeMAPS) which yielded in 72% classification accuracy on reading a text and 80% for monologue. The results were obtained by a mean accuracy of cross-validation with logistic regression classifier. In addition, Narendra and Alku [11] utilized openSMILE-toolkit in order to combine glottal parameters, such as abrupt vibration changes in vocal folds, with more traditional acoustic features. In their research they detected dysarthria using non-words, words and whole sentences from TORGO and universal access speech databases. The results were calculated from the classification accuracies with LOSO cross-validation. They tested various combinations of the features with SVM classifier. As an example, with INTERSPEECH feature set of openSMILE they got 69% classification accuracy as a result.

Vásquez-Correa et al. [1] classified PD using PC-GITA corpus with two different methods. The first method, described as the baseline, used SVM and the feature extraction focused on the motor skills of the onset and offset of spoken words. A segment was taken around these to form transfer speech segments. For the baseline, 12 MFCCs were taken with their first and second derivatives and the classifying was done with 10-fold cross-validation strategy with speaker independent training and testing sets. Their second method used convolutional neural network (CNN) with time-frequency representations as inputs. Based on majority voting strategy either repeated utterances, a text or a monologue were used from the PC-GITA database to teach the CNN model.

In the study of Vásquez-Correa et al. [1] not only the Spanish database was used but also German and Czech. Researchers used a transfer learning method with the CNN and with it they refer to training CNN with one language and using that as a baseline model to train other language. Therefore, the biggest model training was conducted with the first language and the latter language was used to fine-tune the model parameters. The best accuracy for both SVM and CNN without transfer learning method was acquired by Spanish language with accuracies of 73.7% and 71.0% respectfully. Furthermore, using Spanish as the baseline provided the best accuracy for every classification with transfer learning method. For instance, the Spanish baseline improved the accuracy of German language from 63.1% of baseline CNN to 77.3% with the transfer learning method.

The focus of the aforementioned studies relies on motor signs. For comparison, the study of researchers Pérez-Toro et al. [15] focused on the linguistic level of language by analyzing transcriptions of the monologue of PC-GITA database. For methodological approach, they used word-embeddings which is a Natural Language Processing method. The results aligned with the hypothesis that PD also affects non-motor language impairments with the best accuracy of 72% for World2Vec method in which similar words are closely located vectors in a multidimensional space. Additionally, the researchers noted that the topic of the PC-GITA monologues being everyday activities might help to separate the speakers with PD from healthy clients based on the fact that they are much more non-active in their free-time and might had made it harder to find suitable words for the monologue.

3. METHODS

This chapter gives an in depth-analysis of the methods used in this study. First, section 3.1 introduces two different feature sets for a feature extraction which is the second part of the pipeline in figure 2.1. Second, three different machine learning models are introduced in section 3.2. These models are used for classification which is the third phase in the pipeline in figure 2.1. Both feature sets are used separately with all three classifiers.

3.1 Features

Feature extraction is a way to obtain the most relevant information out of a signal, given the processing task in hand, and produce a format for input data that is the most useful for the following machine learning pipeline to use and analyze. In this study, two different features are extracted separately.

3.1.1 Mel-Frequency Cepstral Coefficients

There are many different methods to analyze signals and gather the most important information out of them for speech analysis. One popular method for the feature extraction is extraction of Mel-Frequency Cepstral Coefficients (MFCC) [8]. The coefficients are obtained by performing mel-frequency mapping, logarithm and discrete cosine transform (DCT) on a short-term magnitude of a signal.

In the field of signal processing, windowing is a popular method to analyze signals in short parts [16]. In the windowing method, the audio signal is divided into such small fixed sized segments that the signal is quasi-stationary over the given time frame. Windowing is often done with overlapping between segments to minimize discontinuations. Furthermore, to minimize discontinuations and the aliasing of the signal, the signal segments are smoothened with windowing functions, for instance, with a Hann window which is also known as raised cosine.

The calculations of the MFCC can be divided into three phases [17]. For the first phase, windowed speech segments are turned into frequency domain by discrete Fourier transform. This results in short-time Fourier transforms of the windowed segments. This can be visualized by taking a log-spectra of the short-time Fourier transforms across the whole signal which is also known as a spectrogram. A spectrogram of a monologue by a speaker from the corpus of this study is portrayed in figure 3.1. Secondly, the non-

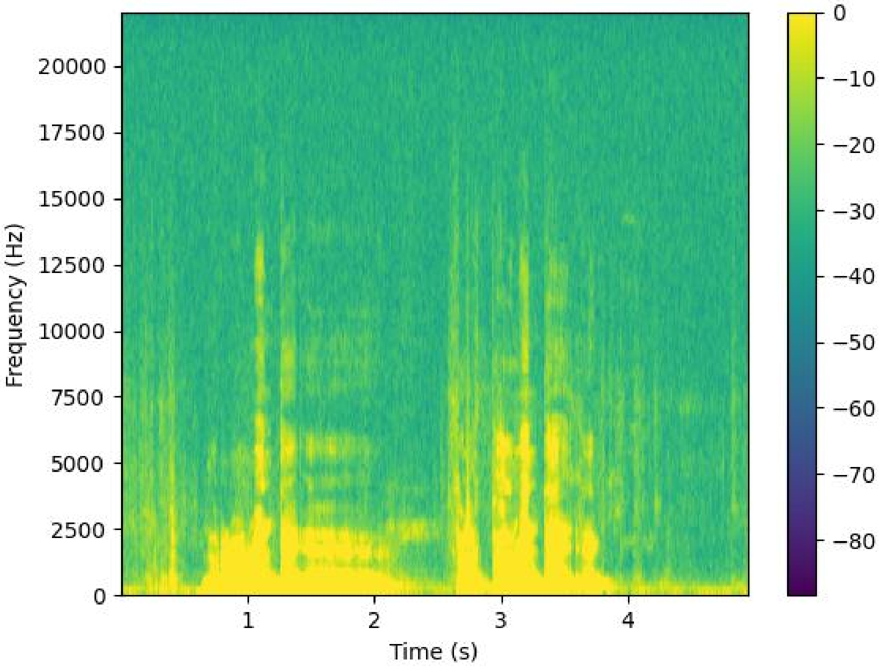

Figure 3.1. *A spectrogram of the first five seconds of a monologue from a male speaker with PD from PC-GITA corpus [13].*

linear frequency resolution of hearing in humans is taken into account by filtering the spectra with a mel-scale filter bank and then taking logarithm of the output. This mel spectrogram is visualized in the figure 3.2. Without mel-scaling, the higher frequencies can have a too large of an effect on the resulting feature vector. Lastly, the cepstral

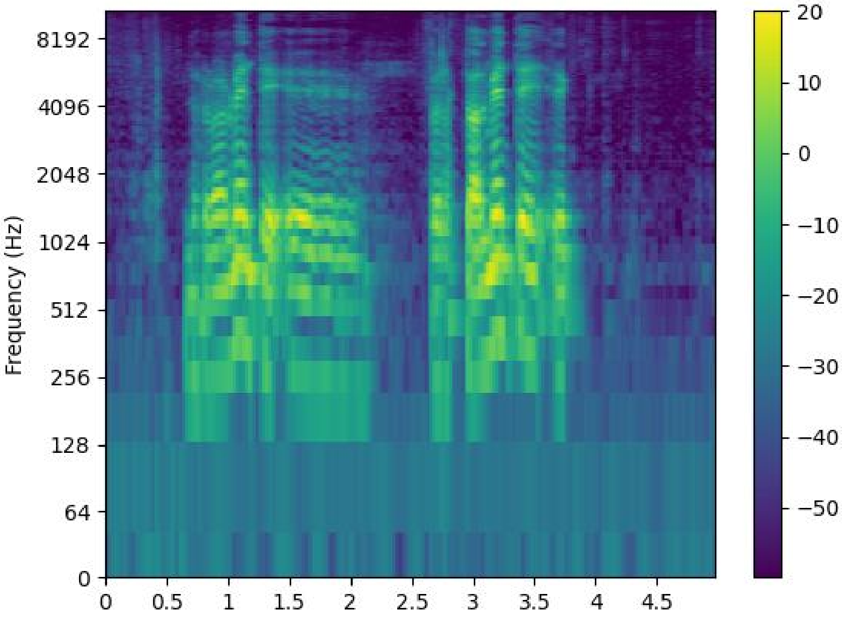

Figure 3.2. *A mel spectrogram of the first five seconds of a monologue from a male speaker with PD from PC-GITA corpus [13].*

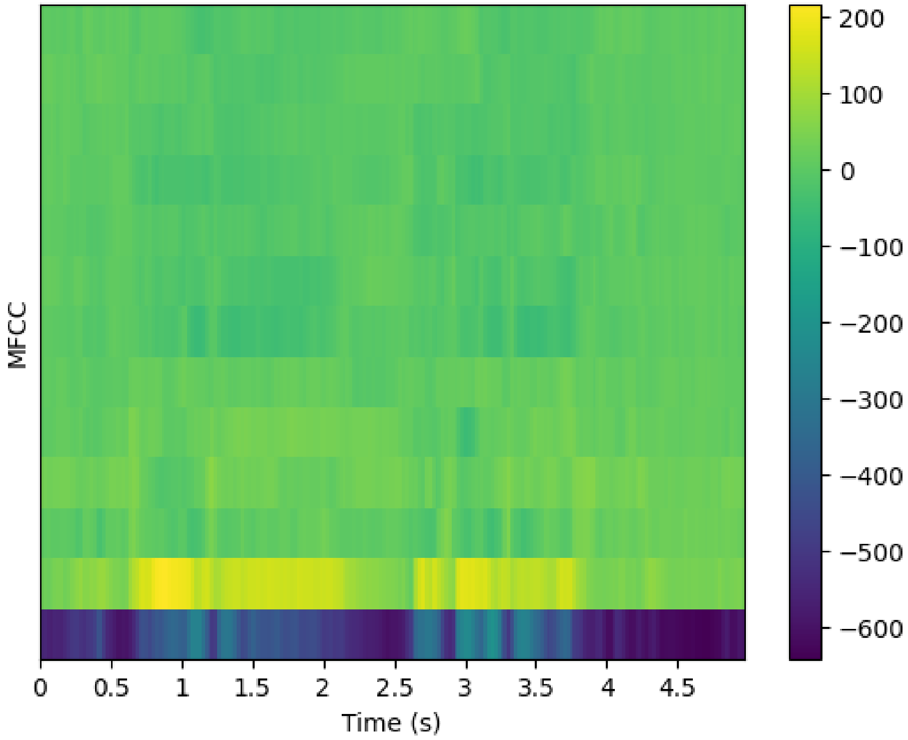

Figure 3.3. *MFCC of the first five seconds of a monologue from a male speaker with PD from PC-GITA corpus [13].*

domain coefficients are created by taking the DCT of the log-mel-spectrum. [8] The resulted coefficients are visualized in figure 3.3. with the same monologue.

For most speech processing tasks, the most relevant information of MFCC lays in the first 10–20 low-level coefficients which are often used for recognition studies. For the most optimal classification phase, the feature vectors are normalized, for instance, with global cepstral mean and variance normalization [18]. To support machine learning tasks, the first and second derivatives of the coefficient can be taken from the low-level coefficients [6]. This results in multitude of different coefficients, for example, with 13 low-level coefficients the number of coefficients is 39 in total.

3.1.2 The extended Geneva Minimalistic Acoustic Parameter Set

The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) was created to be a baseline of feature parameters for different studies based on detecting paralinguistic characteristics. Minimal number of features were chosen to reduce the probability of overfitting and the selection was carried by three main criteria [19]. First, to detect physiological changes in speaking. Secondly, based on the success in earlier studies. Lastly, based on theoretical considerations.

Table 3.3.1. *The minimalistic and extended parameter sets of eGeMAPS [19].*

	Minimalistic Parameter Set	**Extended Parameter Set (includes all parameters from minimalistic version)**
LLD	Pitch Jitter	MFCC 1–4 Spectral flux
	Formant 1, 2, and 3 frequency Formant 1	Formant 2-3 bandwidth
	Shimmer Loudness	
	Harmonics-to-noise ratio (HNR) Alpha Ratio	
	Hammarberg Index Spectral Slope 0-500 Hz and 500-1500 Hz	
	Formant 1, 2, and 3 relative energy Harmonic difference H1-H2 and H1-A3	
Functionals	Arithmetic mean and coefficient of variation of all LLD Hammarberg Index, Spectral Slope 0-500 Hz and 500-1500 Hz and arithmetic mean of Alpha Ratio for unvoiced sections	Arithmetic mean and coefficient of variation of all LLD Arithmetic mean of Spectral flux for unvoiced sections
	8 additional functionals to loudness and pitch 6 temporal features	Arithmetic mean and coefficient of variation of spectral flux and MFCC 1–4 for voiced only regions Equivalent sound level

The baseline of GeMAPS consists of 62 features from which 18 features are low-lever descriptors (LLD), i.e., short-term features, which can be roughly divided into three parameter groups [19]. These groups are related into either frequency, amplitude or spectral. Additionally, functionals are calculated from the LLD across the entire signal of interest together with additional 8 functionals from related to loudness and pitch parameters. These 52 parameters contain only voiced regions, so for the unvoiced segments, four parameters are added. Finally, six temporal features are added to the set, which sums up to a total of 62 parameters for the baseline. All the parameters belonging to the baseline are depicted in table 3.3.1. in more detail.

Furthermore, an expanded version of GeMAPS called the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) is one of the most used feature sets used in social signal processing [11]. It consists of 88 features and 62 of them belong to the aforementioned baseline. The extended version is also defined in detail in table 3.3.1 . The extended parameters are affiliated with cepstral parameters and the final set results in 26 additional parameters. The LLDs are expanded with seven parameters that are either spectral or frequency related. From these parameters functionals are also calculated. The extended set consists also of 11 other descriptors and results in a set of 88 features in total. The LLDs and the statistical functional of eGeMAPS can be extracted with OpenSMILE toolkit [20].

3.2 Classifiers

This section introduces the third part of the pipeline in figure 2.1. Three different machine learning method are used in this study. Each method is tested with both feature sets separately.

3.2.1 Support Vector Machine

One popular supervised learning method often used in pattern recognition is called support vector machine (SVM). SVM can be used to classify both linear and non-linear data [21] into two or more classes. Training samples of SVM are represented as data points in a feature space which dimensions depends on the input data and the chosen version of SVM. In the case of two classes in a three-dimensioned space, SVM separates the space between the data instances with a hyperplane [22]. Therefore, the side in which a test data points lands will determine the class label.

There are multiple solutions to form a hyperplane [23]. Therefore, an ideal solution to the model corresponds to optimization for the best classification accuracy of unknown test data. The advantage of SVM is the fact that the model is improved by maximizing the margin which is the sum of the distances between the hyperplane and the closest data points from each class regarded to the hyperplane. Mathematically a hyperplane in N-dimensional space can be defined as

$$\boldsymbol{w}^T x_i + b, \tag{1}$$

in which $\boldsymbol{w}$ is the feature vector, x_i represents the i^{th} data point and b is the margin [21]. In a perfect classification case, the formula is

$$y_i(\boldsymbol{w}^T x_i + b) > 0, \tag{2}$$

where y_i represents the corresponding i^{th} class label [22]. The optimization of margin is done with the help of support vectors. In the case of two classes, with the equation (2) the data points can be divided into

$$\boldsymbol{w}^T x + b \ \geq 1 \text{ for } y_i = 1, \tag{3}$$

and

$$\boldsymbol{w}^T x + b \ \leq -1 \text{ for } y_i = -1, [22] \tag{4}$$

in which the y_i values of 1 and -1 symbolize the two classes.

However, a perfect margin is not always attainable in practice [23]. Therefore, SVM has a hyperparameter *C* to help to tune the choosing of margin with the help of a measurement for classification error. Thus, SVM is robust to outliers meaning individual outliers don’t modify the model noticeably. Although, in some cases the data can’t be linearly separable, hence it needs to be converted into higher dimension in which the data is linearly separable. This method is called the kernel trick.

3.2.2 Random Forests

Random forests (RF) is an ensemble method for predicting classes in the field of machine learning [23]. The prediction is done by a major class label voting from a collection of decision trees. A decision tree lands on a prediction after multiple condition stages, where each stage makes a decision based on some threshold. These stages, which are called nodes, create a tree-shaped model. The tree starts on a root node and after the internal nodes it ends on leaf nodes which are each linked with a class label.

RF takes labelled feature set as an input, but the tree needs to be built before classifying. Each tree is given a set which has the same dimension as the original feature dataset. The set is formulated from a subset of the original dataset with duplications of some of the features in the subset to acquire the same dimension as the original dataset. During the learning process, random variables from the subset are chosen for splitting. From the most correctly divided split, the node is divided again into two other nodes. Splitting is continued until a chosen threshold of nodes. This method is called the bootstrap aggregation and it is used to avoid overfitting [23]. Decision trees are prone to overfitting [7], which means they model the specifics of the input features in such a detail that the algorithm fails to classify new data when the details differ. In other words, too accurate decision trees can prove faulty. Thus, RF minimizes possible biased or over-fitted decision trees by taking the average of the predictions of numerous individual trees in creation of the final result.

3.2.3 Convolutional Neural Network

Human learning has inspired the structure of a machine learning model called neural network [22]. The biological learning in brains takes place through a nervous system which is formed by neurons which can send signals to other neurons. These neurons and their ability to communicate with each other create a complex network that can process data fast and parallelly to learn and make decisions. These facts have inspired artificial neural networks in machine learning.

Artificial neural networks consist of at least one hidden layer between the input and output layer. In deep learning, the model has multiple hidden layers which together create higher level features. Furthermore, the model itself identifies the most important properties of the input data whereas in more traditional techniques the features and pattern detection are carefully designed by humans [24]. Deep learning has proven to be successful in finding the most useful patterns and it is widely used in PSP.

A hidden layer in artificial neural networks consist of neurons, also known as nodes. Additionally, each node performs a series of operations of three different steps [22]. First, a node has multiple input values which are all the node values from previous layer. The first step is to multiply each input value with a certain weight value. The second step is a summation junction in which all the weighted inputs are summed together. Lastly, with the help of the result value from the summation junction an activation function returns an output based on a certain threshold. In fact, a popularly used non-linear activation function called Rectified Linear Unit (ReLU) function return the same value as it received unless the value is negative. In the case of negative summation value, the output of the activation function is zero.

Convolutional Neural Network (CNN) is specific type of deep learning [24]. The model can process well one-dimensioned and high dimensioned data, such as two-dimensional audio spectrograms. The main benefit of CNN models is its ability to recognize patterns regardless of size, location or translation of the pattern in the input form. Therefore, there is no need to create multiple pattern recognition models, one for each different pattern.

A typical learning process is achieved from plethora of stages [24]. In these stages, typically multiple pairs of convolutional or pooling layers are stacked and the network ends with a fully connected output layer. First, in convolutional layers, a high-dimensional filter, also known as a kernel, glides through the data from previous layer. The values of the filter are then weighted and summed together, and this method is called a filter bank. This results in a feature map. The convolutional layer consists of multiple feature maps with each having an own filter bank. Therefore, this method allows the model to detect local similarities and this information is passed on to the next layer.

Moreover, pooling layers compress the representations in time or space by outputting only the highest activation within the pooling window to the next layer, thus also reducing computational burden of the following layers [24]. Additionally, pooling layers have different types such as maximizing or averaging pooling. The main logic behind the CNN algorithm is that high-level attributes are created from aggregating lower-level features together in a non-linear matter. The results of a classification task are obtained from a fully connected layer with an activation function such as softmax.

4. EXPERIMENTS

The experimental setup of this study is described in this chapter. The experiment follows the PSP pipeline illustrated in figure 2.1. Specifically, the database used in this study is introduced in detail in section 4.1 and the setup is described in section 4.2. The aforementioned feature sets and classifiers in chapter 3 are utilized in this study.

4.1 Data

The data used in this study is from PC-GITA corpus which was published by Orozco-Arroyave et al. in the article “New Spanish speech corpus database for the analysis of people suffering from Parkinson’s disease” in 2014 [13]. The main motive for the creation of the corpus was research purposes of PD. In fact, PC-GITA was the pioneer corpus in Spanish language for PD related research.

PC-GITA corpus [13] consists of 100 native Colombian Spanish speakers from whom half has been diagnosed with PD and the other half is healthy controls (HC). Both halves have 25 women and 25 men. The metadata of the corpus is depicted in table 4.1. As seen from the table 4.1, the average age is around 61 years for both PD and HC speakers. Based on these aforementioned facts the database is reasonably well balanced in terms of demographic properties.

Furthermore, the corpus [13] includes the UPDRS stage and UPDRS speech level of speakers which are represented in table 4.1. The UPDRS scale is from zero to 100 and the speech impairment-focused UPDRS-speech scale ranges from zero to three. The ratings were obtained from professional neurologists. On the contrary to the gender and age, the speech impairment stage of the corpus is unbalanced since most of the speakers with PD have stage one of speech impairment and only four out of all 50 speakers have the stage three in speech impairment. In addition, seven speakers with PD have the stage zero which indicates a lack of speech impairment.

The audio recordings of the PC-GITA corpus [13] were captured in a studio environment and sampled at 44100 Hz with a 16-bit resolution. To be precise, the database has multiple different speech task recordings, and the tasks were originally chosen to represent

***Table 4.1** Metadata of the speakers in PC-GITA corpus [13].*

Women				**Men**			
PD			**HC**	**PD**			**HC**
Age	**UPDRS**	**UPDRS-speech**	**Age**	**Age**	**UPDRS**	**UPDRS-speech**	**Age**
72	19	0	63	64	28	1	67
75	52	2	75	65	32	1	67
66	28	1	65	59	6	0	55
55	30	1	60	60	44	1	55
60	29	1	57	81	50	1	56
57	41	1	63	57	20	0	63
51	38	2	73	68	14	0	42
55	43	2	55	71	93	2	65
57	61	2	68	50	53	2	86
66	28	1	62	75	13	0	63
55	30	1	61	75	75	3	76
62	42	1	65	56	30	1	61
61	21	1	63	50	19	1	51
69	19	0	55	74	40	2	62
59	40	2	63	48	9	1	67
51	23	2	58	68	67	3	68
65	54	1	62	54	15	3	54
59	71	1	61	33	51	2	67
64	40	1	64	69	40	2	71
49	53	3	76	67	28	1	50
73	38	1	61	47	33	2	62
58	57	2	57	65	53	2	68
70	23	1	50	64	45	1	64
54	30	0	49	68	65	2	31
55	29	2	50	45	21	1	42
Total	**Age (PD)** = 61 (SD 9 years; min 33, max 81)		**Age (HC)** = 61 years (SD 9 years; min 31, max 86)		**UPDRS** = 38 (SD 18; min 6, max 93)	**UPDRS-speech** = 1,3 (SD 0,82; min 0, max 3)	

three different voice characteristics. From the three characteristics, phonation and articulation can not only be evaluated from repeated and sustained vowels but also from words and groups of words. Similarly, on top of the phonation and articulation, prosody can be measured with sentences, conversations and spontaneous speech, the last being also known as monologue data. In PC-GITA the average duration of the monologues is 44.86 seconds for which the speakers were asked to describe their everyday activities.

4.2 Setup

The aim of this study is to detect PD from spontaneous speech using machine learning. Particularly, in this study a binary classification is used with the values of *one* meaning PD and *zero* meaning HC. The process of this study follows a typical PSP pipeline which is illustrated in figure 2.1. Moreover, the target language for classifying is Spanish and the data is from PC-GITA corpus [13]. The three classifiers were chosen based on success in related work which are introduced in chapter 2. All calculations of this study were done with Python.

Two feature sets were extracted from the monologues of PC-GITA corpus. The first feature set of MFCC was extracted with Librosa [25] with 25ms segments with a 10ms hop length using a Hann windowing function, totaling to 39 coefficients with first and second temporal derivates included. The overall dimensions of MFCC depended on the duration of the monologue. Secondly, for the eGeMAPS the LLD features and functionals were extracted with openSMILE toolkit [20]. The functionals created a vector with the length of 88 features whilst the LLD create multiple 25 features long vectors on each frame and the vectors of LLD formed a matrix together. The dimension of the matrix varied between the speakers based on the length of the monologue. In other words, even though the same number of features per feature set were extracted from each monologue, the number of feature vectors per speaker varies for MFCC and LLD of eGeMAPS as mentioned before.

Additionally, machine learning classifiers used in this study require exact same dimensioned input segments in order to correctly classify all segments belonging to speakers to classes. Therefore, fixed length segments of the feature matrices were created for each speaker, and these were used as input segments. The matrices of MFCC were divided into one-second-long segments for RF and SVM along with two-second-long segments for CNN. In the case of eGeMAPS only 2-second-long segments were used for the CNN since the one-dimensioned functional vectors were acceptable for classifying RF and SVM. Moreover, since the length of the monologues varied, the possibly uneven last segments of each speaker were not used in the calculations. This resulted in losing maximum of a last second of each monologue.

As for the classification part, this thesis uses three different classifying models which are introduced in chapter 3. All three classifier models were optimized with GridSearch by Scikit-learn [26] with concentrating on the important parameters on each model. In more detail of the GridSearch, radial basis function kernel proved to return more accurate results than linear kernel in the case of SVM. Comparatively, the two other optimized parameters, regularization C and kernel scale coefficient γ, varied from 10^{-3} to 10^{2} in multiplies of 10. The optimization resulted in the best regularization of 10 and kernel scale coefficient of 0.1 for MFCC and 1000 and 0.01 for eGeMAPS respectfully.

In the case of Random Forest, the most optimal value for number of estimators was searched with values from 200 to 10^{3} in steps of 200 and maximum depth with values from 10 to 500 with the step of doubling the last step. The best number of estimators was 400 for both feature sets but max depth varied from 100 of MFCC to 300 for eGeMAPS.

Moreover, the architecture of CNN was inspired from researchers Liu et al. [27] in their article “Automatic Assessment of Parkinson’s Disease Using Speech Representations of Phonation and Articulation” but it was modified based on the performance of the model while experimenting. Based on the article, 10 epochs were chosen for the model. The model of MFCC consists of four convolutional layers with filter size of (3,3) and filter count of 64 with an activation function of ReLU. These layers are divided by max-pooling layers with sizes (5,1) and (2,1). The last average-pooling layer sends the information to a dense layer of 64 units. This is followed by two additional dense layers of 16 and 2 units. The last layer returns the prediction of the binary classification using a softmax activation function. However, the CNN of eGeMAPS consists of one convolutional layer less than the CNN of MFCC to avoid overfitting. In addition, the last average-pooling layers has the size of (2,2). Moreover, both models use learning rate of 0.001 with Adam as an optimizer.

5. RESULTS

The results of the experiments done in this study are presented in table 5.1. In the table the mean accuracies of the cross-validations are stated using feature sets MFCC and eGeMAPS. For both of the sets three classifiers were tested with cross-validation. Overall, the scores implicate that Parkinson's disease can be detected from spontaneous speech of Spanish language.

As seen from the table 5.1. the mean classification accuracies obtained in this study vary from 65.25% to 75.00%. The best classifier for MFCC was CNN with the mean accuracy of 67.40%. Comparatively, RF was the most accurate for eGeMAPS with the mean accuracy of 75.00%. In general, eGeMAPS showed better classification accuracy results than MFCC with over 70% mean classification accuracies compared to over 65% accuracies of MFCC. The numbers are logical since eGeMAPS contains MFCC in addition to other features.

Table 5.1. *The mean classification accuracies of cross-validations.*

Feature set	SVM accuracy (%)	RF accuracy (%)	CNN accuracy (%)
MFCC	66.48	65.25	67.40
eGeMAPS	71.00	75.00	70.87

For better understanding of the cross-validation accuracies, statistics of each classification are presented in table 5.2 for MFCC and eGeMAPS. For each cross-validation, a mean standard deviation and an evaluation metric named F1 score are presented. Standard deviations for the classifications with multiple feature segments per speaker were around 20%. For instance, the best classification accuracy obtained with MFCC had a standard deviation of 20.05%. Alternatively, the high valued standard deviations of eGeMAPS with SVM and RF are explained with the structure of the features extracted. For these classifiers the summarized vector of lower-level descriptions of eGeMAPS was used which resulted in one feature segment of data per speaker. Henceforth, a single cross-validation fold could only return either perfectly or completely falsely classified result which fails to return valuable information for the standard deviation.

The F1 score results in range of zero to one in which the maximum result represents a perfectly classified situation [28]. The measurement is a harmonic mean of true positives divided by true positives summed with false positives and negatives. As seen from the

Table 5.2 *Statistics of the accuracies of MFCC and eGeMAPS.*

Classifier	**F1 score**	**Standard deviation (%) of the accuracy**
SVM (MFCC)	0.69	16.24
RF (MFCC)	0.67	18.06
CNN (MFCC)	0.67	20.05
SVM (eGeMAPS)	0.70	45.38
RF (eGeMAPS)	0.76	43.30
CNN (eGeMAPS)	0.70	26.99

table 5.2 the F1 scores support the performance of the classifiers, with over 0.67 values for MFCC and over 0.70 values for eGeMAPS. In the case of MFCC, the best F1 value of 0.69 was obtained by SVM, but the F1 value of 0.67 for the most accurate classifier of CNN had a similar result. Additionally, eGeMAPS performed better than MFCC in case of F1 values with the best F1 value of 0.76 for RF which was the most accurate classifier for eGeMAPS. The results are manageable taking into account the fact the complex nature of spontaneous speech and the feature sets chosen.

Furthermore, the classification results of MFCC and eGeMAPS are visualized with confusion matrices in figure 5.1 for SVM, figure 5.2 for RF and figure 5.3 for CNN. Each confusion matrix displays the number of correctly and falsely classified data inputs while using *zero* to HC and *one* to speakers with PD. In the matrix, y-axis is the estimation made by classifier and x-axis shows the correct label. The monologues consist of multiple feature segments in all other cases than SVM and RF of eGeMAPS, which used the functionals vector of eGeMAPS.

For both cases of MFCC and eGeMAPS, the majority of the segments were classified in correct classes with a similar number of false estimations as seen from the figures 5.1, 5.2 and 5.3. However, SVM and RF of MFCC resulted in substantially more false positives than negatives. Additionally, the results are relatively balanced which is coherent since the database is well balanced based on the distribution of labels.

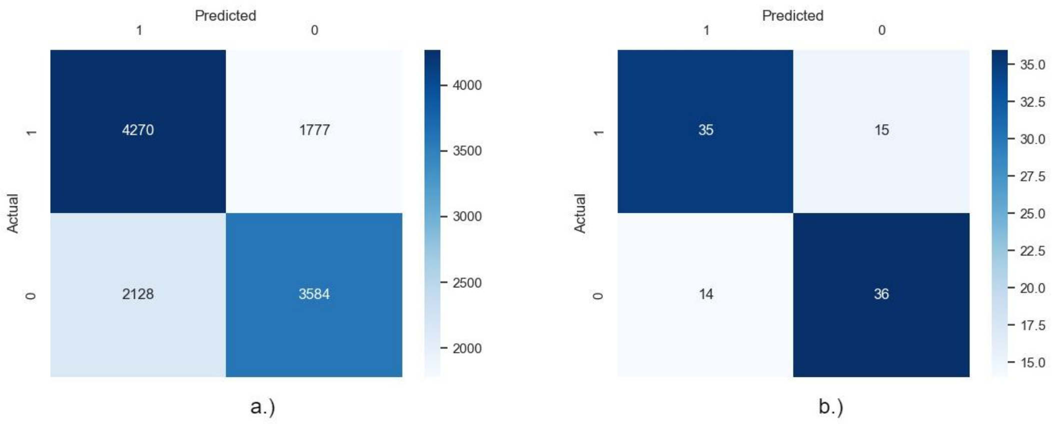

Figure 5.1. *Confusion matrices of SVM when using features of a.) MFCC with multiple feature segments per speaker and b.) eGeMAPS with one feature vector of functionals per speaker.*

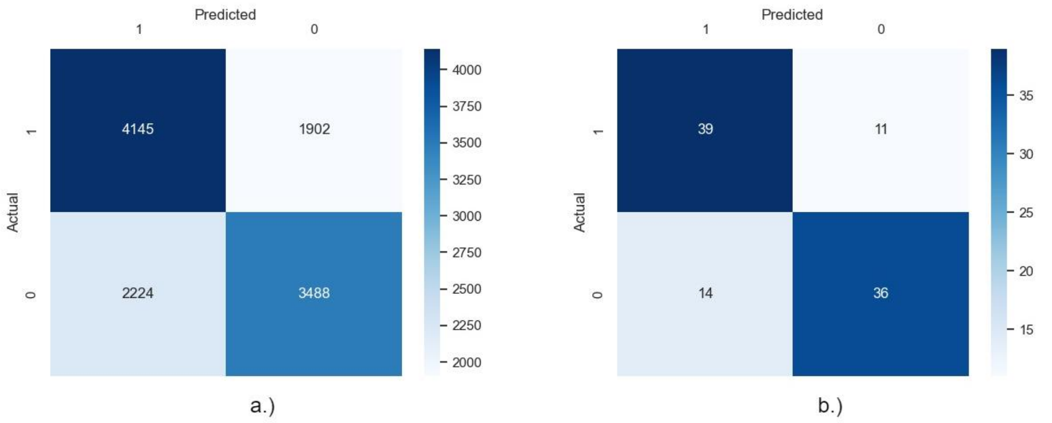

Figure 5.2. *Confusion matrices of RF when using features of a.) MFCC with multiple feature segments per speaker and b.) eGeMAPS with one feature vector of functionals per speaker.*

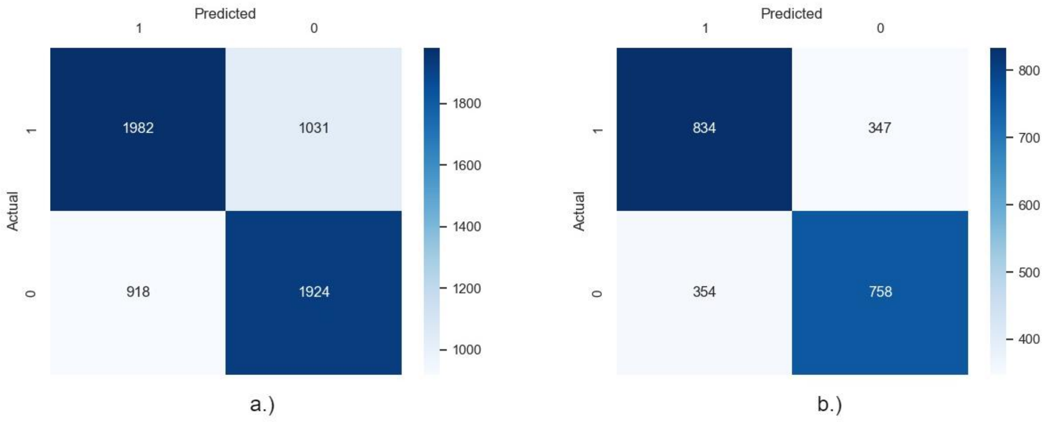

Figure 5.3. *Confusion matrices of CNN when using features of a.) MFCC and b.) eGeMAPS with multiple feature segments per speaker.*

6. CONCLUSIONS

In this study the possibility and accuracy of detecting and assessing PD from spontaneous speech with the help of machine learning techniques was studied. The methodology of this study followed a typical PSP system which is described in chapter 2 and visualized in figure 2.1. The setup of this study consisted of training different machine learning models with features extracted from the monologues of PC-GITA corpus and evaluating classification accuracy with new data. The classification method used in this study was a binary classification between healthy speakers and speakers diagnosed with PD. To avoid overfitting, a cross-validation method was used to train and test a model with all the hundred speakers.

In detail, two feature sets of MFCC and eGeMAPS were tested with three machine learning models separately resulting in suitable cross-validation accuracies with at least over 65% classification accuracy for MFCC and over 70% for eGeMAPS. The best classification accuracy was obtained with RF for eGeMAPS with 75% accuracy and for MFCC the accuracy of 67.40% was acquired with CNN. Additionally, in the case of performance, the classification accuracies with the feature sets with multiple feature vectors per speaker had around 20% standard deviation. Furthermore, F1 values of the models supported the hypothesis that PD is detectable from spontaneous speech with machine learning methods. The best F1 values for each feature set was 0.69 for MFCC and 0.76 for eGeMAPS.

Overall, the results of classification accuracies and model performances acquired in this study are supported by the related works depicted in chapter 2. In the related works similar accuracies and performances were acquired with the same PC-GITA corpus and other similar speech databases with the help of related feature sets and classifiers. Additionally, in chapter 5 the classification results are evaluated and justified.

Nonetheless, the results of this study have scope for improvement, but they are explainable by plethora of reasons. The major challenge of similar speech assessment tasks is the unpredictability and complexity of spontaneous speech while using a small sized speech corpus. Furthermore, from the methodology aspect of this study, an improvement could be gained from richer feature sets or using only the most relevant features for the task of detecting PD. Additionally, the machine learning models could be optimized for the task or pre-taught by a possible bigger speech corpus and then fine-tuned with the target data set.

Moreover, this study concentrated only on the paralinguistic information of speech. The diagnosing process could benefit also from linguistic information. Linguistic speech impairments could be detected, for example, with the help of transcriptions of the monologues. Additionally, the research can be expanded with other languages.

REFERENCES

[1] J.C. Vásquez-Correa, T. Arias-Vergara, C. D. Rios-Urrego, M. Schuster, J. Rusz, J. R. Orozco-Arroyave, E. Nöth, *Convolutional Neural Networks and a Transfer Learning Strategy to Classify Parkinson’s Disease from Speech in Three Different Languages*, Lecture Notes in Computer Science, Springer Nature Switzerland AG, vol.11896, Oct 2019, pp.697-706. DOI: 10.1007/978-3-030-33904-3_66

[2] E. Ronken, G.J.M. Scharrenburg, *Parkinson’s disease*, IOS Press Incorporated, vol.1, 2002.

[3] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel, B. Dubois, R. Holloway, J. Jankovic, J. Kulisevsky, A. E. Lang, A. Lees, S. Leurgans, P. A. LeWitt, D. Nyenhuis, C. Warren Olanow, O. Rascol, A. Schrag, J. A. Teresi, J. J. van Hilten, N. LaPelle, *Movement Disorder Society-sponsored revision of the Unified Parkinson’s Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results,* Movement Disorders, vol.23, vo.15, 2008, pp.2129–2170. DOI: 10.1002/mds.22340.

[4] T. Tykalova, J. Rusz, J. Klempir, R. Cmejla, E. Ruzicka, *Distinct patterns of imprecise consonant articulation among Parkinson’s disease, progressive supranuclear palsy and multiple system atrophy*, Brain and Language, vol.165, Feb 2017, pp.1–9. DOI: 10.1016/j.bandl.2016.11.005.

[5] K. M. Smith, D. N. Caplan, *Communication impairment in Parkinson’s disease: Impact of motor and cognitive symptoms on speech and language,* Brain and Language, vol.185, Oct 2018, pp.38–46. DOI: 10.1016/j.bandl.2018.08.002.

[6] M. H. Shirali-Shahreza, S. Shirali-Shahreza, *Effect of MFCC normalization on vector quantization based speaker identification*, IEEE International Symposium on Signal Processing and Information Technology, 2010, pp.250–253. DOI:10.1109/ISSPIT.2010.5711789.

[7] A. Sosnovshchenko, *Machine Learning with Swift,* Packt Publishing, 1st edition, 2018.

[8] T. Bäckström, O. Räsänen, A. Zewoudie, P. P. Zarazaga, L. Koivusalo, S. Das, E. G. Mellado, M. B. Mansali, D. Ramos, *Introduction to Speech Processing*, 2nd edition, 2022. DOI: 10.5281/zenodo.6821775. URL: https://speechprocessingbook.aalto.fi. (visited: 6.12.2022)

[9] B. Schuller, A. Batliner, *Computational paralinguistics: emotion, affect and personality in speech and language processing,* John Wiley & Sons, 2014.

[10] A. Airola, T. Pahikkala, W. Waegeman, B. De Baets, T. Salakoski, T, *An experimental comparison of cross-validation techniques for estimating the area under the ROC curve*, Computational Statistics & Data Analysis, Vol.55, 2011, pp.1828–1844. DOI: DOI: 10.1016/j.csda.2010.11.018.

[11] N. P. Narendra, P. Alku, *Dysarthric speech classification using glottal features computed from non-words, words and sentences*, Proceedings of Interspeech, International Speech Communication Association, Sep 2018, pp. 3403–3407. DOI: 10.21437/Interspeech.2018-1059.

[12] B. Karan, S. S. Sahu, J. R. Orozco-Arroyave, K. Mahto, *Non-negative matrix factorization-based time-frequency feature extraction of voice signal for Parkinson's disease prediction*, Computer Speech & Language, vol.69, Sep 2021. DOI: 10.1016/j.csl.2021.101216.

[13] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. González-Rátiva, E. Nöth, *New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease,* Proceedings of the 9th International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA), 2014, pp.342–347.

[14] Z. S. Syed, S. A. Memon, A. L Memon, *Deep acoustic embeddings for identifying parkinsonian speech*, International Journal of Advanced Computer Science and Applications, vol.11, no.10, 2020, pp.726–734. DOI: 10.14569/IJACSA.2020.0111089.

[15] P.A. Pérez-Toro, J. C. Vásquez-Correa, M. Strauss, J. R. Orozco-Arroyave, E. Nöth, *Natural Language Analysis to Detect Parkinson's Disease*, Lecture Notes in Computer Science, Springer Nature Switzerland AG, vol.11697, 2019, pp.82–90. DOI: 10.1007/978-3-030-27947-9_7.

[16] K. M. M. Prabhu, *Window functions and their applications in signal processing*, CRC Press/Taylor & Francis, 1st edition, 2014. DOI: 10.1201/9781315216386.

[17] F. Zheng, G. Zhang, Z. Song, *Comparison of different implementations of MFCC*, Journal of Computer Science and Technology, Springer Nature B.V., vol.16, no.6, 2001, pp.582–589. DOI: 10.1007/BF02943243.

[18] M. S. B. A. Ghaffar, U. S. Khan, J. Iqbal, N. Rashid, A. Hamza, W. S. Qureshi, M. I. Tiwana, U. Izhar, *Improving classification performance of four class FNIRS-BCI using Mel Frequency Cepstral Coefficients (MFCC)*, Infrared Physics & Technology, vol.112, 2021. DOI: 10.1016/j.infrared.2020.103589.

[19] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. Andre, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, K. P. Truong, *The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing*, IEEE Transactions on Affective Computing, vol.7, no.2, Apr-Jun 2016, pp.190–202. DOI: 10.1109/TAFFC.2015.2457417.

[20] *OpenSMILE*, audEERING. URL: https://www.audeering.com/research/opensmile/ (visited on 6.12.2022)

[21] J. Terzic, E. Terzic, R. Nagarajah, M. Alamgir, *Ultrasonic Fluid Quantity Measurement in Dynamic Vehicular Applications A Support Vector Machine Approach*, Springer International Publishing, 1st edition, 2013. DOI: 10.1007/978-3-319-00633-8.

[22] S. Chandramouli, S. Dutt, A. K. Das, *Machine Learning*, Pearson Education India, 2018.

[23] A. R. Webb, K. D. Copsey, *Statistical pattern recognition*, John Wiley & Sons, 13th edition, 2011.

[24] Y. Lecun, Y. Bengio, G. Hinton, *Deep learning*, Nature Publishing Group, vol.521, no.7553, 2015, pp. 436–444. DOI: 10.1038/nature14539.

[25] B. McFee, A. Metsai, M. McVicar, S. Balke, C. Thomé, C. Raffel, F. Zalkow, A. Malek, D. K. Lee, O. Nieto, D. Ellis, J. Mason, E. Battenberg, S. Seyfarth, R. Yamamoto, viktorandreevichmorozov, K. Choi, J. Moore, R. Bittner, S. Hidaka, Z. Wei, nullmightybofo, A. Weiss, D. Hereñú, F.-R. Stöter, L. Nickel, P. Friesch, M. Vollrath, T. Kim, *Librosa* 0.9.2. URL: https://librosa.org/doc/latest/index.html (visited on 6.12.2022)

[26] Scikit-learn, *Grid Search CV documentation*. URL: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html (visited on 6.12.2022)

[27] Y. Liu, M. K. Reddy, N. Penttilä, T. Ihalainen, P. Alku, O. Räsänen, *Automatic Assessment of Parkinson's Disease Using Speech Representations of Phonation and Articulation*, IEEE/ACM Transactions on Audio, Speech and Language Processing, Sep 2022. DOI: 10.1109/TASLP.2022.3212829.

[28] C. Davide, G. Jurman, T*he Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation*, BMC genomics, Springer Nature, vol.21, no.1, 2020, pp.6–6.