INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2020IPPAT039

Thèse de doctorat

`

TELECOM Paris

IP PARIS

# Parkinson's Disease Detection by Multimodal Analysis Combining Handwriting and Speech Signals

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Telecom Paris

École doctorale n°626 École Doctoral de l'Institut Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat: Signal, Images, Automatique et Robotique

Thèse soutenue le 26 Novembre 2020, par

**Catherine Taleb**

Composition du Jury:

| | |
|---|---|
| Thierry Paquet<br>Professeur, Université de Rouen | Président |
| Nicole Vincent<br>Professeur, LIPADE, Université de Paris | Rapportrice |
| Björn Schuller<br>Professeur, Université d'Augsburg, Germany | Rapporteur |
| Mounim El-Yacoubi<br>Professeur, Samovar, Telecom SudParis | Examinateur |
| Laurence Likforman-Sulem<br>Professeur Associé, Telecom Paris, IP Paris | Directrice de thèse |
| Chafic Mokbel<br>Professeur, Université de Balamand, Liban | Co-Directeur de thèse |

ECOLE DOCTORALE

**Titre :** Détection de la Maladie de Parkinson par Analyse Multimodale Combinant Signaux d'Écriture et de Parole

**Mots clés :** Maladie de Parkinson (MP), détection précoce, écriture, parole, multimodale, langage-indépendant

**Résumé :** Les troubles dégénératifs tels que la maladie de Parkinson (MP) ont une influence sur les activités quotidiennes en raison de la rigidité des muscles, des tremblements ou des troubles cognitifs. La MP est un trouble complexe caractérisé par plusieurs symptômes moteurs et non-moteurs qui s'aggravent avec le temps et qui diffèrent d'une personne à l'autre. Aux stades avancés de la MP, le diagnostic clinique est clair. Cependant, dans les premiers stades, lorsque les symptômes sont souvent incomplets ou subtils, le diagnostic devient difficile et, parfois, le sujet peut rester non diagnostiqué. Cette difficulté motive fortement la création d'outils d'évaluation informatisés, d'outils d'aide à la décision et d'instruments de test susceptibles de faciliter le diagnostic précoce et la prévision de l'évolution de la MP. L'écriture, la parole et les mouvements oculaires sont affectés par la MP à un stade précoce et peuvent être utilisés pour concevoir des tests cliniques non invasifs et peu coûteux. De nombreuses études dans la littérature ont analysé la caractérisation de l'écriture, de la parole et des mouvements oculaires chez les patients atteints de la MP pour la détection automatique de la maladie en utilisant ces signaux, mais aucune d'entre elles ne s'est concentrée sur la construction d'un modèle langage-indépendant pour détecter la maladie à un stade précoce en utilisant la combinaison de ces trois signaux. Dans ce travail, une base de données multimodale et multilingue comprenant des enregistrements d'écriture, de parole et de mouvements oculaires est construite dans ce but et en raison du manque de telles bases de données. Dans cette thèse, en raison du temps limité, nous nous sommes concentrés sur l'analyse de l'écriture manuscrite et de la parole, tandis que les mouvements oculaires seront étudiés plus tard. Cependant, puisque notre objectif est la détection précoce de la maladie, et que la collection d'une grande base de données aux stades précoces (avec de nombreux échantillons à chaque stade) est très difficile, nous faisons l'hypothèse que les patients atteints de la MP sous traitement ont des imperfections plus proches des stades précoces de la maladie. Pour cette raison, dans cette thèse, les patients atteints de la MP sont étudiés après avoir pris le médicament L-dopa. Un système automatique de détection précoce de la MP par l'analyse de l'écriture est d'abord construit, où deux approches sont considérées, étudiées et comparées: une approche classique d'extraction de caractéristiques et de classification et une approche d'apprentissage profond. Dans l'approche classique d'extraction de caractéristiques et de classification, une combinaison de caractéristiques globales et indépendantes de la langue, telles que caractéristiques de l'accident vasculaire cérébral, caractéristiques cinématiques et temporelles, caractéristiques de pression, caractéristiques d'entropie et d'énergie, et caractéristiques intrinsèques, sont utilisés pour diagnostiquer la MP, où seule la composante "sur papier" est étudiée. La sélection des caractéristiques se fait en deux étapes: la première étape sélectionne un sous-ensemble à l'aide d'une analyse statistique, et la deuxième étape sélectionne les caractéristiques les plus pertinentes de ce sous-ensemble par une approche sous-optimale. Les caractéristiques sélectionnées sont introduites dans un classificateur de machine à vecteur de support (SVM) afin d'identifier les sujets souffrant de la MP. Nous avons constaté que l'écriture manuscrite peut être un outil de diagnostic précoce de la MP avec une performance de prédiction de 96.87 % lorsqu'une combinaison de la cinématique, de la pression et de la corrélation entre la cinématique et la pression est utilisée. Sur la base de ces caractéristiques sélectionnées, un classificateur SVM multi-classes est construit pour la détection de stade, où la segmentation par fenêtre est appliquée pour augmenter le nombre d'échantillons. Un modèle de perceptron multicouche (MLP) est utilisé pour combiner les scores de tous les segments d'un même sujet. De nombreux obstacles sont rencontrés dans cette partie, tels que la grande variabilité des patients en termes de symptômes et de stade de la maladie, le déséquilibre des données et de la distribution des classes, et le nombre limité d'échantillons.

En raison de ces obstacles, nous devons effectuer notre modalité sur un grand ensemble de données équilibrées pour déterminer si notre modèle est efficace ou non.

Dans la deuxième approche, nous avons proposé un modèle basé sur les caractéristiques à court terme et l'apprentissage profond, où le modèle est formé sur toutes les langues afin que le vecteur de caractéristiques ne soit pas biaisé vers une langue spécifique. Des modèles 2D réseaux de neurones convolutifs (CNN) et 1D CNN-BLSTM (bidirectionnelle - mémoire à court terme longue) sont proposés pour la classification de séries temporelles de bout en bout. Nous avons étudié les signaux dynamiques de l'écriture manuscrite entiers afin d'extraire des caractéristiques à la fois dans l'air et sur la surface. Le spectrogramme et le champ angulaire gramian modifié sont appliqués pour encoder les séries temporelles en images pour le modèle 2D CNN, où les signaux bruts sont utilisés directement avec le 1D CNN-BLSTM. Nous avons démontré l'importance des deux éléments suivants: une architecture profonde basée sur la combinaison de couches récurrentes 1D CNN et BLSTM, et un modèle 2D CNN avec des spectrogrammes en entrée pour la détection de la MP, en raison de leur capacité à traiter la variation de l'information dans les séries temporelles, soit en considérant explicitement l'information locale à court terme sur l'axe temporel des signaux d'écriture manuscrite en ligne non stationnaires ou en traitant directement les séries temporelles brutes. Pour faire face aux données limitées, et pour améliorer nos modèles profonds, des approches d'augmentation des données sont appliquées. La tâche difficile de détection précoce de la MP est abordée avec succès en utilisant le modèle 1D CNN-BLSTM avec la combinaison des méthodes d'augmentation de la gigue et des données synthétiques, ce qui donne une précision de 97.62 %.

Après avoir réussi à construire un modèle langage-indépendant pour le diagnostic précoce de la MP en utilisant l'analyse de l'écriture, la troisième partie de cette thèse consiste à construire un ensemble de caractéristiques acoustiques indépendantes du langage et de la tâche pour évaluer les troubles moteurs chez les patients atteints de MP, et à étudier l'influence du taux d'échantillonnage et des sons non vocalisés sur la performance. Seules les caractéristiques artisanales de la phonation et de l'articulation qui peuvent être extraites pour toutes les tâches évaluées sont étudiées. Les effets des sons non voisés et du taux d'échantillonnage sur la performance de classification de la détection de la MP par l'analyse de la voix sont étudiés. Nous avons réussi à construire un modèle SVM indépendant du langage pour le diagnostic précoce de la MP par l'analyse de la voix avec une précision de 97.62 %.

La quatrième partie de cette thèse s'est concentrée sur la construction d'un système multimodal indépendant du langage pour la détection précoce de la MP en combinant l'écriture manuscrite et les signaux vocaux; où les deux modèles SVM classiques et les modèles d'apprentissage profond ont été analysés. Des méthodes de fusion globale au niveau des caractéristiques et de la décision sont appliquées pour former un vecteur multimodal qui sera utilisé pour la détection. Une précision de classification allant jusqu'à 100 % est obtenue lorsque les caractéristiques artisanales des deux modalités sont combinées et appliquées au SVM, où la précision obtenue en combinant les caractéristiques audio et d'écriture apprises en profondeur est similaire à celle obtenue avec les caractéristiques d'écriture apprises en profondeur. Nous avons aussi étudié la corrélation entre les caractéristiques artisanales et les caractéristiques apprises en profondeur afin d'améliorer l'interprétabilité des caractéristiques manuscrites/acoustiques profondes extraites.

L'objectif principal de cette thèse était l'analyse de l'écriture manuscrite, où notre système de diagnostic de la MP à partir de l'écriture manuscrite a été validé sur une base de données publique. En ce qui concerne les analyses de la parole et les analyses multimodales, nous y avons travaillé dans la dernière phase de la thèse. En raison du temps limité, nous avons cherché d'autres formes d'ondes vocales brutes et des bases de données multimodales (écriture manuscrite et échantillons de parole) qui sont disponibles publiquement et ne nécessitent pas de négociation avec leurs propriétaires afin de pouvoir valider nos systèmes de diagnostic de la MP sur d'autres bases de données. En raison de la non-disponibilité d'une autre base de données multimodale publique ou même d'une base de données vocale avec des formes d'onde vocales brutes, nous n'avons pas pu valider nos systèmes de diagnostic de la MP. Malgré les résultats encourageants obtenus, il reste encore du travail à faire avant de mettre notre modèle multimodal de détection précoce de la MP en application clinique, car nous disposons de peu de sujets, par rapport au monde réel où nous aurions des milliers de sujets.

**Title :** Parkinson's Disease Detection by Multimodal Analysis Combining Handwriting and Speech Signals

**Keywords :** Parkinson's disease (PD), early detection, handwriting, speech, multimodal, language-independent.

**Abstract:** Degenerative disorders such as Parkinson's disease (PD) have an influence on daily activities due to rigidity of muscles, tremor or cognitive impairment. PD is a complex disorder characterized by several motor and non-motor symptoms that worsen over time, and that differ from person to another. In advanced stages of PD, clinical diagnosis is clear-cut. However, in the early stages, when the symptoms are often incomplete or subtle, the diagnosis becomes difficult and at times, the subject may remain undiagnosed. This difficulty is a strong motivation for computer-based assessment tools/decision support tools/test instruments that can aid in the early diagnosing and predicting the progression of PD. Handwriting, speech, and eye movements are affected by PD at early stages and can be used to devise noninvasive and low cost clinical tests. Many studies in the literature analyzed the characterization of handwriting, speech, and eye movements in Parkinson patients for the automatic detection of PD by using these signals, but any of them focused on building a language-independent model to detect PD at early stages using the combination of these three signals.

In this work, a multimodal and multilingual database that includes handwriting, speech, and eye movements' recordings is built for this aim and due to the lack of such databases. In this thesis, due to limitation time, we focused on handwriting and speech analysis, where eye movements will be studied later. However, since our target is the early detection of the disease, and since collecting large database at early stages (with many samples at each stage) is very difficult, we make the assumption that PD patients with medication on have imperfections closer to the early stages of the disease. For this reason, in this thesis the PD patients are studied after taking the L-dopa medication. An automatic system for PD early detection through handwriting analysis is first built, where two approaches are considered, studied and compared: a classical feature extraction and classifier approach and a deep learning approach. In the classical feature extraction and classification approach, a combination of global and language-independent features such as stroke, kinematic and temporal, pressure, and advanced handwriting markers based on entropy, energy and intrinsic measures of handwriting, are used to diagnose PD, where only the "on-paper" component are studied. Feature selection is done in two stages; the first stage selected a subset using statistical analysis, and the second step selected the most relevant features of this subset by a suboptimal approach. The selected features are fed to a support vector machine (SVM) classifier to identify the subjects suffering from PD. We found that handwriting can be a tool for PD early diagnosis with a 96.87 % prediction performance when a combination of kinematic, pressure and the correlation between kinematic and pressure is used. Based on these selected features, a multi-class SVM classifier is built for stage detection, where windowed segmentation is applied to increase the number of samples. A multilayer perceptron (MLP) model is used to combine the scores of all segments from one subject. Many obstacles are faced in this part such as: large variability over patients in term of symptoms and stage of disease, imbalanced data and class distribution, and limited number of samples. Due to these obstacles, we need to perform our modal on a large and balanced dataset to assert whether our PD stage detection model is efficient or not.

In the second approach, we proposed a model based on short-term features and deep learning, where the model is trained on all the languages so the features vector will not be biased toward a specific language. 2D convolutional neural network (CNN) and 1D CNN-BLSTM (bidirectional-long short term memory) models are proposed for end to end time series classification; where we have studied the whole handwriting dynamic signals so we can extract both in-air and on-surface features. Spectrogram and modified gramian angular field are applied to encode time series into

images for the 2D CNN model, where the raw signals are used directly with the 1D CNN-BLSTM. We have demonstrated the importance of both: a deep architecture based on the combination of 1D CNN and BLSTM recurrent layers, and a 2D CNN model with spectrograms input in PD detection due to their ability to tackle the variation of information in time series either by explicitly considering the local short term information on the time axis of the non-stationary online handwriting signals or by dealing with raw time series directly. To cope with the limited data, and to improve our deep models, data augmentation approaches are applied. The challenging PD early detection task is successfully tackled using the 1D CNN-BLSTM model with the combination of jittering and synthetic data augmentation methods yielding an accuracy of 97.62 %.

After succeeding in building a language-independent model for PD early diagnosis using handwriting analysis, the third part of this thesis is to build a language and task-independent acoustic feature set for assessing the motor disorders in PD patients, and to study the influence of sampling rate and unvoiced sounds on the performance. Only phonation and articulation handcrafted features that can be extracted for all the tasks under assessment are studied. Unvoiced sounds and sampling rate effects on classification performance of PD detection through voice analysis are studied. We have succeeded to build a language-independent SVM model for PD early diagnosis through voice analysis with 97.62 % accuracy.

The fourth part of this thesis focused on building a language-independent multimodal system for PD early detection by combining handwriting and voice signals; where both classical SVM model and deep learning models were both analyzed. Global feature-level and decision-level fusion methods are applied to form a multimodal vector that will be used for detection. A classification accuracy up to 100 % is obtained when handcrafted features from both modalities are combined and applied to the SVM, where the accuracy obtained by combining both deep learned audio and handwriting features is similar to the one obtained with deep learned handwriting features.

We have also studied the correlation between hand-crafted and deep learned features to enhance interpretability of the extracted handwriting/acoustic deep features.

The main focus of this thesis was on handwriting analysis; where our PD diagnosis system from handwriting has been validated on a publicly available database. For speech and multimodal analyses, we have been working on them in the last phase of the thesis. Due to limitation time, we were only searching for another raw speech waveforms and multimodal (handwriting and speech samples) databases that are publically available and do not need a negotiation with their owners so we could validate our PD diagnosis systems on another databases. Due to the non-availability of another public multimodal database or even a speech database with raw speech waveforms, we could not validate our PD diagnosis systems. Despite the encouraging results obtained, there are still some works to do before putting our PD early detection multimodal model into clinical use due to the fact that we have few subjects, in comparison with the real world where we would have thousands of subjects.

# Acknowledgment

I would like to thank my PhD supervisors, Professor Laurence Likforman-Sulem and Professor Chafic Mokbel. I do not have enough words to thank them for their wonderful support throughout this thesis. It has been a pleasure working under their supervision. Laurence was always there making sure everything is done with precision. Her advices and comments were very valuable to me. Chafic has always been an inspiration to me. I have learned a lot from him throughout these years. He helped me make the right decisions and pointed me in the right direction.

I would also like to thank my committee members, Professor Nicole Vincent, Professor Björn Schuller, Professor Thierry Paquet, and Professor Mounim EL-Yacoubi for serving as my committee members even at hardship. I also want to thank you for letting my defense be an enjoyable moment, and for your brilliant comments and suggestions, thanks to you.

I would especially like to thank the Chairperson of the department of biomedical science at the University of Balamand, Professor Maha Khachab who have been there to support me to get the approval from the institutional review board (IRB) of the University of Balamand and Saint George Hospital University Medical Center to collect the database, and who was a reference to me concerning the neurological part of my thesis.

I would like to take the opportunity to thank Sami Rahi, the neurologist at Saint George Hospital, who allowed me to contact the Parkinson’s disease patients attending his clinic, once taking their approval. Also I would like to extend my thanks to all the patients and control subjects who had participated in my study, thank you for your trusts.

Finally, my family, colleagues and friends, I can not thank you enough for encouraging me throughout this experience and for supporting me during these years.

# Contents

# 1 Introduction, Hypothesis and Goals

## 1.1 Introduction and hypothesis

Parkinson's disease (PD) is a neurodegenerative disorder that was first described in 1817 by Dr. James Parkinson, a medical doctor, as "involuntary tremulous motion, with lessened muscular power, in parts not in action and even when supported; with a propensity to bend the trunk forwards, and to pass from a walking to a running pace: the senses and intellects being uninjured" [Parkinson, 1969]. It is a complex disorder characterized by several motor and non-motor symptoms that worsen over time. In advanced stages of PD, clinical diagnosis is clear-cut. However, in the early stages, when the symptoms are often incomplete or subtle, the diagnosis becomes difficult and at times, the subject may remain undiagnosed. Moreover, monitoring the progression of the disease over time requires repeated clinical visits by the patient [Nilashi et al., 2016]. The difficulty in early detection and monitoring the progression of PD is a strong motivation for computer-based assessment tools/decision support tools/test instruments that can aid in the early diagnosing and predicting the progression of PD. Timely detection, preferably at a stage earlier than currently possible, and subsequent intervention could be hugely beneficial in a way that the patient could have access to disease modifying therapy to slow down the course of PD progression. Studies have shown that 90 % of people with PD show some form of vocal impairment; this impairment may also be one of the earliest indicators for the onset of the illness [Little et al., 2008]. Some other studies have indicated that handwriting can be with advance used for differential diagnosis of PD, since the deterioration of handwriting is one of the first manifestations of PD [Pereira et al., 2016]. In addition, saccadic eye movements can help in developing early biological markers for PD as saccadic eye movements' circuitry involves both cortical and subcortical brain areas and saccadic task manipulation gives an insight into the information processing in impaired brain [Goyal et al., 2014]. These impairments can play a role in the early detection of the disease and in the monitoring of its progression when detected.

## 1.2 Goals and objectives

The previous section clearly shows the importance of having an efficient, cost effective, highly accessible and non-invasive PD early detection system that can also serve in differential diagnosis. Actually, this shall improve the health care service and a close follow-up of the patients. Such a system can use handwriting, speech and eye movements to early detect and monitor the progress of PD.

According to the reviewed literature, a language-independent model to detect PD using multimodal signals has not been enough addressed. The main goal of this thesis is to build a language-independent multimodal system for assessment the motor disorders in PD patients at an early stage based on combined handwriting and speech signals, using machine learning techniques that are becoming popular in biomedicine: logistic regression, random forests, boosted tree, support vector machine (SVM), and deep learning. The resulting system will serve as supporting tool in the differential diagnosis of PD for new patients.

The specific objectives of this thesis are as follows:

- The targeted system is data driven. We rely on PD datasets in order to build, assess and benchmark solutions. For this purpose and due to the lack of a multimodal and multilingual dataset a first objective of this thesis is to build a multilingual and multimodal database equally distributed between healthy control (HC) subjects and PD patients, where the PD patients will be examined before and after taking L-dopa medication. The database must include handwriting, speech, and eye movements' recordings. However, since our target is the early detection of the disease, and since collecting large database at early stages (with many samples at each stage) is very difficult, we make the assumption that PD patients with medication on have imperfections in their handwriting, speech and eye movements closer to the early stages of the disease. For this reason, in this thesis the PD patients will be studied after taking the L-dopa medication.

- The second objective of the thesis is to build PD diagnosis automatic system from handwriting. Two approaches are to be considered, studied and compared: global hand-crafted feature and classical classifier approach and short-term features and deep learning approach. In the first class of approach, global and language-independent handcrafted features need to be defined, extracted and used with a machine learning classifier, e.g. SVM. The second approach consists of building a system based on short-term features and deep learning, where the system need to be trained on all the languages so the features vector will not be biased toward a specific language, and where some data augmentation techniques shall be applied to overcome the limited size of the dataset.

- In this thesis, we focus mainly on PD early detection from handwriting but we also present first experiments on PD detection from speech where the approaches described in the previous point will be also studied here. We then combine both handwriting and speech modalities in order to compensate the lack of data and to enhance the reliability of detecting PD.

The next chapter provides a detailed description of the PD, its physiological origins, the induced symptoms, the clinical diagnosis and the possible medications, and how this disease can affect handwriting, speech, and saccadic eye movements. The thesis proceeds, in chapter 3, with a review of the different methodologies related to the analysis of handwriting, speech and eyes movement in Parkinson patients for the automatic detection of PD by using these signals. Chapter 4 provides a description of our constructed multimodal database. In chapter 5, we start with a discussion of our PD detection and progression model trained on global and language-independent hand-crafted features extracted from online handwriting signals, later we describe the deep learning models implemented for PD detection using online handwriting. The automatic non-invasive PD detection models based on speech are described in chapter 6. Starting with a description of our PD detection model trained on pre-engineered features extracted from speech signals, and the influence of sampling rate on PD prediction. In chapter 7, a description of our deep learning models implemented for PD detection using speech signals is provided, and multimodal assessment of PD is defined: feature-based approach and deep learning approach. Finally in chapter 8, observations and

conclusions obtained in this thesis are listed with a number of perspectives that we would be working on in the future.

# 2 Parkinson’s Disease

This chapter introduces PD, its physiological origins, the induced symptoms, the rating scales used to assess the stage and the progression of the disease, the clinical diagnosis, the possible medications and their side effects on movements, and how this disease can affect handwriting, speech, and eye movements even at the early stages. It also highlights the difficulty of detecting the disease at early stages and the importance of building an efficient and non-invasive system for early detection; where the early detection can be hugely beneficial in a way to slow down the course of PD progression. Hence, this chapter will provide the reader with an overview of the condition and the reasons to seek for new schemes to help clinicians in the diagnosis and assessment.

## 2.1 Parkinson’s disease

PD is today the second most common neurological disorder causing disability after stroke and Alzheimer [Dubayová et al., 2010]. The disease occurs more frequently in men than in women in every decade of life, which is explained by the neuroprotective effects of estrogens. Prevalence and incidence of PD in European countries was estimated at approximately 108 to 257/100,000 and 11 to 19/100,000 per year, respectively. The prevalence in Asian countries is slightly lower, all-age prevalence varied from 51.3 to 176.9/100,000 persons and the incidence from 6.7 to 8.7 per 100 000 persons per year. Prevalence and incidence rates are the lowest in African countries – the crude prevalence varied from 7 to 31.4/100,000 persons and the crude incidence rate of PD was 4.5/100,000 persons per year [Dubayová et al., 2010].

The classical characteristics symptoms of PD are tremor at rest, muscular rigidity, akinesia (a delay in the beginning of movements with long reaction times), bradykinesia (slowness of movement), and postural stability. The most common medication prescribed to treat patients with PD is Levodopa (L-dopa), since the loss of dopaminergic neurons in the substantia nigra in the brain is associated with the appearance of the main motor symptoms in PD. The etiology of PD is also related to age and exposure to free radicals and external toxins, as well as genetic mutations [Weiner et al., 2013].

## 2.2 Pathophysiology

In PD, neurons (nerve cells) of the brain area known as the substantia nigra (Latin for "black substance") are primarily affected (Figure 2.1). When neurons in the substantia nigra degenerate, the brain's ability to generate body movements is disrupted, and this disruption produces signs and symptoms characteristic of PD that will be described in section 2.4 [Weiner et al., 2013].

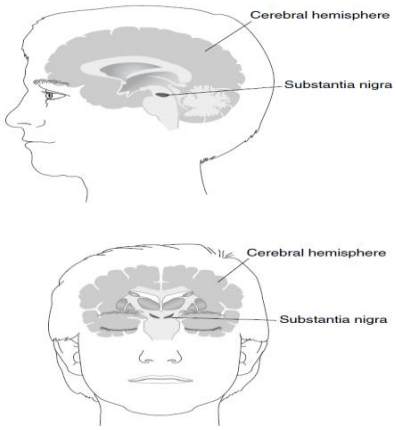



Figure 2.1. Location of Substantia Nigra. This figure shows the location of the substantia nigra (the area of the brain that contains dopamine cells), deep within the brain. The large cerebral hemispheres enclose and cover the substantia nigra as well as other deep midbrain structures [Weiner et al., 2013].

The symptoms of central nervous system diseases are often determined by the location of the neurons that degenerate in a specific location. For example, Alzheimer's disease involves the degeneration of neurons of the cerebral cortex and results in memory loss and mental deterioration. Similarly, the degeneration of the neurons in the substantia nigra, an area of the brain located in the basal nuclei (BG), triggers the appearance of the cardinal motor symptoms in PD [Weiner et al., 2013].

The substantia nigra is a very small area located deep within the brain. There is one substantia nigra on the right side of the brain and one on the left, and often one side is affected before the other. Because of this, people with PD often experience symptoms primarily on one side of their body, particularly in the early stages [Weiner et al., 2013]. The symptoms of PD do not become noticeable until about 80 percent of the cells of the substantia nigra have died, because the human nervous system has multiple safety factors and

redundancies built into it. For a long time, these safety factors have been able to take over the activities of dying cells [Weiner et al., 2013].

In brain biopsies taken from individuals with PD, the brain appears to be relatively normal except that the substantia nigra has lost its usual black pigment. The loss of the dopamine-producing cells in this area of the brain, accompanied by the presence of clumps of alpha-synuclein protein (known as Lewy bodies), has been the hallmark of Parkinson's for decades [Weiner et al., 2013]. The substantia nigra accounts for an extremely small percentage of the brain's weight, but because of its important electrochemical connections with motor centers (brain centers that control movement), it is a vital component in how we move [Weiner et al., 2013]. Specifically, a series of complicated electrical and chemical events within the brain transmits information from neuron to neuron. The chemicals that brain cells use to communicate with one another are called neurotransmitters or, more generally, neurochemicals. The specific neurotransmitter produced and used by the substantia nigra is dopamine [Weiner et al., 2013]. When the cells of the substantia nigra degenerate and die, dopamine is lost and dopamine-relayed messages to other motor centers cannot go through, resulting in dysfunction of the BG [Weiner et al., 2013]. The classic pathophysiological model of the BG for a healthy and a PD patient is summarized in Figure 2.2. The BG consists of four main subnuclei: striatum (comprises the caudate nucleus, putamen, and nucleus accumbens), globus pallidus [internal segment (GPi) and external segment (GPe)], subthalamic nucleus (STN), and substantia nigra [compact (SNc) and reticular (SNr)] [Voytek, 2006]. This model diagrammed a "direct" and "indirect" circuit within the BG. The direct pathway is said to preferentially target striatal D1 receptors, whereas the indirect pathway is said to preferentially target D2 receptors [Rodriguez-Oroz et al., 2009]. These receptors modulate excitation and inhibition in the circuit, respectively. Ultimately, both pathways project to the cortex via the anterior thalamus. In the direct pathway, striatum is directly connected to GPi. This pathway facilitates the movements [Rodriguez-Oroz et al., 2009]. In the indirect pathway, striatum is connected to GPi through the external GPe and STN. This pathway decreases the movements. Dopamine from the SNc facilitates putaminal neurons in the direct pathway and inhibits those in the indirect pathway [Rodriguez-Oroz et al., 2009]. Activation of the direct pathway leads to reduced neuronal firing in the GPi/SNr and movement facilitation, while activation of the indirect pathway

suppresses movements. In PD, dopamine deficit leads to increased activity in the indirect circuit, in which STN hyperactivity is a key characteristic, and hypoactivity in the direct circuit. Together, these actions result in increased GPi/SNr output inhibition of the thalamus and reduced activation of cortical and brainstem motor regions. This explains the motor symptoms of bradykinesia [Rodriguez-Oroz et al., 2009].

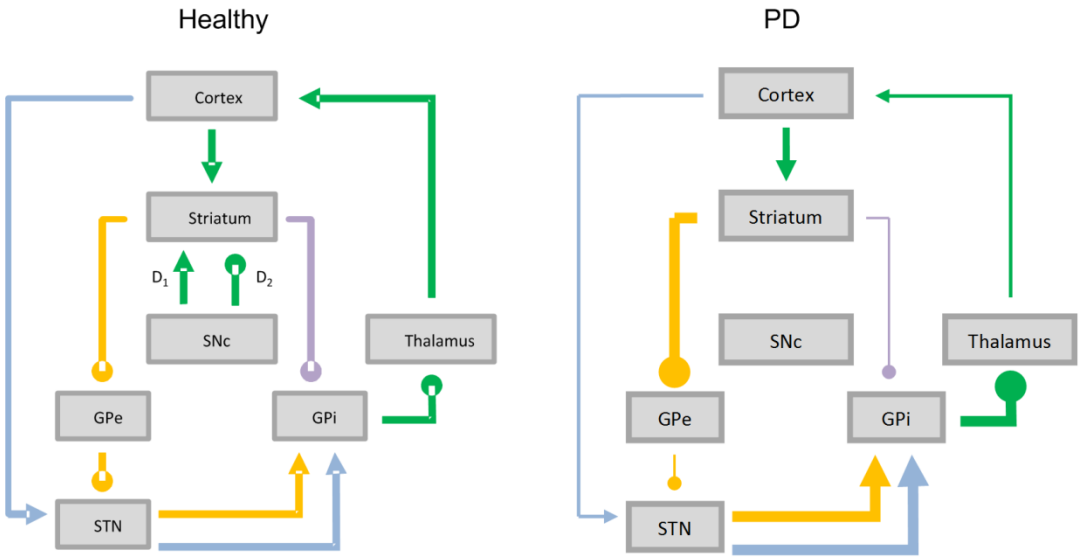



Figure 2.2. Diagram illustrating classical BG connectivity in healthy, and PD states. Arrows indicate excitatory connections; dots indicate inhibitory connections. Arrow width indicates relative strength of connections. Purple, direct pathway; orange, indirect pathway; blue, hyperdirect pathway; green, ubiquitous connection. In this model, PD would release the indirect pathway striatal inhibition and reduce direct pathway excitation. Both effects result in a hyperactive GPi, strongly inhibiting the thalamus and causing a reduction in cortical activity [Voytek, 2006].

PD rigidity is characterized by increased muscle tone to palpation at rest, reduced distension to passive movement, increased resistance to stretching, and facilitation of the shortening reaction [Magrinelli et al., 2016]. However, how these changes are associated with dopamine deficiency and BG output abnormalities, which are stipulated by the classical BG pathophysiological model, is still unclear [Magrinelli et al., 2016].

PD patients can show different tremor types. They include rest tremor, action tremor, and exaggerated physiological tremor [Magrinelli et al., 2016]. Some studies focused on the pathophysiology of rest tremor. The pathophysiology of rest tremor differs from that of bradykinesia and rigidity since rest tremor is not related to dopamine deficiency in the striatum [Magrinelli et al., 2016]. Several hypotheses have been suggested to explain the pathophysiology of rest tremor. One of the tested hypotheses is that resting tremor in PD emerges when dopamine depletion in the output layer of the BG circuitry alters striatopallidal

influences on a core cerebellothalmic circuit [Helmich et al., 2011]. Tremor-related responses have been observed both in the BG and in the cerebellothalmic circuit (ventral intermediate nucleus of the thalamus (VIM) and cerebellum (CBLM)), and interference with either circuit can effectively suppress resting tremor [Helmich et al., 2011]. Dopamine alterations in the pallidum increase the interactions between the BG and a distinct cerebellothalmic circuit, which will trigger hyperactivity in the cerebellothalmic circuit (VIM-MC-CBLM), leading to resting tremor (Figure 2.3).

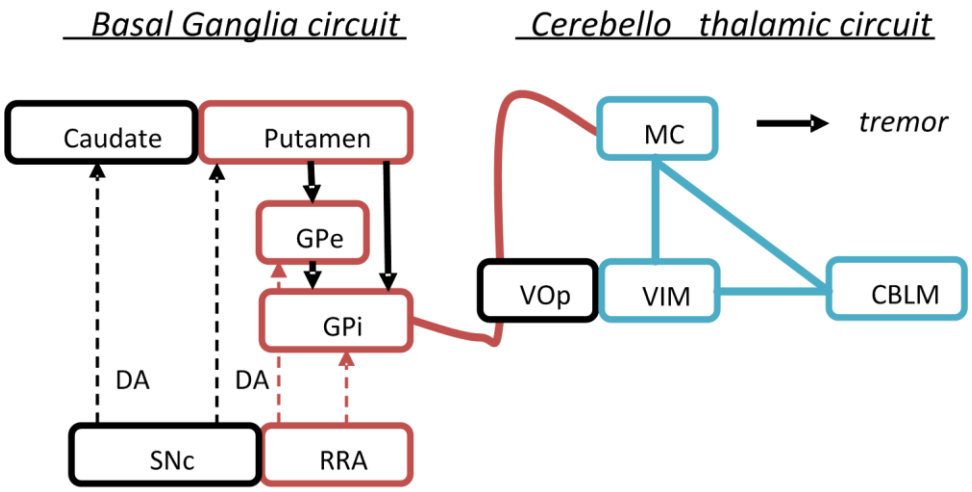



Figure 2.3. A model of cerebral mechanisms underlying PD resting tremor. PD resting tremor emerges from the ventral intermediate nucleus of the thalamus (VIM)–motor cortex (MC)–cerebellum (CBLM) circuit (in blue), when triggered by transient pathological signals from the BG motor loop (in red). In tremor-dominant PD, the BG (globus pallidus internus [GPi], globus pallidus externus [GPe], and putamen) have increased connectivity with the VIM–MC–CBLM circuit through the MC (thick red line). These alterations may be caused by loss of dopaminergic projections from the retrorubral area of the midbrain (RRA; in red) to the GPi and GPe. These alterations are different from the dopaminergic denervation of the striatum, observed across different PD subtypes and associated with bradykinesia and rigidity [Helmich et al., 2011].

## 2.2.1 Parkinson's disease progression

In PD, other areas of the brain beyond the substantia nigra are involved in disease's progression. Researchers have found that areas of the brain stem below the substantia nigra show cell loss and have been found to contain clumps of alpha-synuclein protein, which may form well before those in the substantia nigra [Port, 2019]. Braak et al. [Braak and Tredici, 2017], therefore, developing a staging system that characterizes PD's disease progression. This system is divided into six different stages, with each stage being attributed to abnormal pathology in particular neurological structures. A-synuclein-related pathology appears first in the olfactory bulb and brainstem regions and changes in these areas cause non-motor

symptoms, such as a lessened sense of smell or constipation. As the disease progresses, pathology spreads to the midbrain and basal forebrain cause motor symptoms before finally reaching the cortex where cognitive symptoms arise [Petersen, 2017]. This model suggests that the characteristic progression of PD symptoms corresponds to the regions infiltrated with Lewy pathology (Figure 2.4).

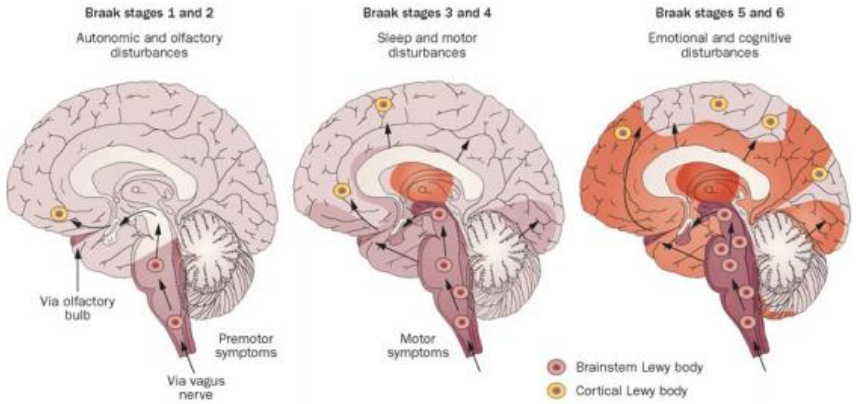



Figure 2.4. Illustration of Braak staging in PD [Petersen, 2017].

## 2.3 Etiology

Degenerative diseases are not usually caused by infections or metabolic disturbances or insufficient blood supply to the affected region (in this case, the substantia nigra). The cause of most neurodegenerative diseases remains unknown [Weiner et al., 2013]. However, three main factors have been shown to be revealed as influential in developing the disease, and they are as follows:

**Age:** The average age of onset of PD is about sixty years. Now, though, we cannot say that PD affects only the elderly, Nevertheless, about 80 percent of all people with PD develop the disorder between forty and seventy years of age, and about 5 percent develop it between twenty and forty [Weiner et al., 2013]. The relationship between normal aging and the course and features of PD is ardently debated [Weiner et al., 2013]. Some researchers argue that everyone would eventually develop PD if they lived long enough because, in the normal process of aging, the dopamine-containing neurons of the substantia nigra, so essential for normal motor function, become dysfunctional and degenerate. This position is not quite

supported by the current evidence, which indicates that normal aging alone does not cause PD. It is true that once the substantia nigra is damaged by whatever may cause PD, the normal age-related-loss of cells in the same area may contribute to the development of signs of the disease and possibly to its progression [Weiner et al., 2013].

**Genetic:** Approximately 15 percent of people with PD have a family history of this disorder. Familial cases of PD can be caused by mutations in the Leucine-Rich Repeat Kinase 2 (LRRK2), Parkinsonism associated deglycase (PARK7), PTEN induced kinase 1 (PINK1), and parkin RBR E3 ubiquitin protein ligase (PRKN) genes [Tran et al., 2020]. It is not fully understood how genetic changes cause PD or influence the risk of developing the disorder. Some gene mutations appear to disturb the cell machinery that breaks down (degrades) unwanted proteins in dopamine-producing neurons. As a result, undegraded proteins accumulate, leading to the impairment or death of these cells [Tran et al., 2020].

**Environmental factors:** Several studies link the disease to the herbicide and pesticide exposures and other toxic substances present in the drinking water or directly suspended in the air [Liou et al., 1997]. For this reason, some studies show a higher incidence of PD in rural and farming environments. However, these are not the only environmental factors. Literature review also suggests that the risk of suffering from PD is higher in subjects who have undergone a Traumatic brain Injury (TBI) [Liou et al. 1997]. The literature also suggests an increased risk linked with high exposure to metals. Protective factors include consumption of caffeine, regular tobacco smoking, consuming vitamin D, and physical activities [Liou et al., 1997].

## 2.4 Symptomatology

PD manifests in very heterogeneous symptoms, from which many are shared with other conditions. This is one of the reasons why a precise diagnosis is complicated and requires long periods of observations. As mentioned in section 2.2, dopamine is so important to the central nervous system control of muscles, that when this neurotransmitter is lost, muscle tone is altered [Weiner et al., 2013]. Rapid muscle tightening at inappropriate times and then muscle release produces tremor. Sometimes the muscles tighten up and become stiff and rigid. With inadequate communication between the brain and the muscles, movement

also becomes slow: muscles cannot make quick, fluid, spontaneous movements [Weiner et al., 2013]. The central mechanism that controls muscle tone does not function adequately for the delicate interplay of muscles required to help in standing, walking, and balancing. In addition, because PD affects the autonomic nervous system (ANS), (the involuntary nervous system that controls body temperature, digestive system, sexual function, and bladder control, among other functions), the functions of several ANS-regulated organ systems are affected [Weiner et al., 2013].

The characteristic symptoms of PD are distributed into motor and non-motor symptoms. Among the motor symptoms, the most common are [Weiner et al., 2013]:

1. Akinesia: a delay in the beginning of movements with long reaction times.
2. Bradykinesia: slowness of movements.
3. Hypokinesia: poor, incomplete or simplified movements. In most cases, the limbs involved do not reach the full extent associated with the action of purpose of the movement. As an example, handwriting can become smaller.
4. Rigidity: rigidity of muscles, causing pain and hampering certain actions and postures.
5. Tremor: it can be postural or resting tremor. In the same manner, it can be constant or intermitted. In some cases, it consists of an internal tremor, not visible but quite disturbing, for the patient.
6. Postural instability: involving balance problems in the standing position or walking.
7. Dystonia: involuntary repetitive twisting and sustained muscle contractions.

Regarding the non-motor symptoms, the most common are [Weiner et al., 2013]:

1. Depression and anxiety. It can involve the inability of the patient to express emotions.
2. Constipation.
3. Loss of smell.
4. Communications problems.
5. Dementia and other cognitive problems such as hallucinations or difficulties on focusing and performing complex tasks.
6. Sleeping problems.

7. Sexual disruptions.

In PD, the first symptoms differ only slightly from the normal state and progress slowly, perhaps over decades. These symptoms can manifest in varying degrees and combinations with different individuals [Weiner et al., 2013].

## 2.5 Diagnosis and rating scales

An early detection of PD might improve and maintain the patients' quality of life and increase their life expectancy. However, the fact is that diagnosing PD based on very early symptoms is seldom possible. Furthermore, there are no efficient and reliable methods capable of achieving PD early diagnosis with certainty [Weiner et al., 2013]. Because the initial symptoms can be very subtle and may be overlooked or mistaken for other medical problems, a firm diagnosis is difficult in the early stages. The only way to diagnose accurately is to wait and see whether the typical Parkinson's symptoms get worse [Weiner et al., 2013]. A physician who diagnoses PD should have expertise in neurologic diagnosis. As the symptoms of PD increase in severity, a definitive diagnosis can usually be made and medications can be prescribed to help the individual continue functioning [Weiner et al., 2013].

No definitive laboratory test or radiologic study is available for diagnosing PD, but some tests or imaging scans can be used to rule out PD, such as genetic or blood tests that indicate the existence of Huntington's disease or Wilson's disease that produce certain similar symptoms with PD [Weiner et al., 2013]. Imaging scans (such as Magnetic resonance imaging (MRI) and computerized axial tomography (CAT) scans) can be useful in helping physicians rule out other causes of the symptoms. But the MRI and CAT scans of the brain of people with PD appear normal. The brain changes that create PD are microscopic, on a chemical level, and are not revealed by these scans [Weiner et al., 2013]. Certain types of positron emission tomography (PET) scan and single photon emission computed tomography (SPECT) scan are used to assess the dopamine system in the brain. However, these two scans are not widely available and are very expensive. In addition, the findings from these imaging techniques are generally similar for PD and for other causes of Parkinsonism (any person who has the signs and symptoms of PD but not necessarily has PD), so they are not a way of

narrowing the diagnosis [Weiner et al., 2013]. Thus, the diagnosis must be based on the clinical judgment of the physician, piecing together historical clues and findings from a thorough physical examination.

**Clinical diagnosis:** when performing a neurologic examination to evaluate a patient with a movement disorder, the physician takes a history and performs a physical examination. First of all, the doctor asks the patient and the family members about symptoms, because family members may notice the initial Parkinson's tremor even before the affected person does, and they may also spot the characteristic Parkinson's posture and gait. Then, the physician observes the patient's finger, hand, and foot movements and observes the patient walk, turn, and resist postural challenges (pushing the patient to see if he or she can avoid falling). The physician also observes the patient to see whether any abnormal movements (e.g., tremor) are visible and whether the patient can easily arise from a chair. Facial expression, eye movements, and speech are also examined. The physician flexes and extends the patient's neck, arms, wrists, and legs to search for abnormal muscle tone. Strength and coordination of the arms are evaluated [Weiner et al., 2013]. The cognitive (mental) function of the patient is also assessed. The mini mental state examination (MMSE) is recommended to measure cognitive impairment [Kurlowicz and Wallace, 1999]. It is also used to estimate the severity and progression of cognitive impairment and to follow the course of cognitive changes in an individual over time. The MMSE is a 30 points questionnaire, including registration (repeating named prompts), attention and calculation, recall, language, ability to follow simple commands and orientation. Any score of 24 or more (out of 30) indicates a normal cognition. Below this, scores can indicate severe (≤9 points), moderate (10–18 points) or mild (19–23 points) cognitive impairment.

Beside the MMSE, several different rating scales may be used to assess the stage and the progression of PD in an individual. The two most widely used are the Hoehn and Yahr scale (H&Y) and the unified Parkinson's disease rating scale (UPDRS) [Perlmutter, 2009]. The UPDRS comprises four main parts: I, intellectual function, behavior and mood; II, activities of daily living; III, motor examination; IV, motor complications. The rating according to this scale is accomplished through interviews with the patient and through clinical observations. The best possible score for each part is 0 whereas the worst one depends on the part: 16 for part I, 52 for part II, 56 for part III and 23 for part IV. Therefore,

the global UPDRS can range between 0 and 147, where it can be claimed that, the larger the value, the higher the affection of the disease. Nevertheless, two patients with the same global rate can exhibit different signs [Perlmutter, 2009].

The H&Y scale comprises several stages whose values can range from 1 to 5. The value of 1 implies that the patient has low or no functional disabilities and 5 imply that patient is totally dependent. This scale was modified to include three intermediate levels (0, 1.5 and 2.5) to account for the intermediate course of the disease as listed [Perlmutter, 2009]:

1. Stage 0: no signs of disease.
2. Stage 1: unilateral disease.
3. Stage 1.5: unilateral plus axial involvement.
4. Stage 2: bilateral disease, without impairment of balance.
5. Stage 2.5: mild bilateral disease, with recovery on pull test.
6. Stage 3: mild to moderate bilateral disease; some postural instability; physically independent.
7. Stage 4: severe disability; still able to walk or stand unassisted.
8. Stage 5: wheelchair bound or bedridden unless aided.

## 2.6 Treatment

Once a subject is diagnosed with PD, it is possible to successfully manage PD symptoms through healthy lifestyle choices, medications, and, in some cases, surgery. Treatment needs change over the course of the disease [Golbe et al., 2012]. At every stage, it is important for the patient to maintain physical activity, eat a healthy diet, and monitor his/her mental health. Early treatment with medication can help most people with PD maintain an active lifestyle and continue working. Medications are adjusted throughout the course of the disease, in order to maintain the best control of symptoms and avoid major side effects. Adjusting medications can be complex and is one of the best reasons to be seen by a movement disorders specialist [Golbe et al., 2012].

## 2.6.1 Medications for motor symptoms

There are several classes of medications available for the successful treatment of motor symptoms throughout the course of PD. Some medications work better than others for specific symptoms of PD [Golbe et al., 2012].

1. Carbidopa-levodopa*:* The most effective treatment for PD is the combination medication of carbidopa-levodopa which is intended to increase brain levels of dopamine, that are deficient in people with PD. Levodopa, which is converted to dopamine in the brain, increases brain levels of dopamine. Levodopa reduces tremor, stiffness, and slow movement in people with idiopathic PD. Carbidopa prevents levodopa from being broken down in the body before it reaches the brain. Therefore, the addition of carbidopa allows levodopa to get into the brain more efficiently. Side effects of carbidopa-levodopa treatment include nausea, orthostatic hypotension, sleepiness (which can be sudden, called sleep attacks), impulse control behaviors, hallucinations, and confusion. Levodopa also contributes to the development of dyskinesia (uncontrolled movements) [Golbe et al., 2012].

2. Carbidopa-levodopa Infusion: An alternative form of carbidopa-levodopa was approved by the food and drug administration (FDA) in 2015. It is intended for people with more advanced disease, whose symptoms are no longer responding well to oral carbidopa-levodopa. Instead of taking a pill, people with PD can receive carbidopa-levodopa in a gel form through an infusion pump. The pump delivers the medication directly into the small intestine through a surgically placed tube. The advantage of a continuous infusion of the carbidopa-levodopa is less immobility or "off" time (when medications are not sufficient to maintain good symptom control throughout the day) from levodopa. The side effects of the carbidopa-levodopa infusion are similar to those of oral carbidopa-levodopa, but may be associated with a higher incidence of peripheral neuropathy (numbness or loss of sensation in the fingers or feet) [Golbe et al., 2012].

3. Dopamine Agonists: are a little different from carbidopa-levodopa in that, instead of increasing dopamine levels in the brain, they mimic the activity of dopamine. They can be given alone in the early stages of PD, or as an adjunct to carbidopa-levodopa or other PD medications. The side effects of dopamine agonists are similar to those of carbidopa-levodopa, although impulse control disorders and sudden onset of sleepiness can be more pronounced [Golbe et al., 2012].

4. Catechol-O-Methyltransferase (COMT) Inhibitors: are sometimes used with carbidopa-levodopa. Like carbidopa, they prevent the breakdown of levodopa before it reaches the brain. The result is that a more reliable supply of levodopa enters the brain, where it can be converted to dopamine. COMT inhibitors are typically prescribed to treat frequent "off" times with levodopa therapy. Sometimes COMT inhibitors can increase the side effects associated with levodopa therapy. Other common side effects of COMT inhibitors are abdominal pain, diarrhea, and discolored bodily fluids such as urine [Golbe et al., 2012].

5. Selective Monoamine oxidase B (MAO-B) Inhibitors: block the MAO-B enzyme in the brain, which breaks down dopamine. This is another way to increase dopamine levels in the brain. MAO-B inhibitors can be used alone or with other PD medications. Selective MAO-B inhibitors may be prescribed to complement carbidopa-levodopa therapy, particularly if individuals experience "wearing-off" symptoms while taking levodopa. Side effects of selective MAO-B inhibitors include mild nausea, dry mouth, lightheadedness, constipation, and, occasionally, hallucinations and confusion [Golbe et al., 2012].

6. Anticholinergics: are often used for the management of PD as adjunct medications to other PD therapies. Anticholinergics are frequently prescribed to reduce the characteristic tremor of PD or to ease the problems associated with the wearing-off of levodopa therapy. Common side effects of anticholinergics include confusion, hallucinations, constipation, dry mouth, and urinary problems. As a result, the use of anticholinergics is typically limited to younger people with PD (under the age of 70). These anticholinergics should also be avoided in

combination with antihistamines, certain psychiatric drugs, and alcohol [Golbe et al., 2012].

7. Amantadine Formulations: has been observed to ease the tremor of PD as well as muscle rigidity. It is typically used as an adjunct medication to other therapies for PD. In addition, it is used to decrease dyskinesia or involuntary movements caused by levodopa and to reduce "off" time. Common side effects include lightheadedness, dry mouth, constipation, vivid dreams, lacy rash, typically on the legs, and swelling of the ankles. It may also interact with or enhance the side effects of anticholinergics and levodopa therapy [Golbe et al., 2012].

8. Adenosine inhibitors: block the effects of the adenosine receptor. Like dopamine, adenosine is a neurotransmitter that works in the BG, the deep structures of the brain that are affected in PD. However, to some degree, adenosine and dopamine have opposite effects, so that *inhibiting* the adenosine receptor improves motor function. It is indicated for use as an add-on treatment to carbidopa-levodopa for those experiencing "off" episodes [Golbe et al., 2012].

### 2.6.2 Surgery treatment

**Deep brain stimulation (DBS):** is a neurosurgical procedure for people with advanced PD who retain a good response to levodopa, but who have developed significant motor fluctuations including dyskinesia. DBS may also be used to treat medication resistant tremor. By stimulating specific points in the motor control circuits in the brain, DBS "rebalances" the circuits, restoring normal movement control to some degree. In most cases, this allows the person with PD to reduce their dosage of levodopa, and thus reduce their dyskinesia, while maintaining good symptom control. DBS involves the implantation of permanent, thin electrodes into selected deep parts of the brain. A battery-operated pulse generator, much like a cardiac pacemaker, is implanted under the skin of the chest or abdomen. The pulse generator is connected to the stimulator electrodes via wires, which are tunneled underneath the skin of the scalp and neck as shown in Figure 2.5. The DBS procedure is associated with a small chance of infection, stroke, bleeding, or complications associated with anesthesia [Golbe et al., 2012]. The equipment is not visible underneath the

cloth and causes no discomfort in daily use. The electrodes are programmed by a remote computer for maximum symptom control, and the batteries can be replaced by an outpatient procedure when necessary, typically after several years.

DBS technology is constantly evolving with the development of rechargeable batteries which last much longer than a traditional battery, as well as more sophisticated programming options to maximize symptom control. The newer DBS systems are also MRI-conditional and can be placed in MRI-mode for the duration of an MRI, making the process of getting an MRI with a DBS system in place much simpler than with older systems [Golbe et al., 2012].

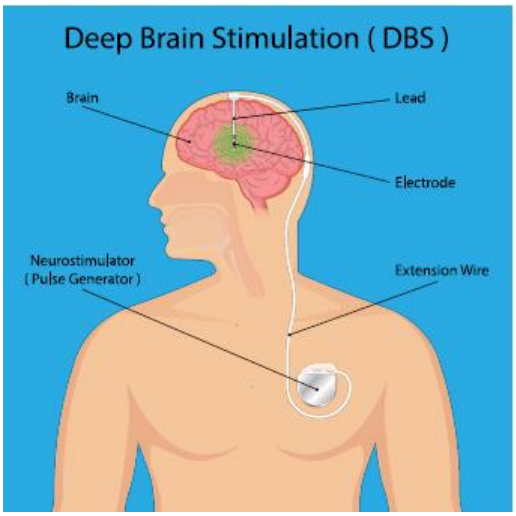



Figure 2.5. Deep Brain Stimulation [Golbe et al., 2012].

**Stem cell transplantations:** Current available treatments to treat Parkinson's focus on increasing dopamine levels in the brain, which alleviates motor symptoms. However, these treatments' efficiency declines over time and are associated with various side effects [Henchcliffe and Parmar, 2018]. Because medications only go part way to fully treat incoordination and movement problems, a possible long-term treatment for Parkinson's is transplanting dopamine-producing stem cells into patients' brains. This approach could lead to long-term relief of motor symptoms and can reduce or even stop the need for medication. Using stem cells made in the lab for transplants would allow the production of enough cells to cover the current demand, ensure the number and type of transplanted cells are always the same, and that they produce the desired amounts of dopamine, and that the risk rejection is

low. Cells used to perform the transplants are extracted from a patient's blood or skin, and are treated to become almost any type of cell in the body [Henchcliffe and Parmar, 2018].

## 2.7 Parkinson's disease and handwriting

As mentioned previously, PD is caused by the degeneration of nigrostriatal neurons resulting in a reduction of the neurotransmitter dopamine [Weiner et al., 2013]. Thus, PD patients are expected to suffer from difficulties in the coordination and control of various muscle systems [Flash et al., 1992]. For example, PD patients show a delay in the onset of the opening of the hand relative to the initiation of the transportation of the forearm [Castiello et al., 1994]. Temporal dissociation has also been observed between the left and the right arms in PD patients [Lazarus and Stelmach, 1992]. In addition, some researchers have found that PD patients show a substantial impairment when performing an isotonic elbow flexion while isometrically squeezing a force transducer [Benecke et al., 1986]. Similarly, others have observed that PD patients do not initiate components of arm movements simultaneously, resulting in angular or curved movement trajectories [Isenberg and Conrad, 1994]. These data suggest that in PD patients, coordination is reduced in movement patterns that require control of a large number of muscles and joints.

PD leads to a disruption in the execution of practiced skills such as handwriting. In [Boisseau et al., 1986], the authors observed that PD handwriting can be characterized by various types of dysfluencies: lack of control, abrupt changes of direction, tremor, slowness, hesitation, rigidity, variability of baseline, and, in some cases micrographia. From the other side, fingers, wrist, and arm generate specific components of the writing movements [Teulings et al., 1997]. The fingers produce the vertical movement as in up-down strokes, toward and away from the body in the horizontal plane. The wrist produces the local horizontal movement as in left-right strokes. The forearm produces the left-to-right horizontal progression as in extended horizontal lines of writing [Teulings et al., 1986]. Teulings et al. [Teulings et al., 1986] found that PD subjects show reduced capability to coordinate the wrist and fingers in handwriting tasks, and wrist flexion showed more irregular acceleration profiles than wrist extension. Therefore, they concluded that coordination impairments in PD patients can be detected in finger and wrist movements, and even in flexions and ulnar deviations of the wrist which may contribute to the handwriting impairments observed in PD

patients [Teulings et al., 1986]. In conclusion, handwritings are considered ideal to study motor control and to detect PD.

## 2.8 Parkinson's disease and speech

Speech requires the integrity and integration of numerous neurocognitive, neuromotor, neuromuscular, and musculoskeletal activities [Duffy, 2013]. These activities can be summarized as follows:

1. When thoughts, feelings, and emotions generate intent to communicate verbally, they must be organized and converted into a code that abides by the rules of language. These combined activities are referred to as cognitive-linguistic processes [Duffy, 2013].

2. The intended verbal message must be organized for neuromuscular execution. These activities include the selection, sequencing, and regulation of sensorimotor "programs" that activate speech muscles at appropriate co-articulated times, durations, and intensities. These combined activities are referred to as motor speech planning, programming, and control [Duffy, 2013].

3. Central and peripheral nervous system activity must combine to execute speech motor programs by innervating breathing, phonatory, resonatory, and articulatory muscles in a manner that generates an acoustic signal that faithfully reflects the goals of the programs. The neural and neuromuscular transmission and subsequent muscle contractions and movements of speech structures are referred to as neuromuscular execution [Duffy, 2013].

The combined processes of speech motor planning, programming, control and execution are referred to as motor speech processes. When the nervous system becomes disordered, so may the production of speech. In fact, changes in speech may be a harbinger of neurologic disease. The speech disorders associated with PD are termed *hypokinetic dysarthria* and lead to reduced speech intelligibility. Dysarthria reflect neuromuscular disturbances of strength, speed, tone, steadiness, or accuracy of the movements that underlie the execution of speech. They also reflect disturbances at any or a combination of the major

components of the speech mechanism, including respiration, phonation, resonance, articulation, and prosody [Duffy, 2013]. More detailed information about each speech related organ and disturbances reflected on them due to PD are presented in chapter 6.

## 2.9 Parkinson's disease and eye movements

As mentioned in section 2.2, PD motor manifestations are caused by dopaminergic cell loss within the SNc, resulting in dysfunction of the BG. Movement disorders such as PD have been traced to BG dysfunction. Consequently, the BG is traditionally classified as part of the extrapyramidal motor system. PD is characterized by the loss of dopaminergic innervation of the striatum from the SNc [Voytek, 2006]. Brain areas that are involved in controlling visual fixation and saccadic eye movements include regions within the cerebral cortex, BG, thalamus, superior colliculus (SC), brainstem reticular formation, and cerebellum [Coe and Munoz, 2017]. SC receives inputs from retina as well from BG and cortex. It plays a major role in fixation and saccadic eye movements [Goyal et al., 2014]. Figure 2.6 shows the neural control of eye movements. Visual inputs to the system arise from the retino-geniculo-cortical pathway to primary visual cortex and from the retinotectal pathway to the superficial layers of the SC (SCs) [Coe and Munoz, 2017]. Visual information is processed through several extrastriate visual areas before it impinges on motor structures to effect action [Coe and Munoz, 2017]. The parietal eye field (PEF) in the parietal cortex is one area at the interface between sensory and motor processing. PEF projects to both the intermediate layers of the SC (SCi) and frontal cortical oculomotor areas including the frontal eye fields (FEF), the supplementary eye fields (SEF), and the dorsolateral prefrontal cortex (DLPFC) [Coe and Munoz, 2017]. The FEF play a critical role in executing voluntary saccades. The SEF play an important role in internally guided decision-making and sequencing of saccades. The DLPFC is involved in 'domain-general' functions (i.e. improvements in individual functions also improve other related functions) such as executive function, spatial working memory, and suppressing automated or reflexive responses [Coe and Munoz, 2017]. These frontal oculomotor regions project to the SCi, which is a critical node in the premotor circuit where cortical and subcortical signals converge and are integrated. The SCi projects directly to the premotor circuit in the brainstem reticular formation to provide the necessary input to guide saccades [Coe and Munoz, 2017]. There are also important pathways through the BG. Frontal

cortical oculomotor areas project to the caudate nucleus (CN). Via the direct pathway, GABAergic neurons in the CN project directly to the SNr. Neurons in the SNr form the major output of the BG circuit: they are GABAergic and they project to the SCi and nuclei in the thalamus that project to frontal cortex. Cortical inputs to the direct pathway lead to disinhibition of the SC and thalamus because these signals pass through two inhibitory synapses. There is also an indirect pathway through the BG, in which a separate set of GABAergic neurons in the CN project to the external segment of the GPe. GABAergic neurons in GPe then project to the STN. Neurons in the STN send excitatory projections to the GPe, which then projects via GABAergic projections to the SNr. There is also a hyperdirect pathway in which regions of cerebral cortex project to the STN, which then projects directly to SNr. These complex sets of excitatory and inhibitory projections within the BG provide a rich set of control mechanisms to help guide voluntary behavior [Coe and Munoz, 2017]. As mentioned in section 2.2, the degeneration of dopaminergic neurons in the SNc leads to increased activity in the indirect circuit and to an increased inhibitory output from BG, which further will leads to inhibition of SC and thalamus. This inhibition of SC helps in the suppression of saccadic eye movements, which is considered to be responsible for the typical eye movements' abnormalities in PD [Terao et al., 2013].

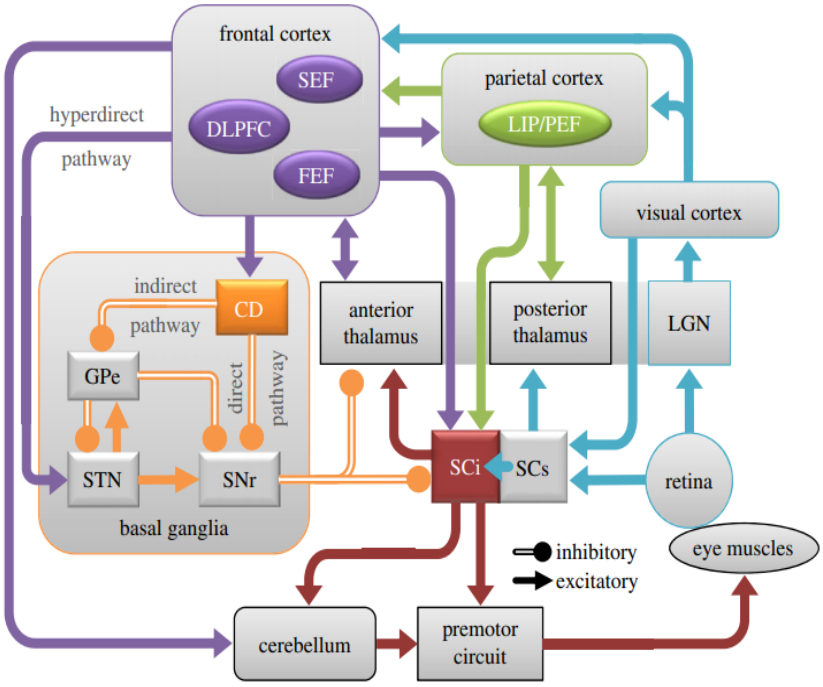



Figure 2.6. The neural control of eye movements [Coe and Munoz, 2017].

# 3 State of Art

This chapter provides a review of the different methodologies existing in literature that are related to the analysis and characterization of handwriting, speech, and eye movements in Parkinson patients for the automatic detection of PD by using these signals, and also highlights the missing analyses that should be addressed.

## 3.1 Handwriting and Parkinson's disease

PD is a long-term degenerative disorder of the central nervous system that mainly affects the motor system. Handwriting involves several cognitive abilities. For this reason, handwriting analysis can be used as an effective tool for early diagnoses of PD [Rosenblum et al. 2013]. Many studies have been proposed that use handwriting for detecting and monitoring PD, since abnormal handwriting is a well-recognized manifestation of PD. Handwriting anomalies may appear years before at the early stages of the disease and thus may be one of the first signs of PD.

### 3.1.1 Influence of PD and L-dopa medication on handwriting

Previous study [McLennan et al., 1972] founds that micrographia or small writing is the most commonly observed handwriting abnormality in patients with PD due to the inability to properly control wrist and finger movements; which will lead to a difficulty in maintaining a constant force: this could explain the reduction in handwriting size [Dounskaia et al., 2009], [Teulings et al., 1997]. Moreover, according to McLennan et al. [McLennan et al., 1972], in approximately 5 % of PD patients, micrographia may be observed even before the onset of the cardinal motor symptoms. Micrographia can be detected with conventional paper-and-pencil tools (offline analysis).

The recent advantages of new technologies enables the acquisition of online handwriting signals, where temporal information is added to the X and Y position beside the pressure over the writing surface and measures of pen inclination and orientation. Moreover, these devices can capture pen movement not only when the pen is in contact with the writing surface, but also when the pen is "in-air". Therefore, by using a digitizing tablet, hidden

features of handwriting which are not visible in the written trace can be determined and the analysis is not limited to spatial features which mainly quantify PD micrographia. We are able to quantify temporal, kinematic, and dynamic manifestations of PD dysgraphia; where PD dysgraphia refers to the motor impairments related to the disease such as slow movement, muscle rigidity, and tremor, which cannot be studied objectively using a classical paper-and-pen method [Letanneux et al., 2014].

Several research teams have explored the impact of quantitative PD dysgraphia analysis utilizing simple handwriting/drawing tasks (separate characters, repetitive loops, circles, words, sentences, figures, and the Archimedean spiral). Several online handwriting databases exist for this purpose. The most consistent ones that are publically available are summarized in Table 3.1, where none of them is multilingual.

The Parkinson's Disease Handwriting database (PaHaW) consists of handwriting samples collected from 37 PD patients (after taking their L-dopa medication) and 38 HC subjects that are aged and gender matched [Drotar et al., 2013]. PaHaW includes eight different handwriting tasks (spiral drawing and words written in Czech). Handwritten dynamics were captured by Wacom Intuos 4 digitizing tablet.

The original HandPD database contains handwriting samples collected from 74 PD patients (at early stages) and 18 HC subjects. This data includes spiral and meander drawings. The new extended version of HandPD is called NewHandPD and contains handwriting data from 31 PD patients and 35 HC subjects. NewHandPD includes spiral and meander drawings. The handwritten dynamics were captured by a biosensor smart pen (BiSP) [Pereira et al., 2016].

The ParkinsonHW consists of 62 PD patients (after taking their L-dopa medication) and 15 HC subjects. Three handwriting tasks were considered: the static spiral test (SST), the dynamic spiral test (DST), and the stability test on a certain point (STCP). In the SST test, subject is asked to retrace the wound Archimedean spirals displayed on the tablet screen, where in the DST test, the Archimedean spiral appears and disappears at certain time intervals. For the STCP test, a red point is displayed in the middle of the screen and subjects are asked to hold the pen on the point without touching the screen [Isenkul et al., 2014].

Table 3.1. The most consistent Handwriting datasets that are publically available.

| Dataset Name | Size | Acquisition device | Tasks |
|---|---|---|---|
| PaHaW | 37 PD<br>38 HC | Wacom Intuos 4M | -Spiral drawing<br>-Repetition of *“l”, “le”, “les”, “ lektorka ”, “ porovnat ”, “ nepopadnout ”, “ Tramvaj dnes už nepo-jede”* |
| HandPD | 74 PD<br>18 HC | BiSP | Spiral and meander drawing |
| NewHandPD | 31 PD<br>35 HC | BiSP | Spiral and meander drawing |
| ParkinsonHW | 62 PD<br>15 HC | Wacom Cintip 12 WX | Static Spiral drawing, Dynamic Spiral drawing, and stability test |

Margolin et al. [Margolin and Wing, 1983], Phillips et al. [Phillips et al., 1991], and Teulings et al. [Teulings and Stelmach, 1991] were among the first to use graphic tablets to assess handwriting in PD. Phillips et al. [Phillips et al., 1991] adopted a simple zig-zag drawing. Results revealed that patients had more difficulties in producing smooth movements rather than in controlling stroke length or duration. This result is confirmed by the one obtained in [Teulings and Stelmach, 1991], where users were asked to produce handwriting modifying speed and dimension.

With the availability of graphics tablets, kinematic and size features have now been investigated and some studies have shown that pen-tip pressure differ between PD patients and HC subjects [Drotár et al., 2016], [Rosenblum et al., 2013] due to rigidity and bradykinesia (unable to retain Pen-pressure). From the other side, Raudmann et al. [Raudmann et al., 2014] found that kinematic features are more affected than size features by PD.

Also many studies have been investigated in order to see the effects of L-dopa medication on handwriting. Some researchers did not find any correlation between improvement in handwriting and medication [Eichhorn et al., 1996], whereas some others found that medication improved the kinematic features [Poluha et al., 1998], [Tucha et al., 2006]. However, this improvement is less effective in case patient is asked to perform many tasks simultaneously.

### 3.1.2 Hand-crafted features and classical classifier for PD detection based on handwriting

Several works proposed automatic systems to detect the PD using handwriting analysis. Drotar et al. [Drotar et al., 2016] extracted global kinematic and pressure features from handwriting samples and used the SVM classifier. They distinguished the PD patients from healthy subjects with an accuracy of 81.3 %. After that, Drotar et al. [Drotar et al., 2015b] proposed to combine global kinematic, entropy and intrinsic features together with an SVM classifier, and they found that this combination return an 88.13 % classification accuracy. To improve these results, the same authors extract also the pressure features beside kinematic, entropy, and intrinsic features ending with an 89 % accuracy [Drotar et al., 2015-a]. The handwriting samples in Drotar et al. studies have been taken from PaHaW database, where the features were extracted only when the pen was touching the paper (or on-paper state).

Also some works studied both in-air and on-surface features. Drotar et al. [Drotar et al., 2013], [Drotar et al., 2014] used global kinematic features based on on-paper and in-air movements with an SVM for PD classification, where Mucha et al. [Mucha et al., 2018] combine global kinematic and temporal features that are extracted for both “on-paper” and “in-air” and apply an SVM for classification. It was found that in-air global features set seems to outperform an “on-surface” one, and combining both features can achieve better results in PD detection. Also the handwriting samples in these studies were taken from PaHaW database.

Also classification of different handwriting tasks was performed with the aim of finding the most effective method and task. Drotar et al. [Drotar et al., 2015-b] found that classification accuracy depends on the choice of handwriting tasks; where some tasks are better representative at symptomatic signs than others.

### 3.1.3 Deep learning for PD detection based on handwriting

Most of the studies mentioned above applied hand-crafted features models for PD classification. Recently, some studies applied deep learning techniques to learn proper

information about the problem instead of hand-crafting features, since deep learning shows performances in areas such as speech recognition, and image classification. Pereira et al. [Pereira et al., 2018] propose to study each handwriting task separately by encoding the whole set of handwritten dynamic measurements of a given task into a 2D representation. The image corresponding to the 6 time series captured by a smart pen (microphone sound, finger grip, axial pressure, tilt and acceleration in X, Y, and Z directions) and taken from HandPD database is analyzed by a convolutional neural network (CNN) for PD detection. The predictions of all the tasks are combined using the majority voting. Two different CNN architectures were proposed: the CNN-ImageNet and the CNN-Cifar-10 represented in Figure 3.1, where two different input image sizes were studied (64×64 and 128×128). The authors did not employ transfer learning since the images used are domain-specific. Six different handwriting tasks were studied separately, and at a later stage majority voting is applied to get the final decision. This approach returned an accuracy of 93.42 % when applying the CNN-ImageNet model with 64×64 input image size, and 95.74 % with 128×128 input image size.

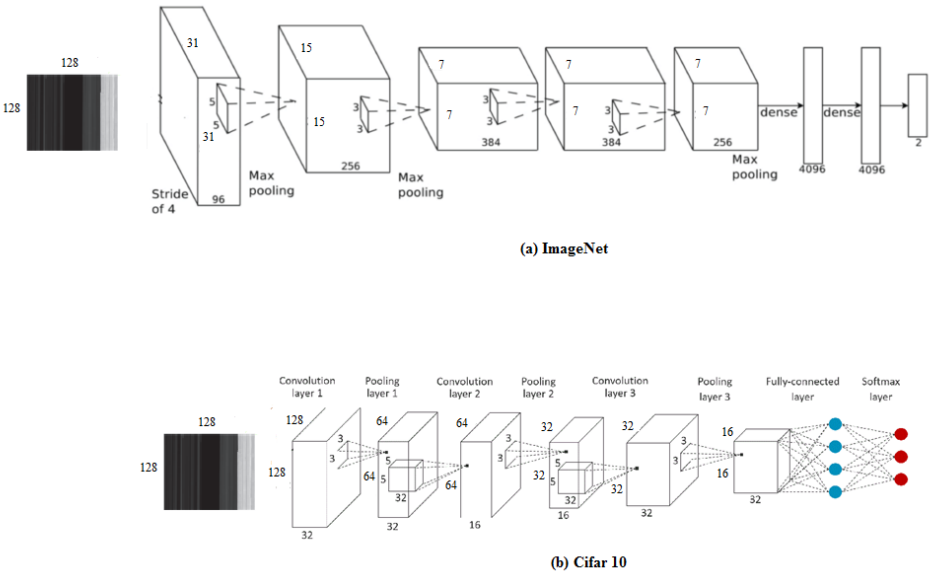



Figure 3.1. The two CNN models studied in [Pereira et al., 2018]: (a) ImageNet [Jeremy, 2018-a], and (b) Cifar-10 [Jeremy, 2018-a].

Moetesum et al. [Moetesum et al., 2019] exploited the static visual attributes of handwriting (a visual image of the drawing obtained by plotting the on-surface (X, Y) coordinates), taken from PaHaW database, to predict PD using CNNs to extract separate

discriminating visual features from the three representations of the input data (the raw image, median filter residual image, and edge image). A pre-trained CNN-Alexnet (Figure 3.1-a) was used for feature extraction purpose. The features extracted from the 3 CNN networks are combined and fed into an SVM for classification. This procedure is repeated for each of the 8 tasks. The predictions of the 8 tasks are combined using the majority voting. The system overview returning an accuracy of 83 % is represented in Figure 3.2.

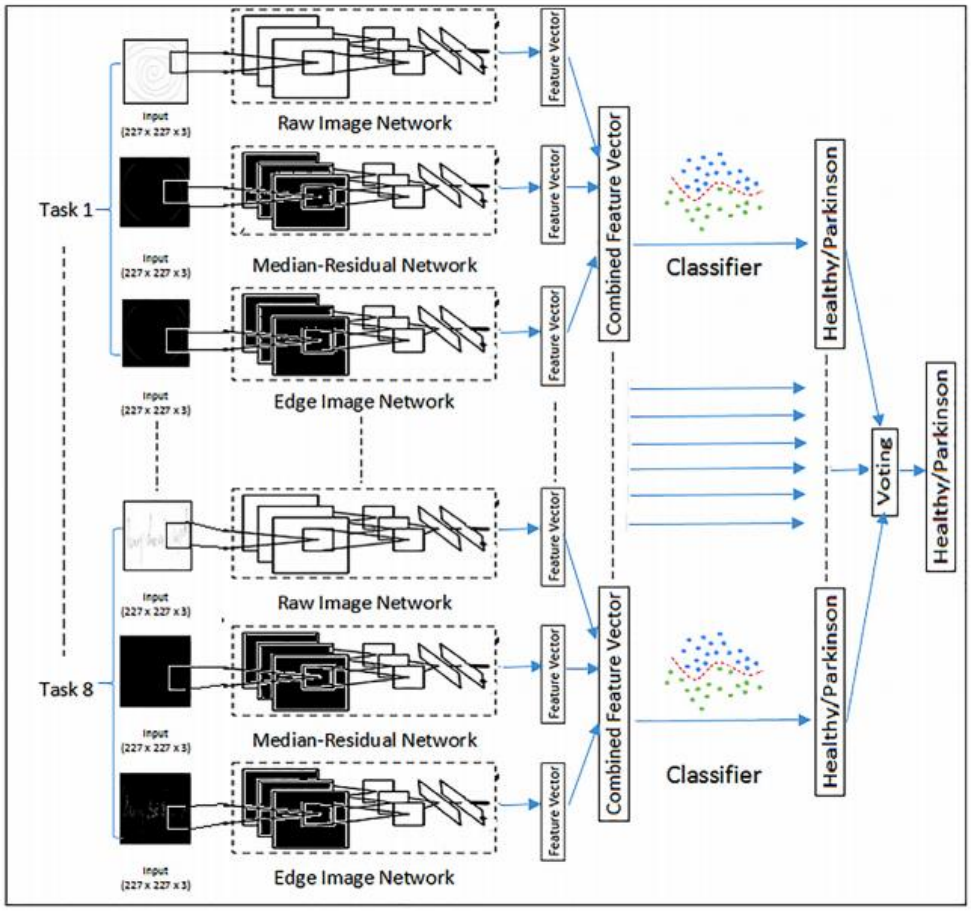



Figure 3.2. Overview of the deep model proposed in [Moetesum et al., 2019].

Khatamino et al. in [Khatamino et al., 2018] applied a CNN model (summarized in Figure 3.3) for PD classification from SST and DST handwriting tests taken from ParkinsonHW database. This CNN architecture included two convolutional, each followed by rectified linear unit (ReLU) activation function, and two maxpooling layers, followed by two fully connected layers. The hidden layer is composed of 128 nodes and ReLU activation function, where the output layer is composed of 2 nodes with softmax activation function. All the convolutional layers employ kernels of size 3×3 with stride of 1 pixel, and the

maxpooling operations are applied on regions of size 2x2, with stride 2. Both dynamic extracted signal-based images and visual attributes of spirals are studied and compared, where the extracted signal-based image is the one proposed by Pereira in [Pereira et al., 2018], and the visual attributes of spirals proposed by Moetesum et al. [Moetesum et al., 2019]. The image size applied is 128×128 pixels. It was found that DST tests predict PD with an accuracy of 88 % just by using visual attributes of spiral or extracted signal-based image.

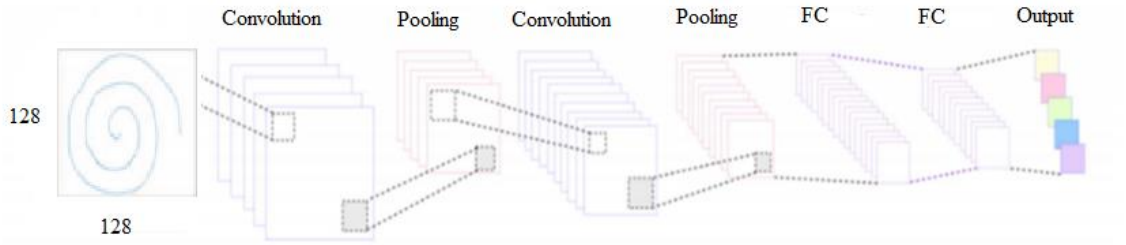



Figure 3.3. A schematic of CNN Architecture used in [Khatamino et al., 2018].

All the models in [Pereira et al., 2018], [Moetesum et al., 2019], and [Khatamino et al., 2018] require a fixed dimension input images, and the variation over the time axis defines a challenge for such models.

From the other side, Gallicchio et al. [Gallicchio et al., 2018] have proposed a novel approach for diagnosis of PD based on Deep Echo State Networks (DeepESN) model where the deep recurrent model is fed by the whole time-series captured from a tablet during the sketching of spiral tests (taken from ParkinsonHW database). The 10 layers DeepESN architecture is shown in Figure 3.4; where the number of layers $N_L$ =10. This model predicts PD with an accuracy of 89.3 %. However, we believe that analyzing handwriting signals without feature extraction can be challenging especially when the signals are noisy.

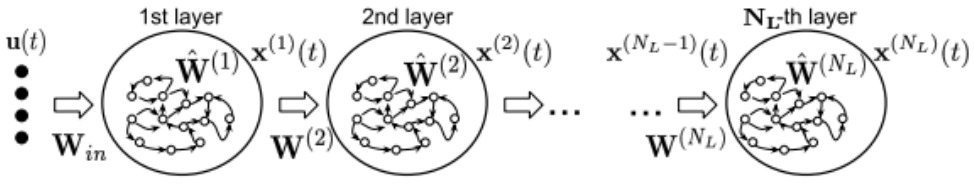



Figure 3.4. The hierarchy of reservoirs in the architecture of a DeepESN [Gallicchio et al., 2018].

## 3.2 Speech and Parkinson's disease

Impairment in voice and speech are related to the rigidity and bradykinesia caused by PD which produce uneven vocal folds vibration, incomplete folds closure, and undershooting of articulatory gestures due to the decreased range of motion of the supra-laryngeal muscles as well as in the respiratory muscles [Hunker et al., 1982], [Rusz et al., 2011-b]. Literature shows a certain number of studies trying to identify the deficiencies caused by PD in the voice or speech of patients or to propose the creation of new biomarkers or automatic detectors to support the diagnosis of this condition. A number of Parkinson voice and speech databases exist. The most relevant ones are summarized in Table 3.2, where all of them are unilingual. It is important to know that only the last four listed databases are publicly available.

Table 3.2. The most relevant speech datasets.

| Reference | Name | Size | Language | Tasks |
|---|---|---|---|---|
| [Fox and Ramig, 1997] | N/A | 30 PD<br>14 HC | American English | Sustained vowels, monologue and picture description |
| [Jiménez et al., 1997] | N/A | 22 PD<br>28 HC | Spanish | Sustained vowel and read sentences |
| [Holmes et al., 2000] | N/A | 60 PD<br>30 HC | Australian English | A sustained vowel, a singing up and down scale, 1 minute monologue |
| [Skodda et al., 2008] | N/A | 121 PD<br>70 HC | German | 4 complex sentences |
| [Rusz et al., 2013] | N/A | 20 PD<br>15 HC (male only) | Czech | Sustained vowels (/a/, /i/, /u/), sentence repetition, read passage and monologue |
| [Bandini et al., 2015] | N/A | 20 PD<br>19 HC | Italian | 10 repetitions of a specific read sentence |
| [Vasquez-Correa et al., 2015] | N/A | 14 PD<br>14 HC | Spanish | A set with six sentences and a read text with 36 words are considered. |
| [Jeancolas et al., 2020] | N/A | 115 PD<br>91 HC | French | Sustained vowel 'a', text reading, glissando, rapid Repetition of syllables, free speech, short phrases repetitions, silence, and rhythm. |
| [Sakar et al., 2013] | Parkinson speech dataset | 20 PD<br>20 HC | Turkish | Sustained vowels, words, and short sentences |
| [Little et al., 2009] | Parkinsons Data Set | 23 PD<br>8 HC | Turkish | Sustained vowels |
| [Sakar et al., 2019] | Parkinson's disease classification Data Set | 188 PD<br>64 HC | Turkish | Sustained vowel /a/ repetition |
| [Naranjo et al., 2016] | Parkinson Dataset with replicated acoustic features Data Set | 40 PD<br>40 HC | Spanish | Sustained vowel /a/ repetition |

## 3.2.1 Influence of PD and L-dopa medication on speech

After a review of the literature, five main speech aspects influenced by PD are identified: phonatory, resonant, articulatory, prosodic, and linguistic; where phonation is defined as the vibration of the vocal folds to create sound, articulation is the movement of speech structures (tongue, lips, and jaw) employed in producing the sounds of speech, and prosodic is varying intonation, stress, and rhythm during speech.

### 3.2.1.1 Phonatory aspect

Amplitude and frequency perturbations (Jitter and Shimmer), as well as noise and their associated features, have been widely used along time in order to see the influence of PD on phonation. Some works demonstrate the influence of PD in Jitter and Shimmer [Skodda et al., 2013], [Rusz et al., 2012], [Bang et al., 2013], where some others find that these features do not provide a significant differentiation between PD patient and HC subjects [Midi et al., 2007], [Rahn et al., 2007].

It is difficult to extract a precise conclusion with respect to some acoustic measurements since different works include a different number of patients in distinct stages of the disease and with dissimilar treatments. But in general terms, a considerable number of works point at the presence of incomplete closure of the vocal folds in patients, producing a breathy voice (presence of noise), and an increment in fold stiffness and instability of vocal folds vibrations.

### 3.2.1.2 Prosodic aspect

The prosodic characteristics are directly linked with the analysis of continuous speech. Disturbances in prosody frequently appear in connected speech tasks that necessitate loudness and pitch variation, such as expressing emotions, reading aloud, and in conversational speech. Typically the speech of PD patients presents reduced fundamental frequency (F0) variability (monopitch) [Midi et al., 2007], reduced stress, reduced rate of speech [Midi et al., 2007], modified syllable rate [Chenausky et al., 2011] and reduced loudness variability (monoloudness) [Duffy 2013]. Although less frequent, other prosodic

characteristics are increased rate of the speech [Blanchet and Snyder, 2009] and inappropriate silences [Darley et al., 1969]. The bradykinesia (reduced speed of muscles) and freezing of movement that accompanies the PD speakers sometimes causes difficulty in the initiation of voluntary speech, as well as inappropriate long silences. In contrast, with the advancement of the disease, festinating speech is sometimes observed, resulting in short rushes of speech together with fast locutions. However, most of the times, the presence of festinating speech is just a perception but not an objective fact, since the rate rarely is over the one of normophonic speakers [Weismer, 2006]. It has been justified by a voluntary increase of the rate of speech to compensate the slow speech caused by the disease. Increasing the speed lets the speaker reach the speech rate of normophonic speakers but causes blurring of articulation, similar to the one that would appear in the general population speaking fast.

#### 3.2.1.3 Resonance aspect

Some works focused on hypernasality aspects, and they found that due to the slow movement and rigidity of the muscles involved in the velopharyngeaal mechanism, the velum will not close properly during the speech causing hypernasality [Duffy 2013] (see Figure 3.5). Studies examining hypernasality in PD have yielded controversial results. Logemann et al. [Logemann et al., 1978] detected hypernasality in only 10 % of PD patients, whereas Theodoros et al. [Theodoros et al., 1995] reported hypernasality in more than 30 % of PD speakers. Currently the most common method for hypernasality estimation is the 1/3-octave spectra; which is based on non-invasive analysis of acoustic speech signal [Kataoka et al., 1996].

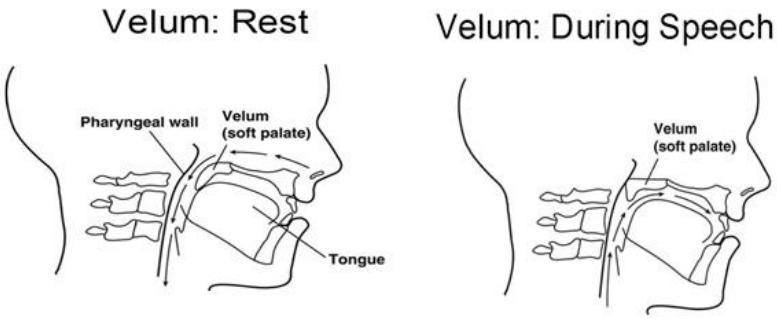



Figure 3.5. Normal Velopharyngeal function (a) Velum at rest (b) Velum during speech [Duffy 2013].

The 1/3-octave spectra method is a type of spectral analysis focused on the examination of spectral changes caused by resonatory speech pathologies. This method is based upon the linear source_filter theory of speech, which was first described by [Fant, 1960]. The linear source–filter theory presents speech as a product of three basic components including source characteristic S(f) connected with fundamental frequency, vocal tract transfer function T(f) associated with formant frequencies and mouth radiation characteristic R(f). The source characteristic approximates the spectrum of vocal folds vibrations, which is filtered by the transfer function of the vocal tract and further modified by the radiation characteristics of the mouth. Based on experiments, previous studies have confirmed the vowel /i/ as an ideal speech task for hypernasality assessment [Kataoka et al., 1996], [Vogel et al., 2009].

#### 3.2.1.4 Articulation aspect

Some studies have showed that the movements of the articulators are reduced compared to normal articulation. At the same time, the timing of the vocal folds closure is imprecise in some occasions, in combination with the rest of articulators during speech. This suggests that the combination of these deficiencies can cause abnormal voiced and unvoiced segment boundaries [Velazquez, 2018]. A considerable number of works study the vowel space area (VSA) for vowels extracted from a running speech between PD and HC; since VSA can be used to reflect the dynamics of the articulators. Authors in [Tjaden and Wilding, 2013] have reported that individuals with PD have a significantly smaller VSA in comparison with HC speakers. Conversely, other studies have reported no statistically relevant differences in this same measure between HC and PD speakers [Skodda et al., 2011], [Skodda et al., 2012]. On the other hand, the Vowel Articulation Index (VAI) is significantly reduced in male and female PD patients as compared with the HC group [Rusz et al., 2013], allowing discrimination with an accuracy around 80 %, although no correlations were seen between VAI and the stage of the disease.

On the other hand, other acoustic measures that can be used for articulatory interpretation are formants. Formants are distinctive frequency components of the acoustic signal produced by speech. The formant with the lowest frequency is called $F_1$, the second $F_2$, and the third $F_3$. Most often the two first formants, $F_1$ and $F_2$, are sufficient to identify the

vowel. The literature reports that PD patients have reduced F1 and F2 slopes compared to the HC groups, associated with the reduced velocity of the articulators since F1 reflects tongue height and F2 reflects tongue advancement [Weismer et al., 1998], [McKell, 2016].

#### 3.2.1.5 Linguistic aspect

Another sign of PD is the linguistic disturbances. These are mainly related with changes in the informative content of the speech (i.e. the expressive dimension of the speech), and are manifested as repetitions and/or substitutions. The prevalence of these disturbances is no more than 30 % of the PD population [Benke et al., 2000]. There are two variants of speech repetitions, one hyperfluent, known as pallilalia, and another dysfluent, stuttering-like. Pallilalia, or the compulsive repetition of words, or even phrases, is sometimes present in PD patients, but no more than in 15 % [Benke et al., 2000] of the PD population. Due to the poor articulation and decreasing loudness, these repetitions become blurred. However, stuttering-like repetitions are usually relatively well articulated at a constant rate and loudness, and their prevalence is similar to the former one. Although the role of repetitions of speech in PD is not well known, they are commonly linked to a deficit of the motor speech control, associated to the freezing symptom, but cognitive and linguistic factors (at the pre-articulatory level) seems to contribute to their generation [Benke et al., 2000]. On the other hand, the authors of [Illes et al., 1988] have found that the number of doubtful silences and their duration is more significant in PD patients than in HC subjects, as well as the number of interjections.

#### 3.2.1.6 Combined aspect

Many works focused on one speech aspect (phonation or articulation or prosody or linguistic), where some others bring together features from two or more different aspects (called combined approach). The combined approaches are the most appropriate to help in the diagnosis of PD since they can cover different types of motor signs at the same time.

Some early works point out that PD has a clearer reflection on the phonation of patients than in articulation or prosody [Bandini et al., 2015], where another work suggests that on early stages, prosody is the most affected speech aspect [Rusz et al., 2011-a]. In

general, although the influence of the disease in most of these aspects is noticeable, there is no consensus on which aspect is more appropriate to help on differential diagnosis in early stages. Therefore, a solution is to use a combined approach containing at least two aspects (articulatory and phonatory or articulatory and prosodic for example).

In the same sense, it was shown that read sentences can help better to obtain accurate differential diagnosis than sustained vowels [Jiménez-Monsalve et al., 2017]. However, these conclusions must not mislead research on avoiding the use of sustained vowels, as the selection of the most appropriate material depends on the analyzed aspect and the feature to be measured.

#### 3.2.1.7 Effect of L-dopa on speech

Some studies have been done to study the effect of dopamine medication on speech in PD patients. Pinho et al. [Pinho et al., 2018] found that the use of levodopa improves vocal parameters such as F0 and jitter; however, vocal intensity remains reduced in both the “on” and “off” states of therapy. Nakano et al. [Nakano et al., 1973] found overall speech improvement after dopamine medication and significant speech intelligibility improvements. Some other researchers [Wolfe et al., 1975] did not find any measured or perceived speech rate changes after taking the L-dopa medication.

### 3.2.2 Hand-crafted features and classical classifier in PD detection based on speech

From the other side, many researchers worked on PD detection based on speech analysis like Vasquez-Correa et al. [Vasquez-Correa et al., 2015] who have studied voiced and unvoiced segments of running Spanish speech recorded in non-controlled noise conditions from a set of 14 PD and 14 HC subjects to detect PD. Voiced and unvoiced segments of the signals were analyzed separately and different sets of audio features were considered achieving 86 % and 99 % detection accuracy for voiced and unvoiced frames, respectively where an SVM was used for classification.

Also some works got a very high PD recognition accuracy from voice recordings (reaching 100 % accuracy). Hariharan et al. [Hariharan et al., 2014] extracted several dysphonia features where feature pre-processing using Model-based clustering (Gaussian mixture model) was applied. Different feature selection techniques were applied: principal component analysis (PCA), linear discriminant analysis (LDA), sequential forward selection (SFS) and sequential backward selection (SBS). For classification, least-square support vector machine (LS-SVM), probabilistic neural network (PNN), and general regression neural network (GRNN) were studied. Speech samples are taken from Parkinsons Data set (see Table 3.2). The proposed model in [Hariharan et al., 2014] gives a very promising classification accuracy of 100 %. However, this result is considered overoptimistic due to the existence of samples in training and test belonging to the same subject.

From the other side, some works have been working on PD stage estimation such as [Schuller et al., 2015] who have estimated PD stage (UPDRS) using a set of handcrafted acoustic features and linear Support Vector Regression Deep learning; where Spearman's Correlation Coefficient ($\rho$) was used as evaluation measure.

### 3.2.3 Deep learning for PD detection based on speech

Some other studies also applied deep learning for PD detection based on speech analysis. These studies are separated into two groups; in the first group the deep models deal directly with the raw speech signals such as Frid et al. [Frid et al., 2016] who proposed a CNN for speech signal processing using One-Dimension for PD as shown in Figure 3.6. The input signal is divided into small (20 ms) processing windows with overlap, and each of them is used as a separate data-point. Majority voting is applied to determine the final classification. Two types of experiments were done: the windows level and participant level. In the first type (windows level) windows are chosen randomly for training and test without regard to the participant, whereas the second type (participant level) windows selection for training and test is done by participant; which means there is no windows in training and testing referring to the same participant. The binary classification between healthy and other stages of PD results at the window level achieved an accuracy of more than 65 % for all the stages. The other binary classification at the participant level results in an accuracy of more than 60 % between subsequent stages of severity of the disease.

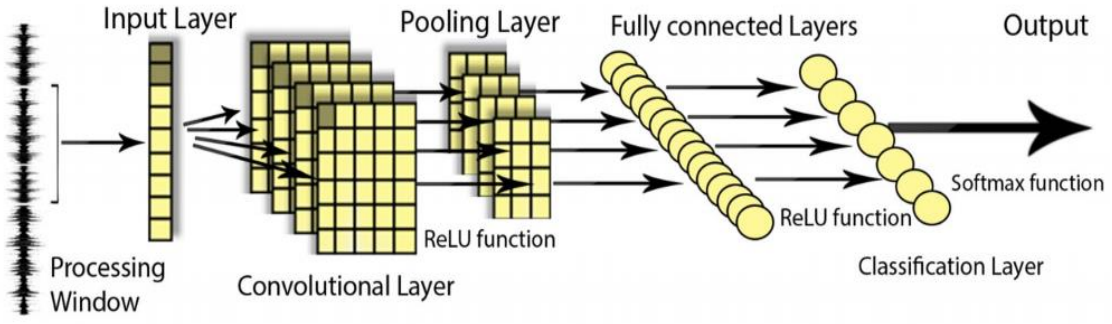



Figure 3.6. CNN model for PD detection from one-dimensional speech signal in [Frid et al., 2016].

In the second group, the deep models deal with some handcrafted acoustic features instead of raw signals such as Caliskan et al. [Caliskan et al., 2017] who proposed a deep neural network (DNN) classifier for PD detection based on speech analysis. This model contains a stacked autoencoder (SAE) and a softmax classifier. The SAE was employed for extracting intrinsic information within the speech features; softmax layer was used for interpreting the encoded features to classify the patients (see Figure 3.7). In order to justify the performance of proposed model, several experiments were conducted with two different datasets (Parkinson speech dataset, and Parkinsons Data Set see Table 3.2), where each subject has different instances. All the instances are combined together, and training and test are randomly chosen. In this case, there will be instances in training and test that belong to the same subject and the results will be overoptimistic. The experimental results showed that DNN classifier was a convenient classifier for of PD diagnosis with an accuracy of 86.1 % when Parkinsons Data set is used, and 65.55 % when Parkinson speech dataset is used.

Gunduz [Gunduz, 2019] also proposed two frameworks based on CNNs to classify PD using sets of vocal features. These vocal features are separated into 4 types: concat (concatenation of baseline features (such as jitter, shimmer, fundamental frequency, etc.), time frequency features (such as intensity, bandwidth, and formants), and vocal fold features (such as Empirical mode decomposition (EMD), Vocal fold excitation ratio (VFER), etc.)), Mel-Frequency Cepstral Coefficients (MFCCs), Wavelet transform based features, and Tunable Q-factor Wavelet transform (TQWT). The first framework combines different feature sets. The combined features are given to 9-layered CNN as inputs (Figure 3.8-a).

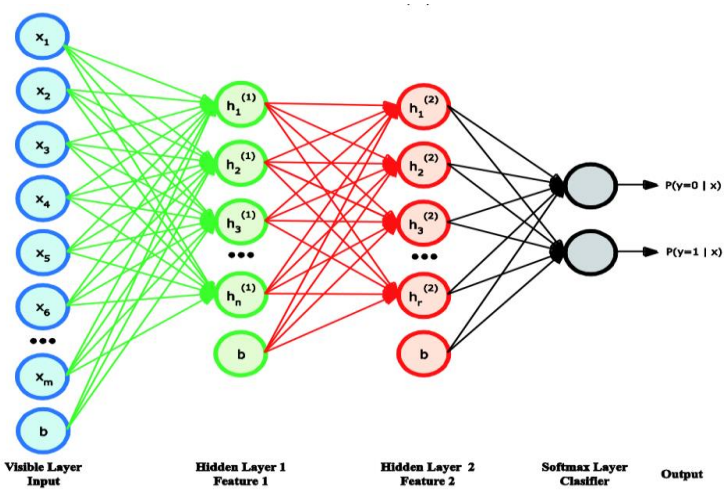



Figure 3.7. The proposed Deep Neural Network in [Caliskan et al., 2017].

In comparison, the second framework passes feature sets to the parallel input layers which are directly connected to convolution layers. Thus, deep features from each parallel branch are extracted simultaneously before combining in the merge layer (Figure 3.8-b). Proposed models are trained with Parkinson's disease classification Data Set that includes 3 voice records per individual (see Table 3.2). Each voice record is analyzed separately, and majority voting is applied for the final decision. Experimental results show the second framework to be very promising (with an accuracy of 86.9 %), since it is able to learn deep features from each feature set via parallel convolution layers.

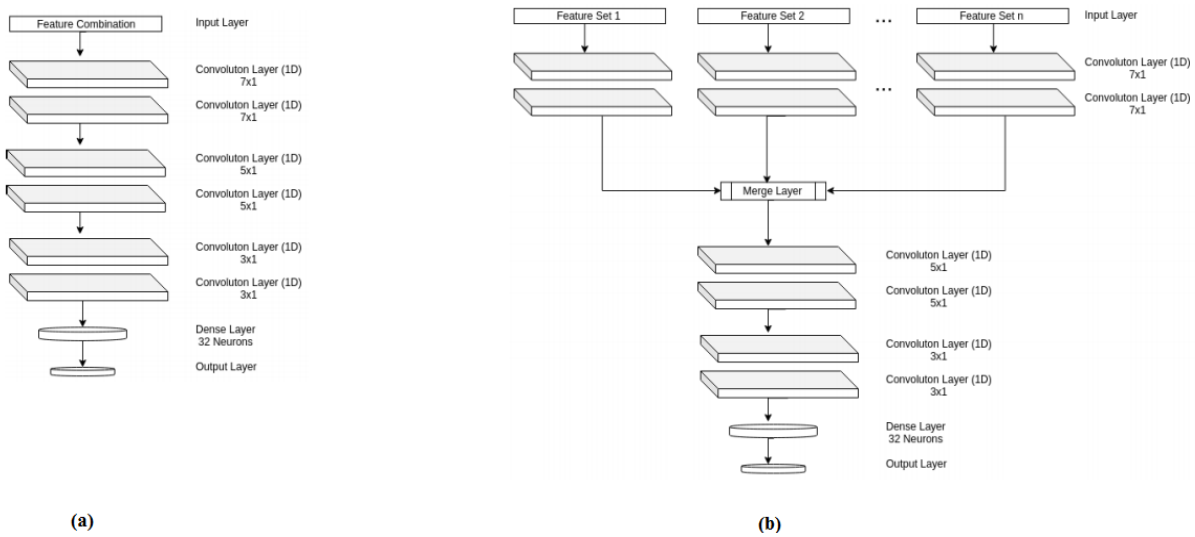



Figure 3.8. The 2 frameworks based on CNN for PD detection proposed in [Gunduz, 2019], (a) feature-level combination, (b) model-level combination.

Jeancolas et al. [Jeancolas et al., 2020] have proposed a new methodology for early detection of PD from voice analysis, called x-vectors. This method consists in extracting embedding features from a DNN taking MFCC as inputs. The goal of this study was to assess whether this technique is better than the MFCC-GMM (Mel-Frequency Cepstral Coefficients - Gaussian Mixture Model). It was found that x-vectors, combined with discriminant analyses, is more relevant than MFCC-GMM classification for text-independent tasks, and particularly suited to women PD detection (with 30 % classification equal error rate (EER)).

Based on the revised studies in this section, it was found that, in case of speech analysis, deep models work better with handcrafted acoustic features than raw signals.

## 3.3 Eye movements and Parkinson's disease

Oculomotor disturbances have been widely studied in patients with PD using various tasks. The majority of recent literature about PD-associated oculomotor changes concerns saccadic dysfunction. A number of Parkinson eye movements' recordings databases exist. The most relevant ones are summarized in Table 3.3.

### 3.3.1 Influence of PD and L-dopa medication on eye movements

Different results perform on reflexive and voluntary saccadic eye movements tasks. No saccadic impairment in reflexive saccades was found but impairment in voluntary saccadic eye movements' task was observed (increased latency, sometimes hypometria, and impaired voluntary suppression of unwanted saccades) [Amador et al., 2006], [Vidailhet et al., 1994]. The reason why voluntary saccades are more affected than the reflexive saccades is that voluntary saccades involves a cortex-BG-SC pathway, whereas reflexive saccades can be generated by direct projection form the cortex to the SC (see Figure 2.6 in chapter 2) [May, 2006]. As mentioned in chapter 2, section 2.2, the degeneration of dopaminergic neurons in the SNc leads to increased activity in the indirect circuit and to an increased inhibitory output from BG, which further will leads to inhibition of SC. Voluntary saccades processing will leads to inhibition of SC, resulting a suppression of saccadic eye movements.

PD patients typically reach a target eye position through a series of discrete short saccades. The latency and velocity of the series of saccades are in the normal range, while the amplitudes are reduced [Kimming et al., 2002]. Impairments of saccades and smooth pursuit eye movements have been reported in about 75 % of PD patients. Other studies focused on eye movements during reading and they found that the reading time was significantly slower in PD compared with HC due to the reduced ability to generate saccades and the prolonged duration of fixations [Jehangir et al., 2018].

Table 3.3. The most relevant eye movements' recordings databases.

| Reference | Size | Acquisition device | Tasks |
|---|---|---|---|
| [Jehangir et al., 2018] | 42 PD<br>80 HC | Not assigned | Reading regular and irregular distributed digit numbers contained in a page |
| [Kimming et al., 2002] | 11 PD<br>11 HC | Infrared light technique | Horizontal eye movements recorded while subjects follow certain targets appearing on a screen. |
| [Amador et al., 2006] | 14 PD<br>11 HC | ISCAN RK-426 eye tracking system | Anti-saccade, delayed anti-saccade, and remembered anti-saccade tasks. |
| [Tseng et al., 2013] | 14 PD<br>24 HC | Not assigned | Watching short videos |

Various studies have been conducted to determine the effect of medication on saccadic eye movements in PD patients. The literature about the effects of levodopa treatment on eye movements' deficits is controversial. Some studies suggests beneficial effects of dopaminergic medication on saccades in PD, where some other studies report no beneficial effect of dopaminergic treatment on saccadic parameters like latency, amplitude, and accuracy [Srivastava et al., 2014].

### 3.3.2 Hand-crafted features and classical classifier in PD detection based on eye movements

Very few studies worked on eye movements' analysis for PD detection. Tseng et al. [Tseng et al., 2013] extracted three types of features: oculomotor-based features (e.g., distributions of saccade amplitudes and fixation durations), saliency-based features correlated participants' gaze to predictions from a computational model of visual salience, and group-based features that capture deviations in participants' gaze allocation. 89.6 % classification accuracy was obtained when applying SVM for classification.

## 3.4 Multimodal assessment of Parkinson's disease

To the best of our knowledge, there are several works considering different bio-signals to assess motor impairment of PD patients, most of these studies consider only one modality. Multimodal analyses (considering information from different sensors) for an accurate prediction of PD disease have not been extensively studied. A small number of multimodal databases for PD exist and are listed in Table 3.4.

Table 3.4. List of multimodal databases for PD.

| Reference | Size | Acquisition device | Tasks |
|---|---|---|---|
| [Prashanth et al., 2016] | 401 PD<br>183 HC | | -University of Pennsylvania Smell Identification Test (UPSIT)<br>-REM sleep Behavior Disorder Screening Questionnaire (RBDSQ)<br>- Biomarkers from Cerebrospinal fluid (CSF)<br>- Striatal Binding Ratio (SBR) from SPECT imaging |
| [Barth et al., 2012] | 18 PD<br>17 HC | - BiSP<br>-3D gyroscopes and 3D accelerometers | Writing tasks<br>On paper:<br>-Drawing twelve circles at the same place<br>-Tracing four preprinted spirals-Tracing four preprinted meanders<br>In the air:<br>-Drawing twelve circles around a virtual point<br>-Performing pronation/supination movements for 20 s<br>-Performing finger tapping on the pen for 20 s<br>Gait tasks<br>10-meter walk, Heel-toe tapping, Circling |
| [Vásquez-Correa et al., 2019] | 44 PD<br>40 HC | -Wacom cintiq 13-HD<br>-eGaIT system (3D gyroscopes and 3D accelerometers) | Speech tasks:<br>-rapid repetition of the syllables /pa-ta-ka/, /pe-ta-ka/, /pa-kata/,<br>/pa/, /ta/, and /ka/<br>- read sentences, read story of 36 words, and a monologue.<br>Gait tasks:<br>-20 meters walking with a stop after 10 meters<br>-40 meters walking with a stop every 10 meters<br>Writing tasks:<br>drawing a circle, a cube, two rectangles, a house, a diamond, the Rey-osterrieth figure, a spiral following a template, a free spiral |

### 3.4.1 Hand-crafted features and classical classifier in PD detection based on multimodal analysis

Prashanth et al. [Prashanth et al., 2016] use the combination of the pre-clinical markers of non-motor features of sleep behavior disorder (RBD) and olfactory loss,

Cerebrospinal fluid (CSF) measurements, and dopaminergic imaging features to perform PD classification using SVM. The classification accuracy obtained is 96.40 %.

In another work, Barth et al. [Barth et al., 2012] combine handwriting and gait features for PD classification at global feature level; where in each modality, a set of global features were extracted for each task and combined together to form a single feature vector as shown in Figure 3.9. At a later stage, Barth et al. evaluated the multimodal data using various machine learning models such as AdaBoost, SVM, and LDA. An excellent classification accuracy of 97 % was reached with the AdaBoost classifier.

Some other authors [Pham et al., 2019] propose a multimodal approach combining voice and handwriting tests at global decision level to enhance the reliability of detecting PD patients (summarized in Figure 3.10). The two data sets used are the Parkinson Speech Dataset and the ParkinsonHW dataset defined in Table 3.1 and Table 3.2.

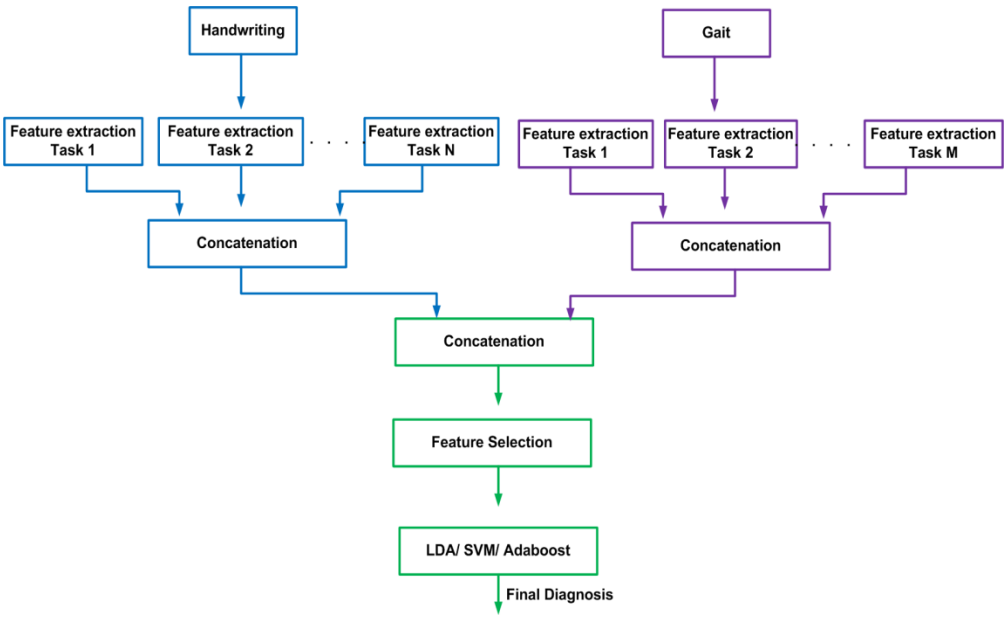



Figure 3.9. The proposed fusion of handwriting and gait modalities by [Barth et al.,2012].

The Parkinson Speech Dataset is composed of 26 speech samples of 20 PD patients and 20 healthy individuals in addition to 2 speech samples collected from a test group made up of a separate 28 PD patients. For the study in [Pham et al., 2019], data from all 68 individuals were mixed to form the overall dataset from which the train and test set were split by the 80:20 ratio. Concerning the spiral test database (ParkinsonHW), it consists of spiral

tests collected from 15 healthy individuals and 62 PD individuals. For all subjects, three types of handwriting recordings, the SST, DST and STCP, are taken. Each handwriting test is studied separately. Samples in both datasets do not belong to the same subjects. Pham et al. have evaluated voice test data using various machine learning models such as SVM, Random Forest, and k-nearest neighbors. The results of these 3 classification models were combined using majority voting. The highest accuracy of 95.89 % is obtained using the combination of the 3 classification models. Moving to spiral test data evaluation, each spiral test is studied alone using the same machine learning models applied speech. For each spiral test, the best model is selected. At a later stage, the results of the best 3 spiral tests are combined using majority voting. The best accuracy of 99.6 % is achieved using the k-Nearest Neighbors classifier in the DST. Pham et al. also built a multimodal system for PD classification based on the combination of the results obtained from the vocal and spiral tests. The final diagnosis is determined by either the vocal or spiral ensemble having a positive diagnosis (PD).

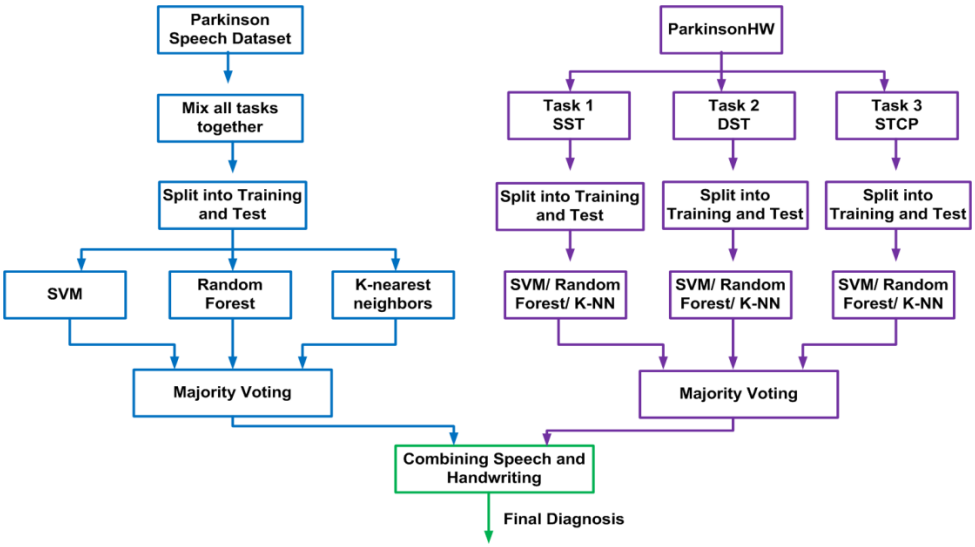



Figure 3.10. The proposed fusion of handwriting and voice modalities by [Pham et al., 2019]

From a personal point of view, we believe that the objective of the multimodal combination method proposed by Pham et al. [Pham et al., 2019] is to deal with the idea that PD symptoms manifest in varying degrees and combinations with different individuals, which means that PD may influence handwriting and not speech or vice versa. However, the multimodal system described is not reliable for two reasons. First of all, the samples taken from both vocal and spiral datasets do not belong to the same subjects. Secondly, in the vocal

test analysis there are samples in training and testing referring to the same participant, so the results obtained are considered overoptimistic.

### 3.4.2 Deep learning for PD detection based on multimodal analysis

The multimodal systems described above use hand-crafted features with classical classifiers for PD classification, where Vásquez-Correa et al [Vásquez-Correa et al., 2019] employs deep learning approach for multimodal assessment of PD. The authors used the combination of handwriting, speech and gait signals at global features level to detect PD using a CNN models for feature extraction and an SVM for classification, where the samples are taken from a dataset that contains recordings of speech, handwriting, and gait collected from 44 PD patients and 40 HC subjects. For speech and gait signals, 2D-CNNs (represented in Figure 3.11-A) are applied with the Short Time Fourier Transform (STFT) as input, where for on-line handwriting, they consider a 1D CNN (represented in Figure 3.11-B) with the raw signals as input. The extracted features (the output of the convolutional layers) for each modal are combined together and applied to an SVM for classification. The results obtained showed that combining the three signals returns a better accuracy (97.6 %) than using each signal separately.

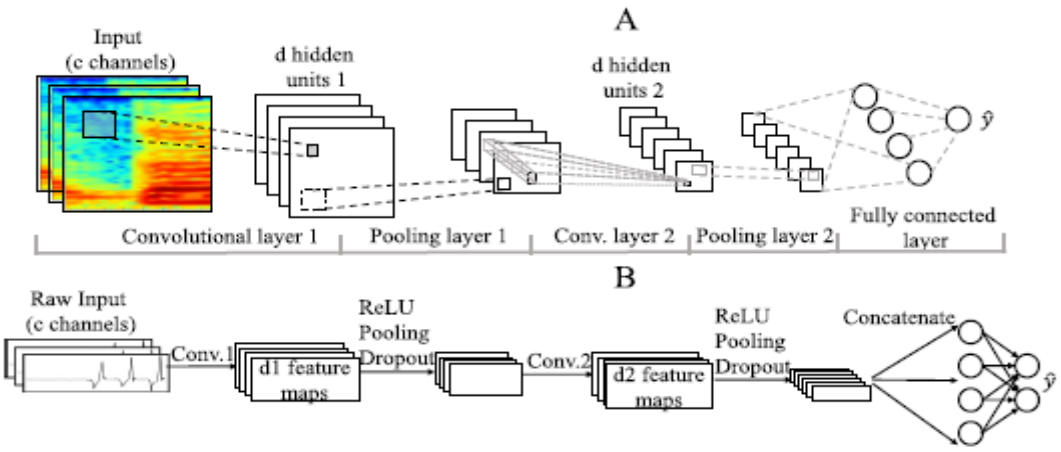



Figure 3.11. The CNN model architecture in [Vásquez-Correa et al., 2019].

## 3.5 Summary and conclusions

Several studies referring to PD detection via handwriting, speech, or eye movements' analysis and summarized in this chapter. These studies are at the state of the art. However, there are some missing analyses that should be addressed. Actually, most of the previous works have analyzed speakers of a single language, where we do not believe that PD is language dependent. A language-independent strategy to evaluate PD from handwriting or speech has not been enough addressed.

High accuracies have been reported for the reviewed systems performing automatic detection of PD. Typically, some accuracies overpass 95 %. We noticed that some works were employing over-optimistic methodologies (for example samples in training and testing belongs to the same subject). This reduces the relevance of the reported performances.

Regarding the multimodal analysis system, a system combining handwriting, speech and eye movements' signals for PD detection is not yet developed, and to our knowledge there exists no multimodal dataset. In general, although the influence of the disease in handwriting, speech, and eye movements is noticeable, there is no consensus on which aspect is more appropriate to help on differential diagnosis in early stages. Therefore, a solution is to use a combined approach containing at least two signals for PD detection.

Finally, in this field it is difficult to collect large databases. Even though databases used in previous studies are small in size, no data augmentation techniques were applied to generate new synthetic samples, and to enlarge the training set.

Based on these findings, a major step toward a unified objective assessment of the disease would consist in developing a robust multilingual computer aided system that can evaluate PD with high performance from a multimodal vector of features, where both global features and classical classifier with efficient feature selection and short term features and deep learning approaches are studied and compared, and where proper data augmentation techniques are applied to overcome the limitation of data.

# 4 Construction of our PD Multimodal Database

As discussed in chapter 2, handwriting, speech, and eye movement impairments happen in the early stage of the disease and can be used for early detection. Based on this and to attain the target of early detection and monitoring the progression of PD via multimodal signals, a multimodal database (that we call it Parkinson's Disease Multimodal Collection (PDMultiMC)) was constructed. It includes online handwriting, speech, and eye movement recordings collected from PD patients and HC subjects, and distributed into 3 different datasets: HandPDMultiMC, SpeechPDMultiMC, and EyePDMultiMC, where the completed tasks were verified by an experienced neurologist at Saint George Hospital-Lebanon. Before starting collecting data, this study has been approved by the IRB of the University of Balamand and Saint George Hospital University Medical Center. After that, PD patients were selected from those attending an experienced neurologist at Saint George Hospital.

It is worth noting that the data are collected from PD patients in two conditions: medication off and medication on so the influence of the medication on classification can be studied. Moreover, to monitor the progression of the disease, the rating scales and the cognitive impairment scores defined in chapter 2 must also be collected. Actually, the major objective of the thesis is the early detection of PD. However, it is very hard to identify early stage PD patients. We make the assumption that PD patients with medication on have imperfections in their handwriting, speech and eye movements closer to the early stages of the disease.

## 4.1 Participants

Prior to their inclusion in this study, all subjects should inform consent to participate in the study. The consent form provides a description of the overall purpose of the research, the specific details of the study, confidentiality, the right to withdraw from participation in data collection, and a list of the study investigators who will be available to address any questions about the study. The PD patients selected were examined in their "off-state" (before taking the dopaminergic medication), and "on-state" (1 hour after taking their regular dose of dopaminergic medication). The total number of PD patients in the dataset is 21 (16 Males and

5 Females). Another 21 HC subjects (5 Males and 16 Females) were selected from my entourage and included in the dataset. The distribution of samples between genders in both groups (PD and HC) was by chance and not by choice. PD and HC subjects are matching for age, years of education, and hand dominance since these variables may have been shown to affect handwriting performance. This database includes samples in three languages: Arabic, French, and English. For confidentiality, each subject is represented by an ID number and not by name. Demographics characteristics (age, gender, hand dominance, years of education), and clinical characteristics (MMSE, UPDRS, the stage of PD, the years of the disease, the state of the patient (medication on or medication off) and the dopamine dosage) concerning each subject are saved. Statistics about the demographic and clinical characteristics are summarized in Table 4.1.

Table 4.1. Demographic and clinical characteristics of PD and HC (means ± standard deviation (STD)).

| Characteristic | HC (n=21) | PD off state (n=21) | PD on state (n=21) |
|---|---|---|---|
| Age [years] | 63.91 ± 6.33 | 67.52 ± 7.54 | |
| Female [no. ( %)] | 16 (76.19) | 5 (23.81) | |
| Years of education | 13.62 ± 3.98 | 12.09 ± 5.07 | |
| Right-handed [no. (%)] | 21 (100) | 20 (95.24) | |
| MMSE | 29.71 ± 0.46 | 27.67 ± 1.62 | 28.29 ± 1.15 |
| UPDRS-I | | 3.24 ± 1.48 | 2.62 ± 1.16 |
| UPDRS-II | | 12.67 ± 5.71 | 10.67 ± 3.85 |
| UPDRS-III | | 12.86 ± 5.93 | 10.95 ± 5.13 |
| Total UPDRS | | 28.76 ± 11.57 | 24.24 ± 8.77 |
| H&Y Stage | | 1.81 ± 0.77 | |
| Disease duration (Months) | | 80.29 ± 77.12 | |
| L-Dopa Dosage (mg) | | | 672.62 ± 361.22 |

# 4.2 Handwriting tasks

The first step was to choose tasks in a manner to highlight as much as possible the differences between PD and HC. Each subject was asked to complete handwriting tasks according to a prepared template. A completed task sheet is shown in Figure 4.1-a. The handwriting template was composed of two parts: part I is the free writing, where subjects were asked to write their name and family name in their familiar language 5 times with their own speed and size each time on a different line. In the copying task, 3 different patterns loop (repetitive letter ‘l’, triangular, and rectangular waves) with “Monday” and “Tuesday” words in 3 different languages are printed on the left of the sheet paper placed on the tablet. Subjects

were asked to start copying the patterns and to proceed from left to right until they completed 10 cycles, then, they were asked to copy “Monday” and “Tuesday” 5 times consecutively with their familiar language (see Figure 4.2).

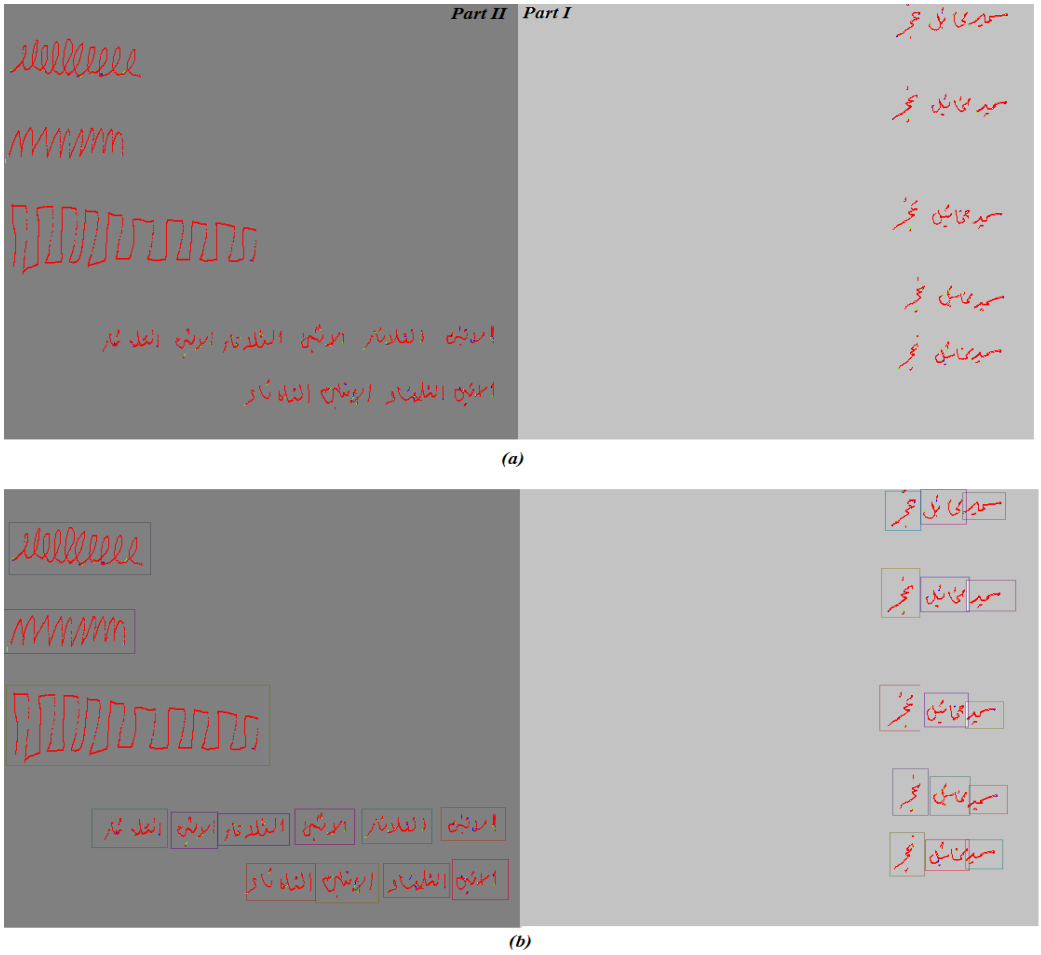



Figure 4.1. (a) Handwriting sample used to assess the handwritten skills of a given individual, (b) word and pattern segmentation.

We have chosen these 3 patterns loop because they demand one continuous pen movement without the brief break between writing words, which will emphasis abnormal movements of dystonia, hypokinesia and tremor [Alty et al., 2017]. From the other side, dystonic posturing of the wrist and fingers may become more obvious with longer periods of writing; which will lead to a difficulty in maintaining the size and shape of the handwriting causing what is known by micrographia or small writing. As a result, micrographia needs an extended writing task to manifest [Alty et al., 2017]. That is the reason of choosing the word repetition tasks in the database.

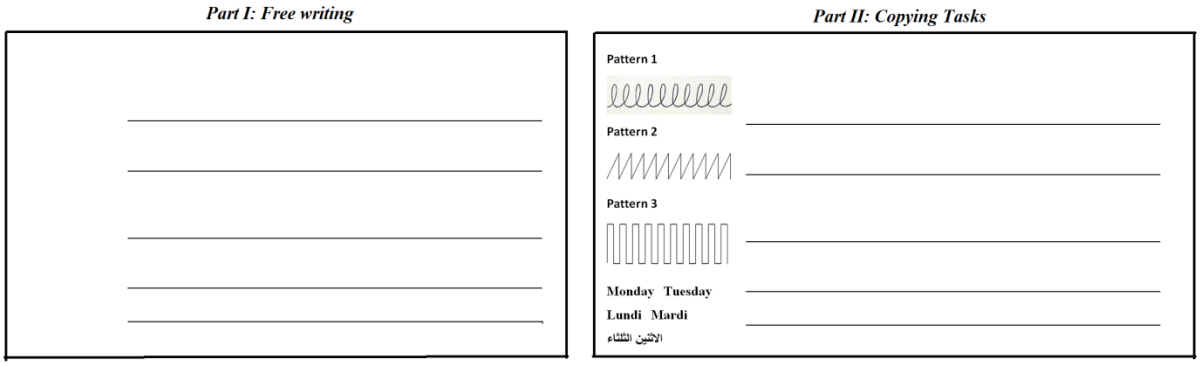



Figure 4.2. The handwriting template proposed to the scripters when collecting the PD database.

Handwriting analysis is divided into 2 types: online analysis and offline analysis. Online analysis requires a stylus and an electronic digitizer tablet connected to a processing computer to grab dynamic information. Offline analysis, on the other hand, deals with the morphological or the external information contained in the handwriting [Kekre and Bharadi, 2010]. In offline analysis, we are having the handwriting template coming from an imaging device like scanner, hence we have only static characteristic of the writings. However, in online approach we can capture the dynamic properties of the handwritings. Online analysis was chosen instead of offline analysis because it was realized that the information content in it is more significant as compared to offline collection [Kekre and Bharadi, 2010]. Wacom Intuos5 pen tablet is used for online handwriting data collection (Figure 4.3). Typical features of the tablets are as follows [Naik, 2012]:

1. Active Area (W x D): 223.52 x 139.70 $mm^2$
2. Connectivity-USB connectivity
3. Pressure levels -2048
4. Sensor pen without battery
5. Sample rate- 197 Points per second
6. LPI - lines per inch resolution-5080 lpi

Each captured point has multi-dimensional information. It contains information about the trace of the pen tip (X-Y-Z coordinate), the pressure of the pen tip on the surface, the angles of the pen relative to the tablet (altitude and azimuth), and timestamp [Naik, 2012]. The Z coordinate is not collected by all the digitizing tablets, but it is useful for the PD

detection task, as will be shown in chapter 5. The movement of the pen can be recorded while the pen is touching the surface, and when the pen is in the proximity of the surface (in-air movements); where the maximum height at which the pen tip is detected is approximately 2cm. The total amount of captured points available for the analysis of this database is 1,386,894 (510,513 points from PD in the "on-state", 510,066 points from PD in the "off-state", and 366,315 points form HC), with approximately 22,014 average points per subject. Details of the amount of captured points per task are shown in Table 4.2.

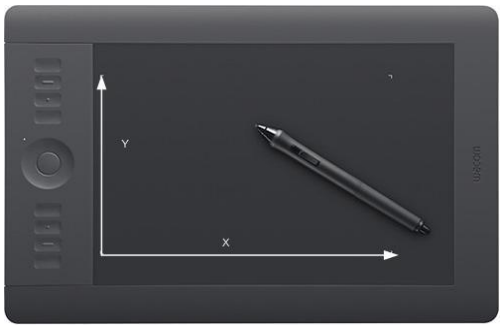



Figure 4.3. Wacom intuos 5 tablet and pen device [Naik, 2012]. The tablet captures X and Y positions, but also z position when the pen tip is close to the tablet. The tablet also captures pen altitude and azimuth, and pen pressure on the surface at each time stamp.

Table 4.2. Details of the amount of captured points per task. The mean captured points per subject and per task is between parentheses.

| **State** | **Nb. Of subjects** | **Repetitive letter 'l' (2,476)** | **Triangular wave (2,343)** | **Rectangular wave (4,342)** | **"Monday" (3,971)** | **"Tuesday" (3,463)** | **"Name" (3,188)** | **"Last Name" (2,230)** | **Total (22,014)** |
|---|---|---|---|---|---|---|---|---|---|
| **PD "on state"** | **21** | **64,096** | **57,597** | **96,105** | **91,790** | **81,482** | **67,404** | **52,039** | **510,513** |
| F | 5 | 21,944 | 16,684 | 24,455 | 24,902 | 23,494 | 15,804 | 13,837 | 141,120 |
| M | 16 | 42,152 | 40,913 | 71,650 | 66,888 | 57,988 | 51,600 | 38,202 | 369,393 |
| **PD "off state"** | **21** | **50,951** | **50,065** | **98,410** | **92,568** | **86,770** | **74,145** | **57,157** | **510,066** |
| F | 5 | 16,580 | 12,977 | 18,993 | 30,045 | 29,999 | 17,073 | 14,792 | 140,459 |
| M | 16 | 34,371 | 37,088 | 79,417 | 62,523 | 56,771 | 57,072 | 42,365 | 369,607 |
| **HC** | **21** | **40,938** | **39,976** | **79,046** | **65,795** | **49,943** | **59,320** | **31,297** | **366,315** |
| F | 16 | 32,737 | 9,308 | 58,547 | 50,243 | 36,778 | 45,557 | 23,859 | 257,029 |
| M | 5 | 8,201 | 30,668 | 20,499 | 15,552 | 13,165 | 13,763 | 7,438 | 109,286 |
| **Total** | **63** | **155,985** | **147,638** | **273,561** | **250,153** | **218,195** | **200,869** | **140,493** | **1,386,894** |

## 4.3 Speech tasks

The recording protocol considers different tasks which were designed to analyze several aspects of the voice and speech of people with PD. All participants were asked to perform two speaking tasks that represent natural speech and reflect motor speech disorders comprehensively:

**Sustained vowel 'a':** participant is instructed to produce a single vowel "a" at a habitual level (habitual pitch and loudness) and hold the pitch of this as constant as possible, for as long as possible on one comfortable breath. Different studies have reported instability in the phonation (vibration of the vocal folds) of PD speakers and articulatory (resonances in the vocal tract) deficits, that affect the speech production and its intelligibility [Hemmerling et al., 2016]. This task is chosen by most of the studies about Parkinson's speech because it is easy to be reproduced by elderly subjects and because it isolates the respiratory and laryngeal mechanisms in order to evaluate the ability to produce a sustained tone at normal pitch, quality, loudness, duration, and steadiness [Duffy, 2000]. This task provides information about the phonatory and articulatory processes of speech production [Hemmerling et al., 2016]. We selected the vowel 'a' because in [Orozco-Arroyave et al., 2015] it was found that the vowel 'a' reports the highest classification accuracy to detect PD when benchmarking different sustained vowels.

**Text reading:** the participant is instructed to read loudly a standardized text of approximately 150 words appearing on the PC screen written with his familiar language. PD is associated with prosody deficits, where prosody is the term applied to reduced loudness, breathy voice, monotone pitch, voice tremor, intermittent rapid rushes of speech, and imprecise production of consonants, which significantly impacts a patient's ability to communicate. Such speech task permits judgments of speech rate, phrase length, voice quality, resonance, and precision of articulation [Duffy, 2000]. Connected speech also permits assessment of the prosodic features of speech by studying the variation of F0 and intensity during the text [Duffy, 2000]. The text red by the subjects is as follows:

- ✓ *text:*

> *"These days technology has brought many changes to our lives, especially in the field of education and communication. In communications, there are significant changes and the major changes happen in the way we communicate with other people. We do not need to meet the person face to face to discuss ideas. We can now simply call them or do video chat using internet and PC. In the past, we spent a long time traveling to a distant place, but now we only need few hours or minutes to go to the desired place using transportations. In addition, technology has brought changes and benefits to students and teachers. For example, students can do their homework faster and easier by using the internet. Teachers also get some advantages from it. They can combine their teaching skills with the addition of color slides and graphics to show and prove how things happen. In conclusion technology itself has given us several advantages to improve and develop the quality of our lives."*

These two speech tasks are repeated only one time. The total amount of speech data available for the analysis of this database is approximately 2 hours (44 mins from PD in the "on-state", 42 mins from PD in the "off-state", and 38 mins from HC subjects), with approximately 2 mins average recording per subject. Details of the amount of speech per task are shown in Table 4.3.

Table 4.3. Details of the amount of speech per task (in seconds). The mean amount of speech per subject and per task is between parentheses.

| **State** | **Nb. Of subjects** | **Sustained vowel 'a' (9s)** | **Text (108s)** | **Total (117s)** |
|---|---|---|---|---|
| **PD "on-state"** | **21** | **206** | **2,406** | **2,612** |
| F | 5 | 26 | 580 | 606 |
| M | 16 | 180 | 1,826 | 2,006 |
| **PD "off-state"** | **21** | **199** | **2,320** | **2,519** |
| F | 5 | 23 | 510 | 533 |
| M | 16 | 176 | 1,810 | 1,986 |
| **HC** | **21** | **175** | **2,086** | **2,261** |
| F | 16 | 124 | 1,586 | 1,710 |
| M | 5 | 51 | 500 | 551 |
| **Total** | **63** | **580** | **6,812** | **7,392** |

## 4.4 Eye movements tasks

Different tasks were designed to analyze several aspects of the saccadic eye movements of people with PD. In this thesis, only one task was selected and performed by the participants:

**Text reading:** the same text used in speech part is used here, the participant is asked to read loudly the text appearing on the PC screen while the webcam is recording the face characteristics changes. PD saccades are known to be hypometric and tend to show a prolonged latency in cognitively impaired PD patients [Waldthaler et al., 2018]. When reading, the eye moves through the text in a series of fixations and saccades. A fixation occurs when the eye temporarily rests on a word. In between fixations, the eye makes a rapid movement called a saccade. Saccades can move the eye forward through the text (a forward saccade) or backward [Fraser et al., 2017]. Such task permits judgement of the affected reading performance due to the saccadic eye movements' deficits in PD [Waldthaler et al., 2018].

## 4.5 Acquisitions

To collect this database, an application[1] written in C# language was developed. This application is composed of four parts: the first part consists of filling a questionnaire with demographic and clinical characteristics of each participant as shown in Figure 4.4. The MMSE [Kurlowicz and Wallace, 1999] score is calculated after answering 20 questions by the participant, where the UPDRS rating scale [Perlmutter, 2009] is accomplished through interviews with PD patient and some family members.

The second part of the application consists of capturing the online handwriting data while accomplishing the handwriting template described in section 4.2. Handwriting signals were acquired using the digitizing tablet Wacom Intuos 5. The tablet itself does not provide visual feedback; therefore it was overlaid with white paper and the tip of the ink pen was replaced by a pencil lead so that the ink pen could be held in a normal fashion and allow for

[1]The application can be used for Parkinson, Alzheimer, and Parkinsonism (defined in chapter 2)

full visual feedback during writing. For interfacing we have used WinTAB compatible VBTablet ActiveX component [Bharadi and Kekre, 2009]. VBTablet is a library programmed in visual basic and developed for interfacing digitizers; this provides objects for scanners which are Wintab compatible. The Wintab driver is a windows driver provided by the tablet manufacturers to communicate between the digitizing tablet and Windows programs [Bharadi and Kekre, 2009]. Handwriting data are displayed in a picture box; where a connect button is used to attach the context of the digitizer to the picture box. Captured pen strokes at different pressure levels and some of the captured features are displayed in Figure 4.5-a, b, while Figure 4.5-c shows the plot of the dynamic data captured from the online handwriting given in Figure 4.5-a.

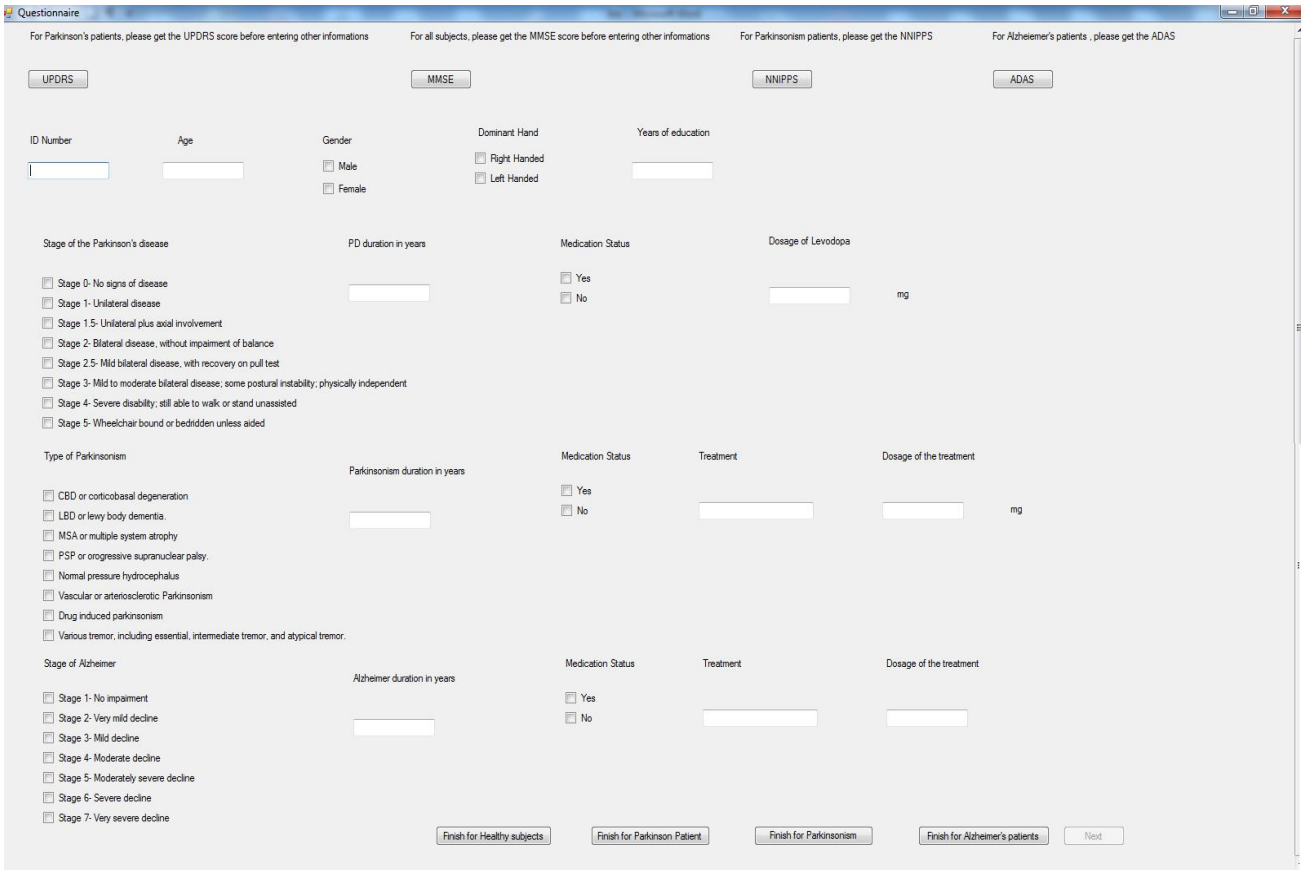



Figure 4.4. Questionnaire form including demographic and clinical information.

Online and offline handwriting data are saved for each subject. The online data will be used in the detection of the disease, where the offline digital documents will be used in word and pattern segmentation. Each word or pattern in the template is enclosed in a bounding box as shown in Figure 4.6. This type of segmentation was chosen to extract different features

(that are extracted from the online signals and not the offline image) related to words or patterns existing in the segment. Segmentation is manually done using Groundtruthing Environment for Document Images (GEDI) tool [Doermann et al., 2010]; where the image is ordered from top to bottom, and left to right, as illustrated in Figure 4.6. The segmented handwriting sample is represented in Figure 4.1-b. The output data file of this tool is an extensible markup language (XML) file (Figure 4.7) that contains the characteristics of each segment such as: the ID number of each segment, the content (text or pattern), the top left coordinate, the width, and the height of each segment (see Figure 4.6).

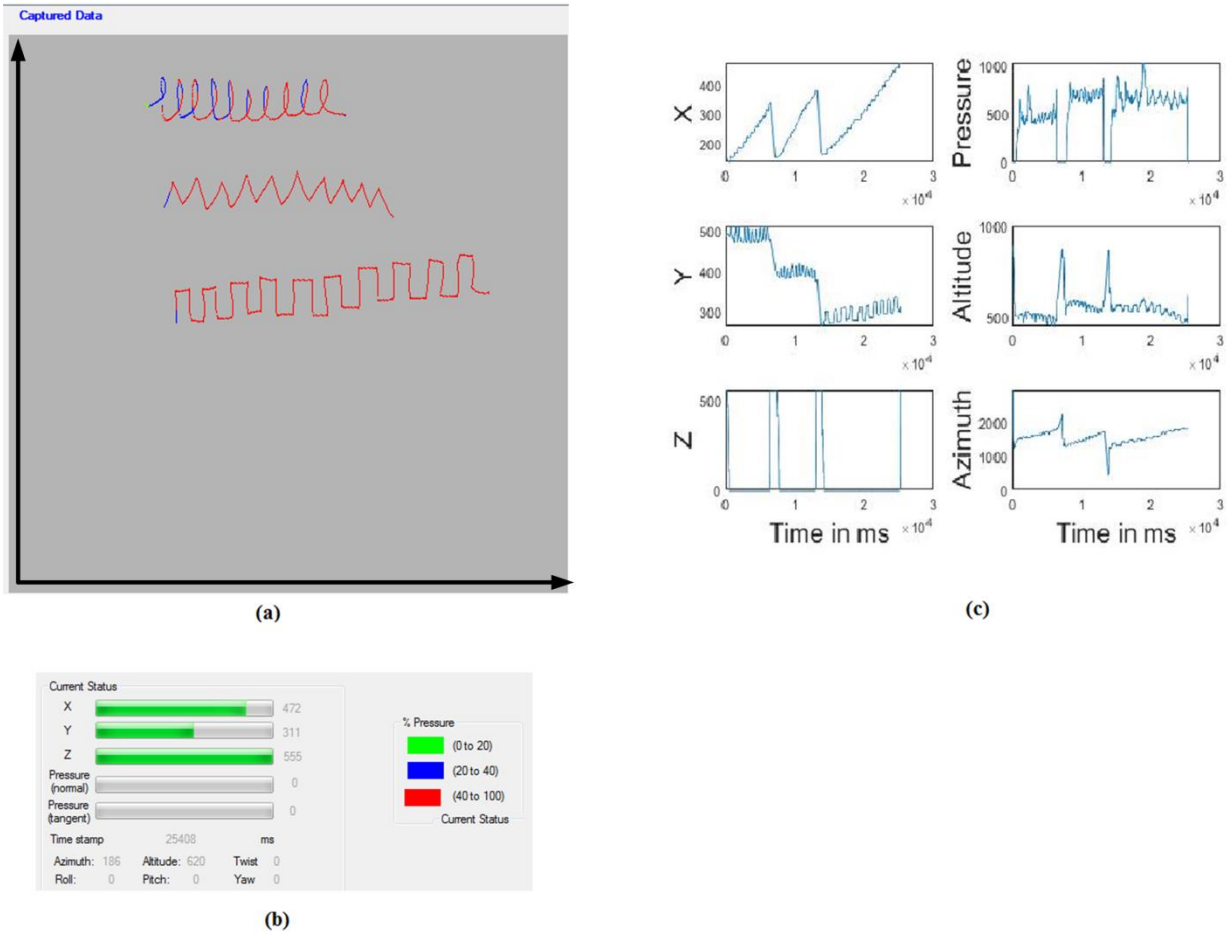



Figure 4.5. Online handwriting sample and its dynamic features: (a) captured pen strokes at different pressure levels, (b) the captured features, and (c) the plot of the dynamic data captured from the online handwriting.

In order to get the online data for each segment, an application written in C# (Figure 4.8) takes the "gedi" XML file and all the text files containing the online data of the whole image as inputs and returns a text file for each segment containing its specific online data. The size of the image is required in order to change the direction of the y axis (as shown

in Figure 4.9) before extracting the online data of each segment. Seven handwriting tasks were studied in this thesis; where the repetitive cursive letter (letter 'l'), the triangular wave, the rectangular wave, the repetitive "Monday", the repetitive "Tuesday", the repetitive subject's name, and the repetitive subject's last name represent the seven tasks *respectively*.

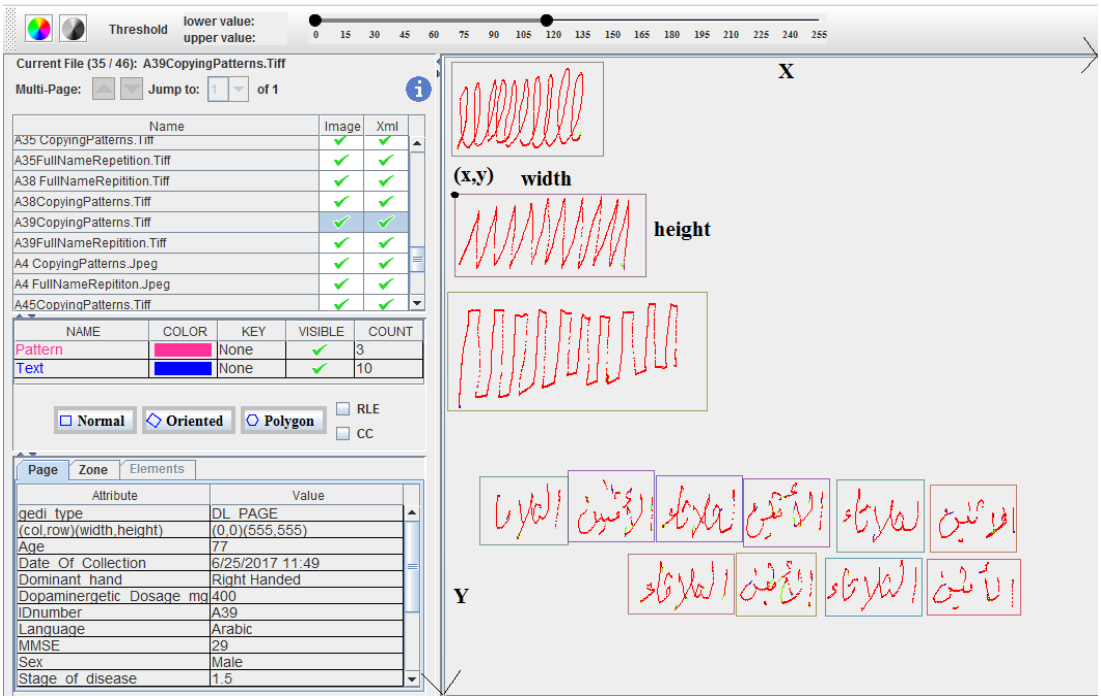



Figure 4.6. GEDI interface.

```
<?xml version="1.0" encoding="UTF-8"?>
<!--GEDI was developed at Language and Media Processing Laboratory, University of Maryland.-->

<GEDI xmlns="http://lamp.cfar.umd.edu/media/projects/GEDI/" GEDI_version="2.4" GEDI_date="07/29/2013">
	<USER name="catherinetaleb" date="2/5/2020 16:03" dateFormat="mm/dd/yyyy hh:mm">	</USER>
	<DL_DOCUMENT src="A39CopyingPatterns.Tiff" NrOfPages="1" docTag="xml">
		<DL_PAGE gedi_type="DL_PAGE" src="A39CopyingPatterns.Tiff" pageID="1" width="555" height="555" Dominant_hand="Right Handed" Dopaminergetic_Dosage_mg="400" IDnumber="A39"
Language="Arabic" MMSE="29" Sex="Male" Stage_of_disease="1.5" Status="Parkinson's patient medication ON" Total_UPDRS="22" UPDRS_I="3" UPDRS_II="8" UPDRS_III_="11" Years_of_Disease="10"
Years_of_education="5" Age="77" Date_of_Collection="6/25/2017 11:49">
			<DL_ZONE gedi_type="Pattern" id="1" col="6" row="4" width="146" height="91"> </DL_ZONE>
			<DL_ZONE gedi_type="Pattern" id="2" col="9" row="132" width="184" height="80"> </DL_ZONE>
			<DL_ZONE gedi_type="Pattern" id="3" col="2" row="229" width="250" height="115"> </DL_ZONE>
			<DL_ZONE gedi_type="Text" id="4" col="467" row="414" width="83" height="65" segmentation="word" contents="الثلاثين"> </DL_ZONE>
			<DL_ZONE gedi_type="Text" id="5" col="377" row="409" width="84" height="70" segmentation="word" contents="الاربعين"> </DL_ZONE>
			<DL_ZONE gedi_type="Text" id="6" col="287" row="408" width="83" height="66" segmentation="word" contents="الثلاثين"> </DL_ZONE>
			<DL_ZONE gedi_type="Text" id="7" col="203" row="405" width="83" height="64" segmentation="word" contents="الاربعين"> </DL_ZONE>
			<DL_ZONE gedi_type="Text" id="8" col="118" row="400" width="83" height="69" segmentation="word" contents="الثلاثين"> </DL_ZONE>
			<DL_ZONE gedi_type="Text" id="9" col="33" row="406" width="85" height="66" segmentation="word" contents="الاربعين"> </DL_ZONE>
			<DL_ZONE gedi_type="Text" id="10" col="464" row="486" width="90" height="54" segmentation="word" contents="الثلاثين"> </DL_ZONE>
			<DL_ZONE gedi_type="Text" id="11" col="366" row="485" width="93" height="56" segmentation="word" contents="الاربعين"> </DL_ZONE>
			<DL_ZONE gedi_type="Text" id="12" col="280" row="480" width="77" height="61" segmentation="word" contents="الثلاثين"> </DL_ZONE>
			<DL_ZONE gedi_type="Text" id="13" col="176" row="481" width="102" height="58" segmentation="word" contents="الاربعين"> </DL_ZONE>
		</DL_PAGE>
	</DL_DOCUMENT>
</GEDI>
```

Figure 4.7. An example of GEDI XML file.

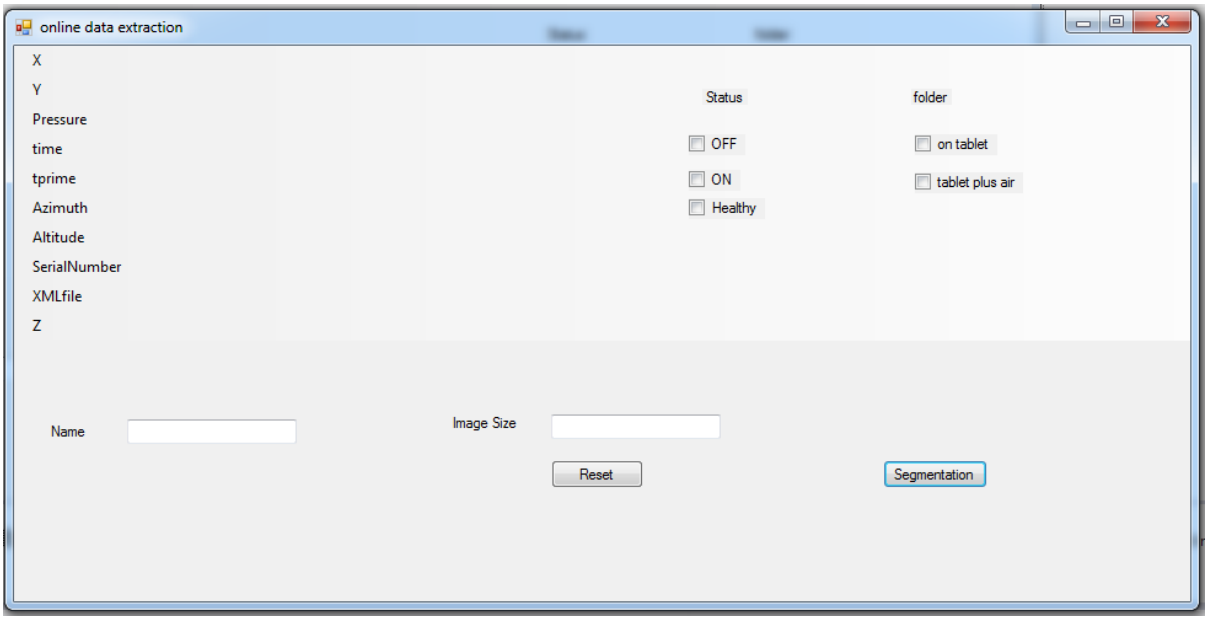



Figure 4.8. Application form used for online data extraction for each segment.

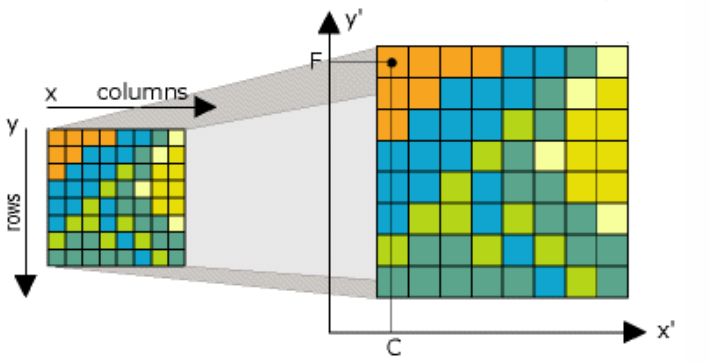



Figure 4.9. Changing the direction of the y axis of the image before extracting the online data for each segment.

The third part of the application is voice recording. As shown in Figure 4.10, this part is composed of two tasks (sustained vowel ‘a’ and text reading as mentioned in section 4.3), and for each task voice is recorded and saved as a wave sound file. The refresh sources button is used to select the source of recording audio; in our case the internal microphone of the Laptop (hp Elittbook 8570 w) is selected that produces a 2 channels sounds with 16-Bit depth and 44.1 KHz sampling rate. The familiar language of each participant should be selected for text reading part. Once the start button is pressed, a wave audio file will be opened where the recorded sound will be saved. Recording will start the time the start button is pressed until the

stop button is pressed. The tab control placed on the bottom of the form is for controlling the font size of the text to be read.

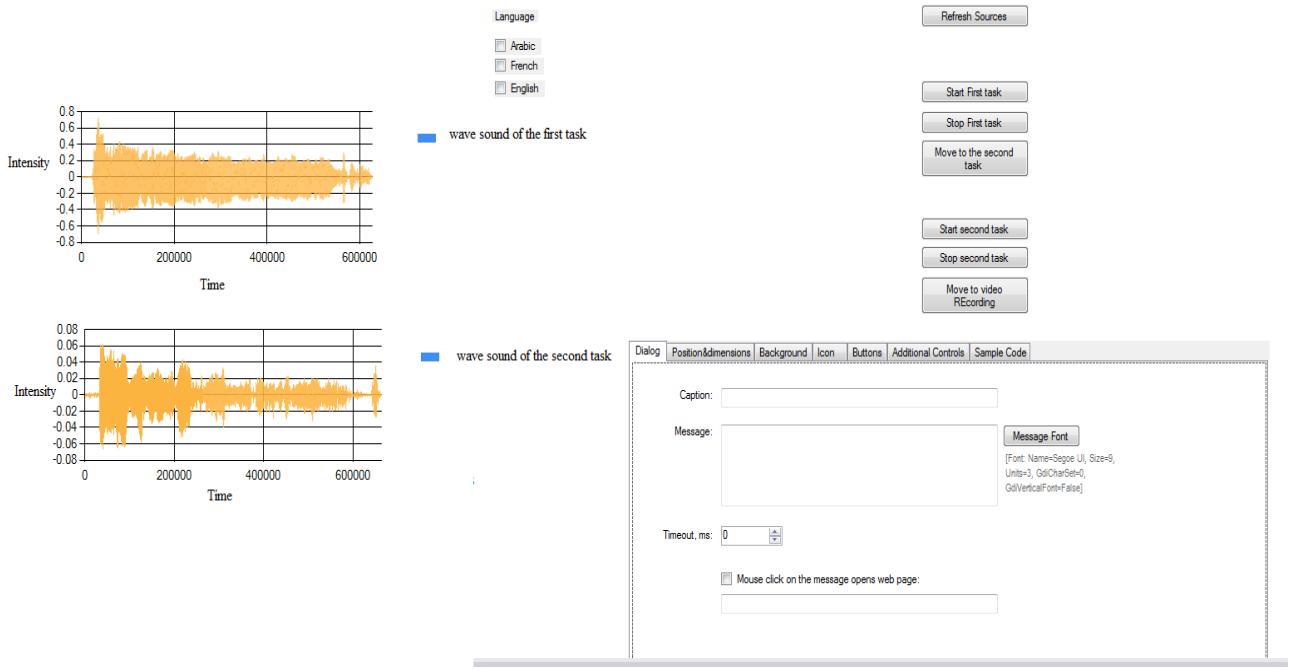



Figure 4.10. Voice recording application form.

Finally, the last part of the application is the face characteristics changes recording. In this part the participant will be asked to read a text that will appear on the screen of the PC while the webcam will start recording the variation of face characteristics. As shown in Figure 4.11, a text paragraph will appear in a textblock with the familiar language chosen by the subject in the previous part. The WebcamCtrl is used for displaying the video. Two combobox are used for displaying the audio and the video devices. In our case the video device is the Laptop Webcam (with a frame rate of 30 frames per second and a resolution of 640×480 pixels) and the audio device is the Microsoft expression encoder device (producing 2 channels sounds with 2-Bit depth and 48 KHz sampling rate). The start button will open a Windows Media Audio/Video file where the recorded video will be saved and start recording until the stop button is pressed. The Repeat button is used in case there is a need to repeat the record.


Figure 4.11. Face characteristics changes recording application form.

In conclusion, PDMultiMC is multimodal and multilingual database that includes seven handwriting tasks, two voice tasks, and one eye movements recording task; where the language representation is not balanced (31 Arabic, 9 French, and 2 English samples). This database will be soon released on the international association for pattern recognition technical committee number 11 (IAPR TC11). In this thesis, due to limitation time, we focused on handwriting and speech analysis for PD detection. Eye movements' recordings constitute a supplementary part in the database that was not studied in this thesis, but can be used for future work.

# 5 Automatic non Invasive PD Early Detection based on Handwriting

As mentioned in chapter 2, fingers, wrist, and arm play an important role in handwriting movement. The up-down strokes are controlled by fingers, where the left-right strokes are produced by the wrist. The forearm is responsible of extending the horizontal writing lines. PD patients have difficulties in coordinating wrist and fingers during writing. This coordination impairment may contribute to handwriting impairments. Based on this, handwritings are considered ideal to study motor control and to detect PD.

The detection of PD using handwriting analysis represents the largest part of our thesis, where it is divided into two parts: global hand-crafted features and SVM classifier, and short-term features and deep learning. This chapter includes a description of all the methodologies employed in this thesis for PD early detection based on handwriting analysis, where the handwriting samples are taken from the HandPDMultiMC subset described in chapter 4. Our goal here is to build a language-independent model for PD detection at the early stages where the motor symptoms are not severe, based on handwriting features. Here comes the importance of studying PD patients in their “on-state”, as mentioned in chapter 4, since dopamine treatment may reduce the motor symptoms, and building a general purpose feature set, which is language-independent and suitable for each task under assessment.

## 5.1 Classification of PD vs HC using global engineered features and support vector machine

While the direct goal of biological modeling is to describe data, it ultimately aims to find ways of fixing systems and enhancing understanding of system objectives, algorithms, and mechanisms [Kording et al., 2017]. Thanks to engineering applications, machine learning is making it possible to model data extremely well, without using strong assumptions about the modeled system. Machine learning can usually better describe data than biomedical models and thus provides engineering solutions.

The vast field of machine learning is a radically different way of approaching modeling that relies on minimal human insight. We focus here on the most popular sub-discipline, supervised learning, which assumes that the relationship between the measured variables and those to be predicted is in some sense simple. Supervised learning is where we have the input variables and the output labels and an algorithm is used to learn the mapping function from the input to the output. Machine learning techniques mostly differ by the nature of the function they use for predicting. The standard use for machine learning is to make a prediction based on something that can be measured. These developments in machine learning make it an important tool in biomedical research [Kording et al., 2017].

In this work, the seven handwriting tasks (described in chapter 4) recorded for each of the 32[2] subjects, 16 HC and 16 PD patients in their "on-state", are studied and analyzed. Since our database is small in size, we chose an SVM model for PD classification to start with that can be trained by a small number of samples, contrary to deep learning models that require larger number of training samples. The main contribution of this study is to find a feature selection approach for an improved PD early detection based on handwriting features suggested by Drotar et al. [Drotar et al., 2015a], [Drotar et al., 2015b].

### 5.1.1 Feature extraction

Raw data acquired by the digitizer are not enhanced by means of standard signal processing algorithms: filtering, noise reduction, and smoothing, because this could result in the loss of important information that can play an important role in PD detection. According to chapter 2, the pathophysiology of bradykinesia (slowness of movement), rigidity and tremor and how these symptoms are associated with PD were discussed in section 2.2. In addition, we have mentioned previously that the existence of micrographia can be related to the difficulty in controlling wrist and finger movements; leading to a difficulty in maintaining a constant force while writing [Dounskaia et al., 2009], [Teulings et al., 1997]. These motor symptoms can manifest in varying degrees and combinations with different individuals. Analyzing only one of these symptoms is insufficient to detect PD, since not all the patients develop the same symptoms. To deal with this, we tried to investigate what handwriting

[2]During this work, the HandPDMultiMC size was 32 (16 PD and 16 HC), then it has been enlarged to 42 (21 PD and 21 HC).

characteristics permit to best assess most of the motor symptoms. The relationship between the characteristic motor symptoms and the handwriting measurements is shown in Figure 5.1. To assess micrographia, writing size should be investigated. Bradykinesia or slowness of movement can be detected via movement speed and time measurements. Muscles rigidity can be detected by the amount of pen tip pressure during writing task, since patients will be unable to retain Pen-pressure due to rigidity. Finally, tremor or irregular muscle contractions introduce randomness to the movement during handwriting [Smite et al., 2014].

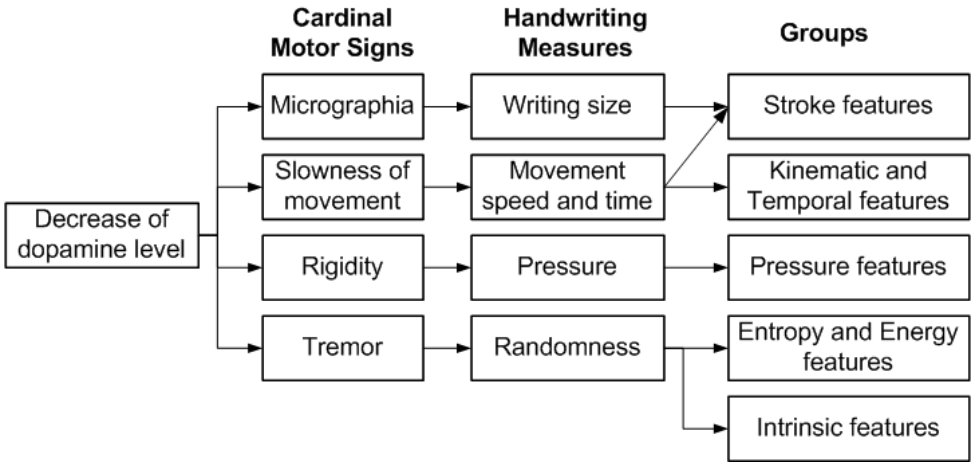



Figure 5.1. Relationship between PD motor symptoms and handwriting measurements.

As shown in Figure 5.1, the measurements that reflect the decrease of Dopamine levels are distributed into 5 groups: stroke features, kinematic and temporal features, pressure features, entropy and energy features, and intrinsic features. Traditional measurement methods to process handwriting signals are used for feature extraction. The extracted features can be either a single value or a sequence of values extracted through time [Drotar et al., 2013]. In case there is a resulting sequence, 5 basic functional features are computed to represent it: mean median, STD, 1st percentile, and 99th percentile [Drotar et al., 2015a]. These features can either be extracted when the pen is touching the screen (on-paper) or when the pen is in the proximity of the surface (in-air). In this work only on-paper features are studied.

### 5.1.1.1 Stroke features

A stroke is defined as a continuous line written by the subject. It can be one of two types: ‘on-paper’ stroke, which is a stroke, written on the paper, or ‘in-air’ stroke, which is a stroke that the pen creates when it does not touch the paper. In order to capture ‘in-air’ and ‘on-paper’ strokes, a button status binary vector is needed. The values of this vector are 0 for pen-up (in air movement) when the pressure is equal to zero and 1 for pen-down (on paper movement) when the pressure is not zero. An ‘on-paper’ stroke starts when the button status moves from 0 to 1 and ends when the button status moves from 1 to 0. Dually; the ‘in-air’ stroke starts when the button status moves from 1 to 0 and ends when the button status moves from 0 to 1. As mentioned before, only the on-paper strokes are considered. The list of computed stroke features proposed by [Drotar et al., 2015b] are provided in Table 5.1, where single value features are denoted as s and vector features are denoted as v. For each stroke, we extract its speed, time, height, width, number of changes in the velocity (NCV), and number of changes in the acceleration (NCA). In definitive, we first compute get vectors of features that are replaced in a second by their statistics mentioned before. The total number of extracted stroke features is equal to 6×5=30.

Table 5.1. List of stroke features extracted.

| Feature Class | Feature | s/v | Description | Calculation |
|---|---|---|---|---|
| Stroke Features | Stroke speed | v | Trajectory during stroke divided by stroke duration | |
| | Stroke time | v | Stroke duration | |
| | Stroke height | v | Difference between the maximum y position and the minimum y position of a stroke | |
| | Stroke width | v | Difference between the maximum x position and the minimum x position of a stroke | |
| | NCV | v | Number of changes in the velocity per stroke | The number of zero crosses in the acceleration of each stroke. |
| | NCA | v | Number of changes in the acceleration per stroke | The number of zero crosses in the jerk of each stroke. |
| Total number of features | 30 | | | |

### 5.1.1.2 Temporal and kinematic features

The second group of features proposed in [Drotar et al., 2015b] includes kinematic and temporal features captured during handwriting task. These include: velocity, acceleration, jerk, horizontal and vertical velocity, horizontal and vertical acceleration, and horizontal and vertical jerk. These features are defined and described in Table 5.2. The total number of extracted features is equal to 9×5+1=46.

Table 5.2. List of Kinematic and temporal features extracted.

| Feature Class | Feature | s/v | Description | Calculation |
|---|---|---|---|---|
| Kinematic and temporal features | Horizontal and vertical velocity | v | Velocity in horizontal direction and velocity in vertical direction | $v_x[n] = \frac{x[n] - x[n-1]}{t[n] - t[n-1]}$ $v_y[n] = \frac{y[n] - y[n-1]}{t[n] - t[n-1]}$ |
| | Horizontal and vertical acceleration | v | Acceleration in horizontal direction and vertical direction | $a_x[n] = \frac{v_x[n] - v_x[n-1]}{t[n] - t[n-1]}$ $a_y[n] = \frac{v_y[n] - v_y[n-1]}{t[n] - t[n-1]}$ |
| | Horizontal and vertical jerk | v | Jerk in horizontal and vertical direction | $j_x[n] = \frac{a_x[n] - a_x[n-1]}{t[n] - t[n-1]}$ $j_y[n] = \frac{a_y[n] - a_y[n-1]}{t[n] - t[n-1]}$ |
| | Velocity | v | Rate at which the position of a pen changes with time | $v[n] = \sqrt{v_x[n]^2 + v_y[n]^2}$ |
| | Acceleration | v | Rate at which the velocity of a pen changes with time | $a[n] = \frac{v[n] - v[n-1]}{t[n] - t[n-1]}$ |
| | Jerk | v | Rate at which the acceleration of a pen changes with time | $j[n] = \frac{a[n] - a[n-1]}{t[n] - t[n-1]}$ |
| | Movement time(on surface) | s | Time spent on surface during writing | |
| Total number of features | 46 | | | |

### 5.1.1.3 Pressure features

The pressure of each stroke starts with rising edge, continues with slowly varying main part and ends with falling edge as shown in Figure 5.2. The boundary between edges and main part is given by median of signal pressure. According to [Drotar et al., 2015a], for each part of the stroke (rising edge, main part, and falling edge) we calculate the number of changes in pressure (NCP), and the correlation coefficients between pressure and vertical velocity, pressure and horizontal velocity, pressure and vertical acceleration, and pressure and

horizontal acceleration. The extracted features are resumed in Table 5.3. The total number of extracted pressure features is equal to 5×5×3=75.

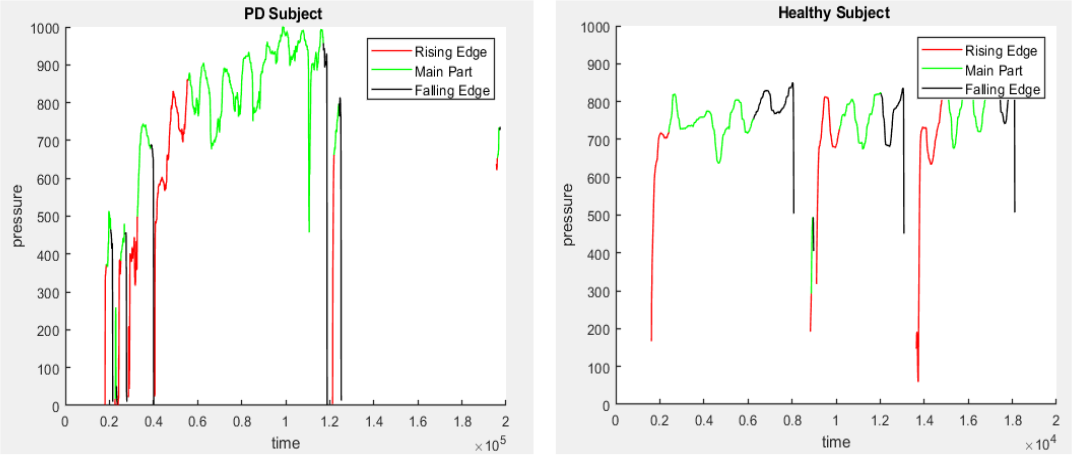



Figure 5.2. Pressure plot of task1 for HC and PD.

Table 5.3. List of Pressure features extracted.

| Feature Class | Feature | s/v | Description | Calculation |
|---|---|---|---|---|
| Pressure features extracted for each part of the stroke | NCP | v | Number of changes in pressure per stroke | The number of zero crosses in the pressure velocity of each stroke |
| | Correlation coefficient between pressure and horizontal velocity | v | Correlation between stroke pressure and stroke horizontal velocity | |
| | Correlation coefficient between pressure and vertical velocity | v | Correlation between stroke pressure and stroke vertical velocity | |
| | Correlation coefficient between pressure and horizontal acceleration | v | Correlation between stroke pressure and stroke horizontal acceleration | |
| | Correlation coefficient between pressure and vertical acceleration | v | Correlation between stroke pressure and stroke vertical acceleration | |
| Total number of features | 75 | | | |

### 5.1.1.4 Entropy and energy features

The digital representation of handwriting as a physiologically based time series is the result of several interacting physiological mechanisms like tremor, or irregular muscle contractions that introduce randomness to the movement during handwriting. This randomness is difficult to analyze using only kinematic measures. Entropy and energy features are used to measure the randomness or uncertainty of a signal as proposed by Drotar et al. [Drotar et al., 2015b].

#### 5.1.1.4.1 Entropy features

Many entropy definitions exist. The widely established and well known are Shannon and Rényi entropies defined as follows [Drotar et al., 2015b]:

$$H_{Shannon}(L) = -\sum_{l\varepsilon\, L} p(l) log_2 p(l) \tag{5.1}$$

$$H_{Rényi,r}(L) = \frac{1}{1-r} log(\sum_{i=1}^{n} {p_i}^r) \tag{5.2}$$

where $L$ is the signal, $p(l)$ refers to the probability density function, n represents the length of the signal $L$, and finally $r \geq 0$ is defined as Rényi entropy order. Shannon entropy is a particular case of Rényi entropy where r tends to be equal to 1. In our work, Shannon entropy and second and third order Rényi entropy are calculated for both X and Y signals and are summarized in Table 5.5. The probability density function $p(l)$ is computed using kernel density estimation with a Gaussian kernel [Walter, 2003].

##### 5.1.1.4.1.1 Kernel density estimation

The objective of many investigations is to estimate the probability density function defined by $p(l)$ from a sample of observations $l_1$, $l_2$, …., $l_n$ [Walter, 2003]. The method to estimate $p(l)$ is based on smoothing using a kernel. From the definition of the pdf or a random variable $L$, one has that [Walter, 2003]:

$$P(l-h < L < l+h) = \int_{l-h}^{l+h} p(t)dt \simeq 2\mathrm{h}\, p(l) \tag{5.3}$$

and hence

$$p(l) \simeq \frac{P(l-h < L < l+h)}{2h} \tag{5.4}$$

The above probability can be estimated by a relative frequency in the sample, hence

$$\hat{p}(l) = \frac{1}{2h} \frac{numberofobservations\ in(l-h, l+h)}{n} \tag{5.5}$$

An alternative way to represent $\hat{p}(l)$ is:

$$\hat{p}(l) = \frac{1}{n} \sum_{i=1}^{n} w(l - l_i, h) \tag{5.6}$$

where the weighting function $w(t,h)$ is of the form:

$$w(t,h) = \frac{1}{h} \mathrm{K}(\frac{t}{h}) \tag{5.7}$$

K is a function of a single variable called the kernel. The kernel determines the shape of the weighting function, and the parameter h (called the bandwidth or smoothing constant) determines the amount of smoothing applied in estimating $\hat{p}(l)$. The most common kernel functions are mentioned and defined in Table 5.4 [Walter, 2003].

Table 5.4. Most common kernel functions.

| **Kernel** | ***K(t)*** |
|---|---|
| Epanechnikov | $\frac{\frac{3}{4}\left(1-\frac{1}{5}t^2\right)}{\sqrt{5}}\ for\ \lvert t\rvert < \sqrt{5}, 0\ otherwise$ |
| Biweight | $\frac{15}{16}(1-t^2)^2\ for\ \lvert t\rvert < 1, 0\ otherwise$ |
| Triangular | 1-$\lvert t\rvert$ for $\lvert t\rvert$<1, 0 otherwise |
| Gaussian | $\frac{1}{\sqrt{2\pi}} e^{-(\frac{t^2}{2})}$ |
| Rectangular | $\frac{1}{2}\ for\ \lvert t\rvert < 1, 0\ otherwise$ |

Kernel estimation of pdfs is characterized by the kernel K and the bandwidth h. These two components must be selected in a way to optimize the properties of $\hat{p}(x)$. The properties of kernel estimators are defined by the mean squared error (MSE), the bias, and the variance defined as follows [Walter, 2003]:

$$\begin{aligned} \mathrm{MSE}(\hat{p}(l)) &= \mathrm{E}\left(\hat{p}(l) - p(l)\right)^2 \\ &= (E(\hat{p}(l) - p(l))^2 + \mathrm{E}(\hat{p}(l) - E(p(l)))^2 \\ &= \mathrm{Bias}^2(\hat{p}(l)) + \mathrm{Var}\,(\hat{p}(l)) \end{aligned} \tag{5.8}$$

A measure of the global accuracy of $\hat{p}(l)$ is defined by the mean integrated squared error (MISE):

$$\text{MISE}(\hat{p}) = \int_{-\infty}^{+\infty} \text{Bias}^2(\hat{p}(l))dl + \int_{-\infty}^{+\infty} Var(\hat{p}(l))dl \tag{5.9}$$

Based on equations (5.6) to (5.9), the MISE ($\hat{p}$) can we written as:

$$\text{MISE}(\hat{p}) \approx \frac{1}{4}h^4 {k_2}^2 \beta(p) + \frac{1}{nh} j_2 \tag{5.10}$$

where $k_2 = \int z^2 K(z)dz$, $j_2 = \int K(z)^2 dz$, and $\beta(p) = \int p''(l)^2 dl$.

According to equation (5.10), MISE is a function of the kernel and the bandwidth. For a given kernel, there is an optimal value of h which minimizes MISE. This optimal bandwidth can be found by minimizing equation (5.10) with respect to h by setting the first derivative equal to zero. The optimal bandwidth for a given kernel and pdf is defined in equation (5.11) where $\gamma(K) = j_2 {k_2}^{-2}$.

$$h_{opt} = \left(\frac{1}{n}\frac{\gamma(K)}{\beta(p)}\right)^{\frac{1}{5}} \tag{5.11}$$

As mentioned in Table 5.4, a range of kernel functions are commonly used. Here in this work the Gaussian kernel is applied due to its convenient mathematical properties. In this case, the optimal bandwidth defined in equation (5.12) where $\sigma$ refers to the sample's STD [Walter, 2003].

$$h_{opt} = \left(\frac{4}{3n}\right)^{\frac{1}{5}}\sigma \tag{5.12}$$

To apply (5.12) one has to estimate $\sigma$. Calculating the sample variance is not robust since it may overestimates $\sigma$ if some extreme observations are present, which will increase the estimated bandwidth even more. To overcome these problems, some authors proposed to estimate $\sigma$ from the median absolute deviation (MAD), since MAD is a more robust estimator of scale than the sample STD [Rousseeuw et al., 1993]. In order to use MAD as a consistent estimator of $\sigma$, equation (5.13) must be used, where I is a constant scalar factor, which depends on the distribution, and MAD is defined in (5.14). For normally distributed data, I $\approx 1.4826$ [Rousseeuw et al., 1993].

$$\hat{\sigma} = I.MAD \tag{5.13}$$

$$\text{MAD} = \text{median}(|\text{L-median(L)}|) \tag{5.14}$$

#### 5.1.1.4.2 Energy features

In the traditional signal processing literature, energy is known as the average of the sum of the squares of the magnitude of the signal L. This energy is known as the conventional energy and defined as follows:

$$CE(L) = \sum_{i=1}^{n} {l_i}^2 \tag{5.15}$$

Using the conventional view of the energy, it is easy to see that two tones at 10 Hz and 1000 Hz of unit-amplitude have the same energy. However, the energy required to produce the signal of 1000 Hz is much greater than that for the 10 Hz signal. The energy to generate a sinusoidal signal depends on both amplitude and frequency. Here comes the definition of Teager Kaiser energy (TKE) that is a function of both amplitude and frequency and defined in equation (5.16), where r≥ 0 is the order of the TKE operator [Boudra and Salzenstein, 2018].

$$TKE_r(L) = \sum_{i=r}^{n-r} {l_i}^2 - l_{i+r}.l_{i-r} \tag{5.16}$$

Conventional energy and first order TKE are calculated for both X and Y coordinates signals. After calculating the signals energy, the signal to noise ratios can be obtained once the noise variance is estimated. What we mean by noise here is the tremor and the irregular muscle contractions. Then signal to noise ratios are defined as:

$$SNR_{CE} = \frac{CE(L)/n}{\text{N}(L)} \tag{5.17}$$

and

$$SNR_{TKE} = \frac{TKE(L)/n}{\text{N}(L)} \tag{5.18}$$

In order to get the signal to noise ratios in (5.17) and (5.18), the noise variance $\mathcal{N}(L)$ is estimated by using robust smoothing method [Gracia, 2010].

#### 5.1.1.4.2.1 Robust smoothing method for noise variance estimation

Smoothing is used to reduce noise while keeping the most important data. The one-dimensional noisy signal can be represented as follows [Gracia, 2010]:

$$l=\hat{l}+\epsilon \tag{5.19}$$

where $\epsilon$ represents a Gaussian noise with zero mean and unknown variance (that we need to estimate) , and $\hat{l}$ supposed to be smooth. Smoothing *x* must be applied in order to find the best estimate of $\hat{l}$. One of the smoothing data techniques consists in minimizing a criterion that balances the fidelity to the data, measured by the residual sum of squares (RSS) and a penalty term (P) that reflects the roughness of the smooth data [Gracia, 2010]. This leads to minimize (5.20), where || || refers to the Euclidean norm; s is a real positive scalar that controls the degree of smoothing, and the penalty P is defined in (5.21). D is a tridiagonal square matrix defined in (5.22), where $h_i$ represents the step between $\hat{l}_i$ and $\hat{l}_{i+1}$.

$$F(\hat{l}) = RSS + sP(\hat{l}) = \left|\left|\hat{l}-l\right|\right|^2 + sP(\hat{l}) \tag{5.20}$$

$$P(\hat{l}) = \left|\left|D\hat{l}\right|\right|^2 \tag{5.21}$$

$$D=\begin{pmatrix} \frac{-1}{{h_1}^2} & \frac{1}{{h_1}^2} & 0 & \dots & 0 \\ \frac{2}{h_1(h_1+h_2)} & \frac{-2}{h_1h_2} & \frac{2}{h_3(h_2+h_3)} & 0 & 0 \\ 0 & \frac{2}{h_2(h_2+h_3)} & \frac{-2}{h_2h_3} & \ddots & 0 \\ \vdots & 0 & \ddots & \ddots & \frac{2}{h_{n-1}(h_{n-2}+h_{n-1})} \\ 0 & \dots & 0 & \frac{1}{{h_{n-1}}^2} & \frac{-1}{{h_{n-1}}^2} \end{pmatrix} \tag{5.22}$$

Using both equations (5.20) and (5.21), minimizing $F(\hat{l})$ returns the following linear system [Gracia, 2010]:

$$(I_n + sD^T D)\hat{l} = l \tag{5.23}$$

As shown in (5.23), $\hat{l}$ depends on the degree of smoothing parameter s. In order to avoid over- or under-smoothing as much as possible, s is selected in a manner of minimizing the generalized cross validation (GCV) score defined in (5.24). $(\lambda_i^2)_{i=1,\dots,n}$ refer to the eigenvalues of $D^T D$.

$$GCV(s) = \frac{n\sum_{i=1}^{n}(\hat{l}_i - l_i)^2}{(n - \sum_{i=1}^{n}(1 + s\lambda_i^2)^{-1}))^2} \tag{5.24}$$

From (5.24) it is obvious that GCV is a function of noise variance that we need to estimate. To solve this issue, equation (5.23) is solved in order to get $\hat{l}$, where the data is assumed equally spaced ($h_i$ =1, $\forall\, i$) for simplicity. The GCV score given by equation (5.24) becomes [Gracia, 2010]:

$$GCV(s) = \frac{n\sum_{i=1}^{n}\left(\frac{1}{1 + s\lambda_i^2} - 1\right)^2 DCT_i^2(l)}{\left(n - \sum_{i=1}^{n}\frac{1}{1 + s\lambda_i^2}\right)^2} \tag{5.25}$$

where $DCT_i$ refers to the $i^{th}$component of the discrete cosine transform.

The optimal smoothing parameter s is obtained by minimizing the GCV defined in (5.25). The noise variance is then obtained by using both equations (5.24) and (5.25) as follows [Gracia, 2010]:

$$\sum_{i=1}^{n}(\hat{l}_i - l_i)^2 = n\sum_{i=1}^{n}\left(\frac{1}{1 + s_{opt}\lambda_i^2} - 1\right)^2 DCT_i^2(l) \tag{5.26}$$

The extracted features are provided in Table 5.5, where the total number of features extracted is equal to 14.

Table 5.5. List of Entropy and energy features extracted.

| Feature Class | Feature | s/v | Description | Calculation |
|---|---|---|---|---|
| Entropy and energy features | H(X), H(Y) | s | Shannon entropy of X and Y position | $H(L) = -\sum_{l \varepsilon L} p(l) log_2 p(l)$ |
| | $H_{R,r}$(X), $H_{R,r}$(Y) | s | Rényi entropy second and third order of X and Y position | $H_{R,r}(\mathrm{L}) = \frac{1}{1-r} \log(\sum_{i=1}^{n} {p_i}^r)$ |
| | CE(X), CE(Y) | s | Conventional energy of X and Y position | $CE(L) = \sum_{i=1}^{n} {l_i}^2$ |
| | $TKE_1$(X), $TKE_1$( Y) | s | First order Teager-Kaiser energy of X and Y position | $TKE_r(\mathrm{L}) = \sum_{i=r}^{n-r} {l_i}^2 - l_{i+r}.l_{i-r}$ |
| | $SNR_{CE}$ (X), $SNR_{CE}$ (Y) | s | Signal to noise ratio calculated from the conventional energy for both X and Y positions | $SNR_{CE} = \frac{CE(L)/N}{\mathrm{N}(L)}$ |
| | $SNR_{TKE}$ (X), $SNR_{TKE}$ (Y) | s | Signal to noise ratio calculated from the Teager-Kaiser energy for both X and Y positions | $SNR_{TKE} = \frac{TKE(L)/N}{\mathrm{N}(L)}$ |
| Total number of features | 14 | | | |

### 5.1.1.5 Intrinsic features

Handwritten dynamics signals are multi-frequency, non-periodic, and arbitrary. They are considered as non-stationary signals, and non-stationary signals have statistical properties that vary as a function of time and should be analyzed differently than stationary data [Kaslovsky and Meyer, 2010]. The non-stationary data can be represented as a superposition of fast oscillations onto slow oscillations. The goal of Empirical Mode Decomposition (EMD) is to represent a signal as an expansion of adaptively defined basis functions with well-defined frequency localization. Each basis function called an Intrinsic Mode Function (IMF) should be physically meaningful, representing one frequency (nearly monochromatic) [Kaslovsky and Meyer, 2010]. An IMF resulting from the EMD shall satisfy only the following requirements:

1. The number of IMF extrema (the sum of the maxima and minima) and the number of zero-crossings must either be equal or differ at most by one.
2. At any point of an IMF, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima shall be zero.

To decompose the signal, the EMD algorithm works as follows [Kaslovsky and Meyer, 2010]:

- The local maxima and minima of the signal are calculated and form 2 envelopes.
- The mean of the 2 envelopes is calculated and subtracted from the original signal.
- A number of iterations should be done until the number of extrema is equal to the number of zeros plus or minus one, and the mean of the envelopes defined as a function of local maxima and minima is zero. The first IMF is obtained.
- The next IMF can be obtained by subtracting the previously extracted IMF from the original signal and repeating the above described procedure once again.
- This continues until all IMFs are extracted.
- The sifting process usually stops when the residue, for example, contains no more than two extrema.

The number of IMFs depends on the nature and length of a signal. Figure 5.3 shows the IMFs decomposition of X signal for a HC and a PD**.** We can see that the first two IMFs are wide band and capture most of the noise, while the following IMFs contain both signal and noise, where the last IMFs are nearly monochromatic. The noise in the signal seems to correspond to the tremor or jerk, and therefore is used to identify the presence of the disease. This explains the reason of focusing on the first two IMFs.

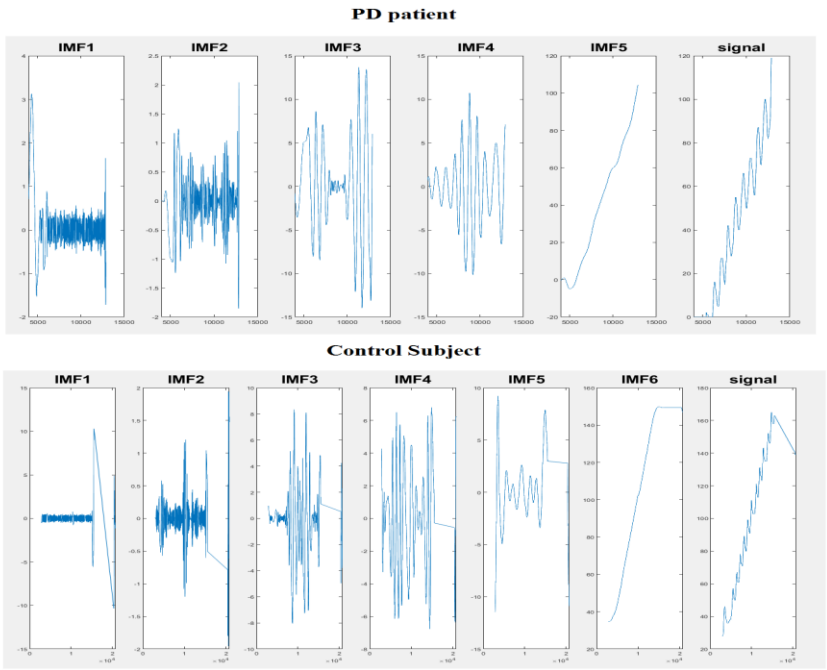



Figure 5.3. IMFs decomposition of X-coordinate in task 1 for HC and PD.

Shannon and Rényi entropies as well as conventional and Teager-Keiser energies described in section 5.1.1.4 are calculated for the first 2 IMFs of each of the X and Y position [Drotar et al., 2015b]. For the SNRs, the higher IMFs are assumed to contain useful signals, where the first 2 IMFs contain noise. They are calculated based on equation (5.27). The intrinsic features extracted are summarized in Table 5.6, where the total number of features is 24.

$$SNR_{ICE} = \frac{\sum_{k=3}^{N_{IMF}} CE(IMF[k])}{\sum_{k=1}^{2} CE(IMF[k])} \qquad SNR_{ITKE} = \frac{\sum_{k=3}^{N_{IMF}} TKE(IMF[k])}{\sum_{k=1}^{2} TKE(IMF[k])} \tag{5.27}$$

The total number of extracted global features is 189. As mentioned in chapter 4, tasks 1, 2, and 3 consist of single pattern segment, where tasks 4 to 7 consist of different word segments. For each segment (whether word or pattern), the 189 features described above are extracted. The aim here is to form one feature vector per task with information of all segments. To do this, for each task we extract features per segment, and then we calculate the average across the different segments. Once the feature vectors are extracted for each task, we get the mean feature vector across the different tasks. An overview of the feature extraction system is shown in Figure 5.4.

For task1, task2, and task3 the number of loops differ from subject to another. Entropy and intrinsic features depends on the amount of information existing in the data. In order to get more accurate features, we have proposed to divide each cursive repetitive task into loops, and overlaying all the loops together as shown in Figure 5.5. To do that, the local maximum and minimum in each task are found using scale-space theory method instead of the local derivative method. This approach is designed by [Liutkus, 2015] to find the local maximum of a given data, and can also be applied to find the local minimum by finding the local maximum of the negation of data. This approach performs iterative smoothing of the input data with increasing length-scales and then defines a peak as a data point that remains a local maximum for many such filtering. Formally, the local maxima are identified after each filtering operation and then associated to the maxima identified with the previous length-

scales. A score is then added to the criterion for these latter points that notably depends on the length scale. This strategy enforces picks that remain local maxima even after many smoothing operations. At the end of the process, the peaks are identified as the points having the largest score. The local maximum and minimum are the red and green dots respectively in Figure 5.5. Once the extrema are found, each loop must be composed of 1 local maximum and 2 local minimum (one at the start and one at the end). The X coordinates in each loop are then rescaled by subtracting the local maximum's X coordinates. Entropy and intrinsic features are calculated for the new X-position data.

Table 5.6. List of Intrinsic features extracted.

| **Feature Class** | **Feature** | **s/v** | **Description** | **Calculation** |
|---|---|---|---|---|
| Intrinsic Features | H(IMF1xposition), H(IMF1yposition) | s | Shannon entropy of the first empirical mode decomposition of X and Y position | |
| | H(IMF2xposition), H(IMF2yposition) | s | Shannon entropy of the second empirical mode decomposition of X and Y position | |
| | $H_{R,n}$(IMF1xposition), $H_{R,n}$(IMF1yposition) | s | Rényi entropy second and third order of first empirical mode decomposition of X and Y position | |
| | $H_{R,n}$(IMF2xposition), $H_{R,n}$(IMF2yposition) | s | Rényi entropy second and third order of second empirical mode decomposition of X and Y position | |
| | CE(IMF1xposition), CE(IMF1yposition) | s | Conventional energy of first empirical mode decomposition of X and Y position | |
| | CE(IMF2xposition), CE(IMF2yposition) | s | Conventional energy of second empirical mode decomposition of X and Y position | |
| | $TKE_1$(IMF1xposition), $TKE_1$(IMF1yposition) | s | Teager-Kaiser energy of first empirical mode decomposition of X and Y position | |
| | $TKE_1$(IMF2xposition), $TKE_1$(IMF2yposition) | s | Teager-Kaiser energy of second empirical mode decomposition of X and Y position | |
| | $SNR_{ICE}$ (X), $SNR_{ICE}$ (Y) | s | Signal to noise ratio based on intrinsic conventional energy | $SNR_{ICE} = \frac{\sum_{k=3}^{Nimf} CE(IMF[k])}{\sum_{k=1}^{k=2} CE(IMF[k])}$ |
| | $SNR_{ITKE}$ (X), $SNR_{ITKE}$ (Y) | s | Signal to noise ratio based on intrinsic Teager-Kaiser energy | $SNR_{ITKE} = \frac{\sum_{k=3}^{Nimf} TKE(IMF[k])}{\sum_{k=1}^{k=2} TKE(IMF[k])}$ |
| Total number of features | 24 | | | |

## 5.1.2 Feature selection

The set of features obtained is large, while the database includes a limited number of samples which suggests an existing risk of falling in a curse of dimensionality. A two-stage feature selection approach was applied to remove the irrelevant features. The first stage consists of a pure statistical analysis of the data in order to reduce the features set to a smaller subset identified as necessary and sufficient to describe the target concept. The set of selected features by statistical tests remains large. It is further reduced to a smaller subset of features in the second stage by applying a suboptimal approach that provides a kind of benchmark of the relevance of the features in the desired task.

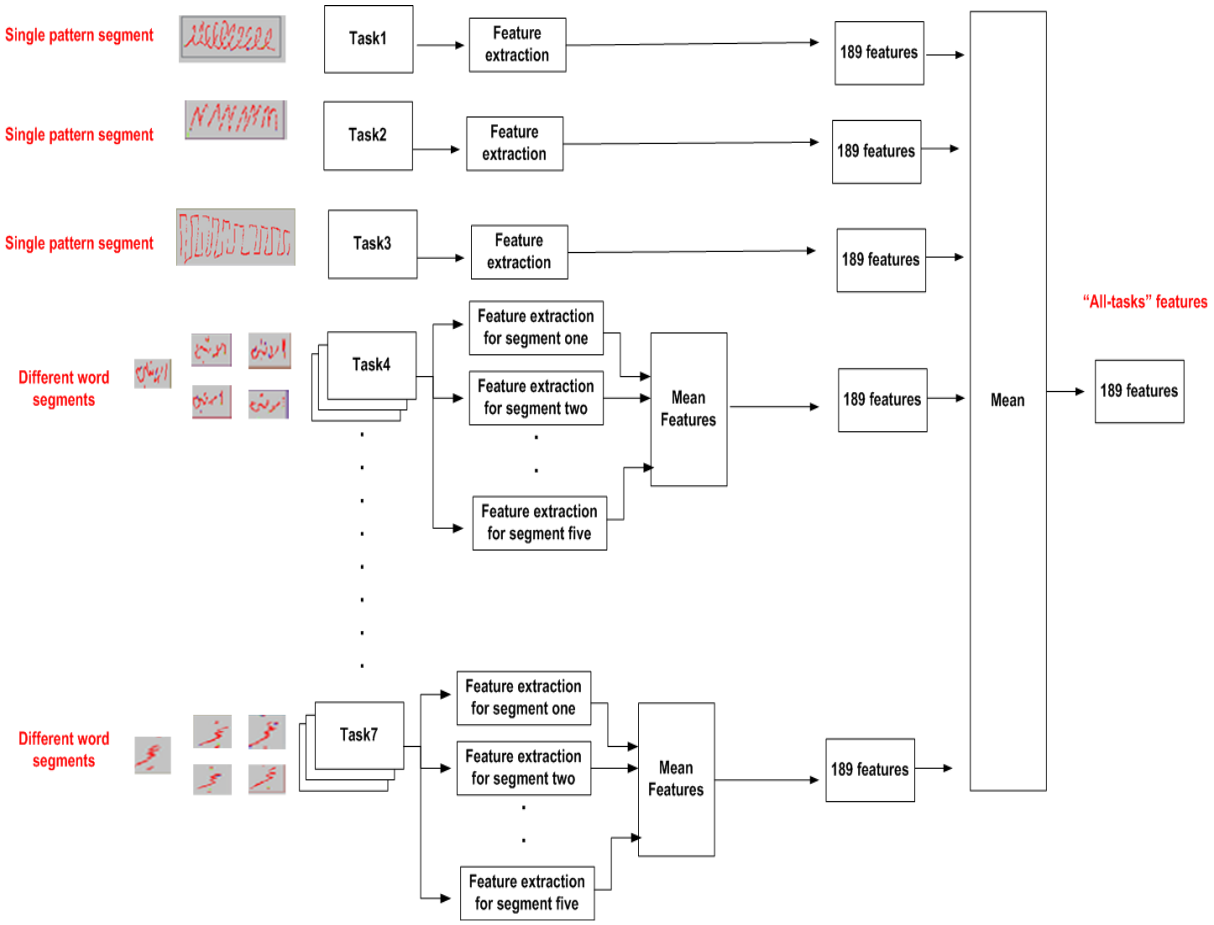



Figure 5.4. Overview of the feature extraction system.

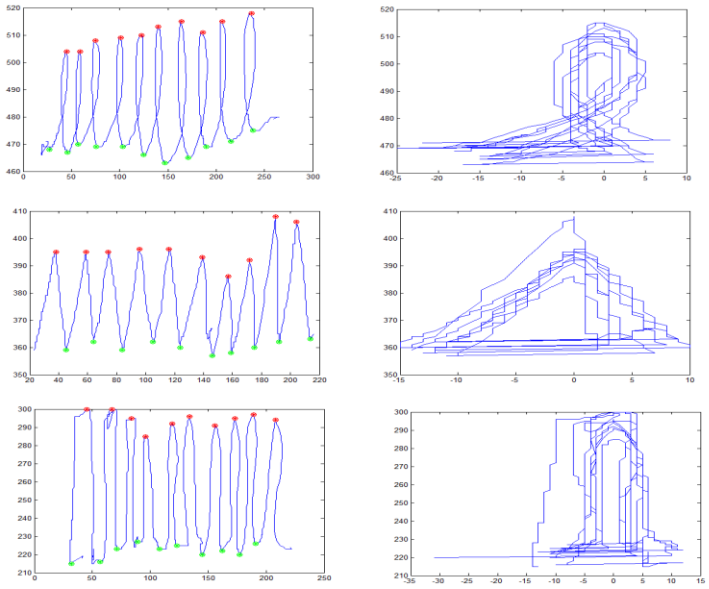


Figure 5.5. Overlaying loops in cursive repetitive tasks.

### 5.1.2.1 Feature selection based on statistical tests

The key idea here is to apply a statistical test on each feature with the hypothesis to validate being if the underlying processes for PD and HC subjects are independent. Statistical tests should conform to the sample features distribution and pairing. But in order to select the best test, the number of groups should also be considered. The main tests for each situation are summarized in Figure 5.6.

Shapiro-Wilk test should be used to decide whether the sample features distribution is normal or not [Royston, 1993]. After that and based on the normality test, two or more sets of data will be compared to determine if the sets of extracted features significantly differ groups. This data can be either unpaired (or independent when the sets of data arise from different individuals) or paired (or dependent when the sets arise from the same individuals at different time).

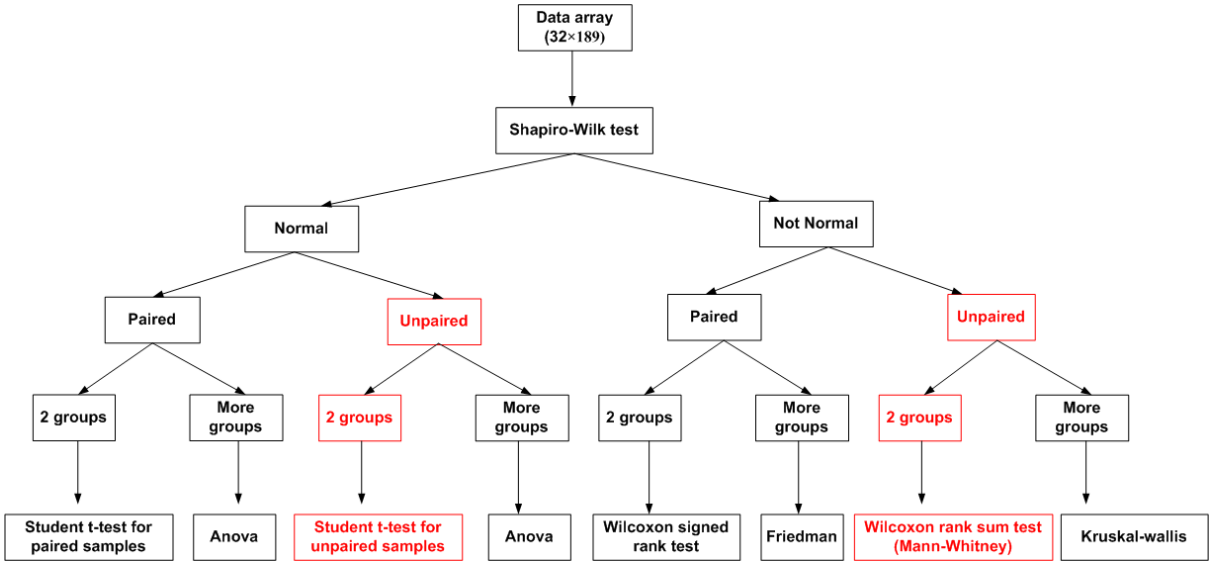



Figure 5.6. The main statistical tests for each situation.

In this work, we are working with unpaired data, where the number of groups is equal to two (PD or HC). Based on our data and Figure 5.6, for features that are normally distributed multiple independent student t-test are used for each feature separately [Koepf and Masjed-Jamei, 2006], and for features that are not normally distributed, multiple independent Mann-Whitney tests are used [Nachar, 2008]. Feature selection using statistical tests depends on a threshold parameter called alpha value or significance level. It is the probability of rejecting the null hypothesis when it is true. Features that passed the statistical test with a probability less than the significance level are kept. It is important to know how to select this parameter. For this reason, in this work feature selection based on statistical tests with sequence of alpha values between 0 and 1 were tested on each of the 7 tasks and "All-tasks" separately as shown in Figure 5.7, and the one with the best validation accuracy was picked.

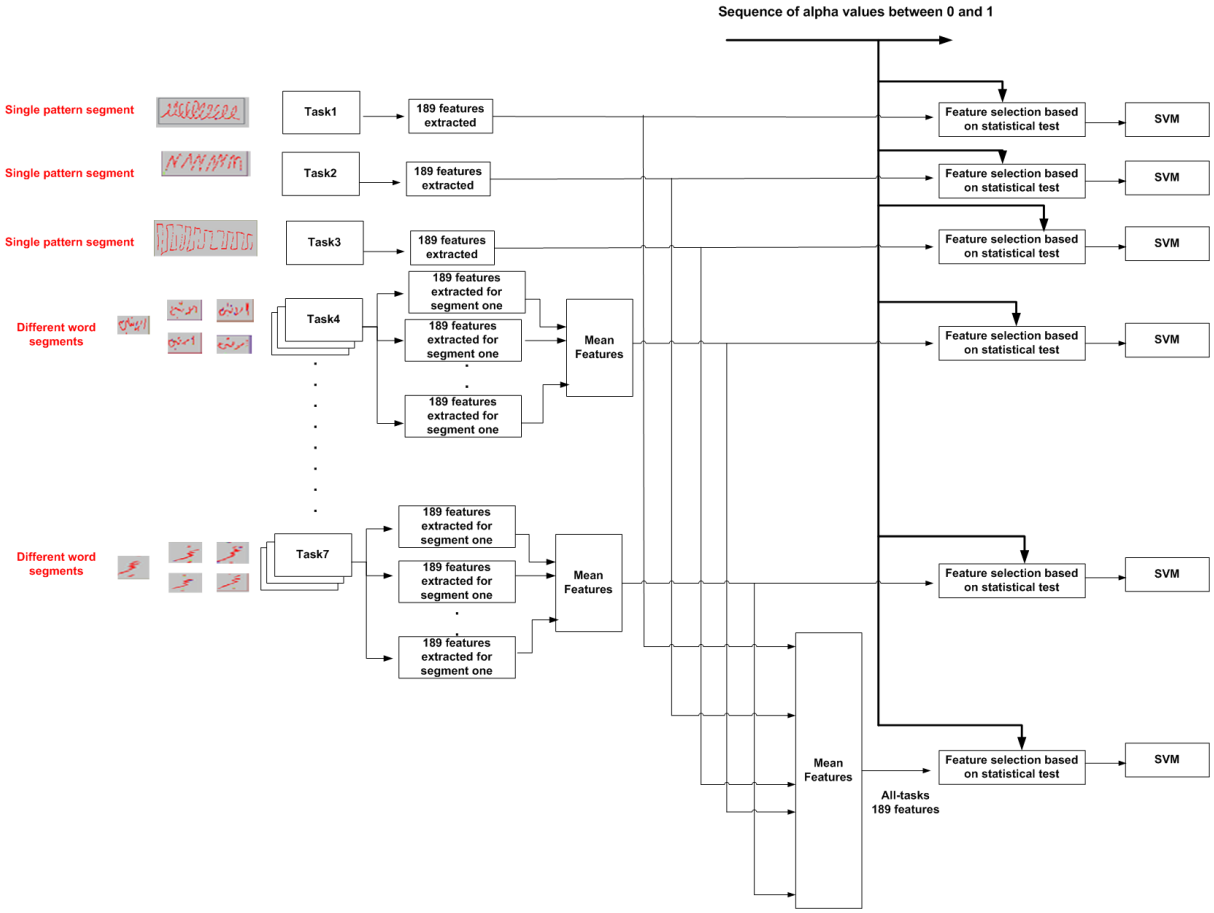



Figure 5.7. Feature selection based on statistical tests overview.

#### 5.1.2.1.1 Shapiro-Wilk test

The objective of this test is to check whether a distribution *L* is normal or not, where a correlation between the data to be tested and a given normal distribution is calculated. A correlation close to 1 would suggest a good fit to normality whereas a correlation much less than 1 would suggest non-normality [Royston, 1993].

The correlation W is defined in (5.27), where the weight vector a= $a_1, a_2 \ldots., a_n$ depends on the covariance matrix of the order statistics of a sample of n standard normal random variables V with expectation vector m (see equation (5.28)) [Royston, 1993].

$$W=\frac{(\sum_{j=1}^{n} a_j L_j)^2}{(\sum_{j=1}^{n}(L_j-\bar{L})^2)} \quad (5.27)$$

$$a = (m^T V^{-1} V^{-1} m)^{\frac{-1}{2}} m^T V^{-1} \quad (5.28)$$

The distribution of W is unknown. The probability p-value of W is defined informally as the probability of obtaining a result equal to or "more extreme" than what was actually observed, when the null hypothesis is true. The aim here is to find a transformation g such that g(W) is approximately a normal distribution with μ mean and $\sigma^2$ variance [Royston, 1993]. In this case, the probability p-value is then 1-normcdf (g(W)); where normcdf is the normal cumulative distribution function. In case of two tailed test, the p-value obtained is then multiplied by 2. Here the 1 tailed test is selected. Once p-value is obtained, a comparison with the alpha value is done. If the probability is less than the alpha value, this means that the null hypothesis (data is normal distributed) is rejected, otherwise the null hypothesis is accepted.

#### 5.1.2.1.2 Student t-test

The t-test is used for features that are normally distributed to determine if the difference between the 2 groups is significance by comparing the means of the 2 groups. The t-distribution is symmetric and bell-shaped like the normal distribution but has heavier tails [Brereton, 2015]. The pdf of the t-distribution is represented in equation (5.29), where df is the number of degree of freedom defined in equation (5.30); where $n_1$ $and$ $n_2$ represent the 2 groups sample sizes, and $\beta$ is the beta function defined in equation (5.31) [Koepf and Masjed-Jamei, 2006]. The larger the samples size, the more the distribution resembles a normal distribution (see Figure 5.8) [Brereton, 2015].

$$f(x) = \frac{1}{\sqrt{df}\,\beta(\frac{1}{2},\frac{df}{2})}\,(1+\frac{x^2}{df})^{\frac{-df+1}{2}} \quad (5.29)$$

$$df = n_2 + n_1 - 2 \quad (5.30)$$

$$\beta(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\,dt \quad (5.31)$$

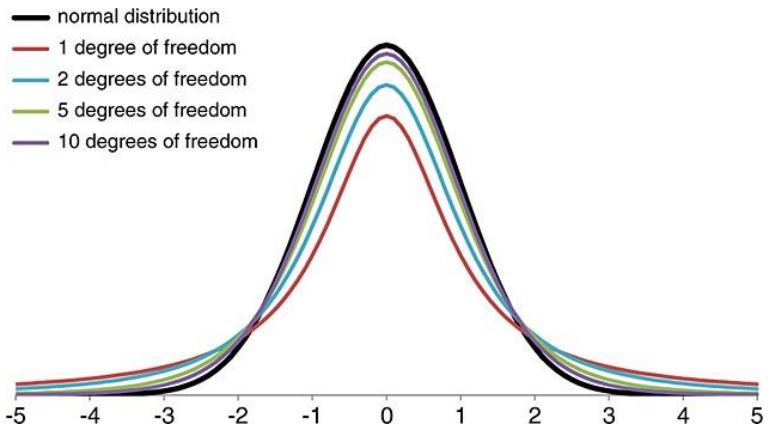



Figure 5.8. The probability density function for the t-distribution for different degrees of freedom [Brereton, 2015].

The t-test formula is described as below, where [ $\overline{X1}$ , $\overline{X2}$ ] and [*var1*, *var2*] refer to the means and variance respectively of the two samples to be compared.

$$t = \frac{|\overline{X2} - \overline{X1}|}{\sqrt{\frac{var1}{n1} + \frac{var2}{n2}}} \tag{5.32}$$

The probability value p associated with the obtained t ratio is equal to p=$(\int_t^{+\infty} f(x)dx) * tails$; where $f(x)$ is defined in (5.29), and tails can be either 1 or 2.

When using a two-tailed test, we are testing for the possibility of the relationship in both directions. In our case, we are comparing two means; the null hypothesis here is that means are equal. A two tailed test will test both if $\overline{X1} < \overline{X2}$ and if $\overline{X1} > \overline{X2}$, where the one-tailed test we test in one direction only. In this work we focused on the one-tailed test.

Once p-value is obtained, a comparison with the alpha value is done. If the probability is less than the alpha value, this means that the null hypothesis (equal means) is rejected; otherwise the null hypothesis is accepted.

#### 5.1.2.1.3 Mann-Whitney test

The Mann-Whitney test is used for features that are not normally distributed to determine if the 2 samples ($X_{11}$, . ., $X_{1n1}$ and $Y_{11}$, . . . , $Y_{1n2}$) come from the same distribution.

The Mann-Whitney U test initially implies the calculation of a U statistic for each group. The U statistics are calculated as follows [Nachar, 2008]:

- ✓ Combining data from both groups together ($X_1$ and $Y_1$)
- ✓ Arrange the combined data in ascending order
- ✓ Assigning rank that is equal to the average position of the scores in the ordered sequence
- ✓ Get the is the sum of the ranks assigned to each group separately (defined by $R_i$)

The U statistics for each group are represented in (5.33) and (5.34), where $n_1$ and $n_2$ refer to samples size.

$$U_1 = n_1\, n_2 + n_1\, (n_1 + 1)/2 - R_1 \tag{5.33}$$

$$U_2 = n_1\, n_2 + n_2\, (n_2 + 1)/2 - R_2 \tag{5.34}$$

In other words, both U equations can be understood as the number of times observations in one sample precede or follow observations in the other sample when all the scores from one group are placed in ascending order [Nachar, 2008]. If the samples sizes are large enough ($n_1\, n_2$>20 for example), the sample's distribution approaches a normal distribution with $\mu_U = \frac{n_1 n_2}{2} = \frac{U_1+U_2}{2}$ and $\sigma_U = \sqrt{n_1\, n_2(n_1 + n_2 + 1)/12}$. The corresponding equation becomes:

$$Z_1 = |U_1 - \mu_U|/\, \sigma_U \text{ and } Z_2 = |U_2 - \mu_U|/\, \sigma_U \tag{5.35}$$

$Z_1$ and $Z_2$ are equal since $\mu_U = \frac{U_1+U_2}{2}$, the p-value is then equal to 1-normcdf ($Z_{1\,or\,2}$) [Nachar, 2008]. From the other side, if the samples sizes are small then the exact Mann-Whitney distribution is used, where the p-value is defined in (5.36), where $c(u|\, N,n_1)$ is the number of possible arrangements of $n_1$ $X_1$s and $n_2$ $Y_1$s that give a value of $R$ that does not exceed $r$ ($r$=min($R_1, R_2$)), N=$n_1 + n_2$, and n=min($n_1, n_2$) [Cheung and Klotz, 1997].

$$\text{P}= \frac{n1!n2!}{(n1+n2)!}\, c(r|N,n) \tag{5.36}$$

In case of two tailed test, the p-value obtained from (5.36) is multiplied by 2. Here we choose 1 tailed test. If the probability is less than the alpha value, this means that the null hypothesis (the 2 groups come from the same distribution) is rejected; otherwise the null hypothesis is accepted.

#### 5.1.2.2 Feature selection based on suboptimal approach

The set of selected features by statistical tests for "All-tasks" remains large. It is further reduced to a smaller subset of features in the second stage. A suboptimal approach that provides a kind of benchmark of the relevance of the features in "All-tasks" has been used for this purpose. Each feature resulting from the statistical tests is used alone to classify a cross-validation set. The feature providing the highest validation performance is first selected. To this one, features are added incrementally to the final selected features set by selecting, at every iteration, the one yielding the highest validation performance. The iterations stop when no more increase in performance is observed. It is worth noting, that for selection as well as for detection, an SVM classifier is being used.

### 5.1.3 Support vector machine

SVM is a useful technique for data classification [Vapnik, 2010]. A classification task usually involves separating data into training and testing sets. Each instance in the training set contains one "target value" (i.e. the class labels) and several "attributes" (i.e. the features or observed variables). The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes.

The idea of SVM is simple: The algorithm determines a linear separator, a hyperplane, which separates the classes in case e.g. in a classification problem. To separate the two classes of data points, there are many possible hyperplanes that could be chosen. An optimal linear classifier maximizes the margin (or distance) between the points on either side of the so called decision line. The benefit of this process is that after the separation, the model can easily guess the target classes (labels) for new cases [Berwick and Idiot, 2016].

The simplest form is where classes are assumed to be linearly separated as shown in Figure 5.9.

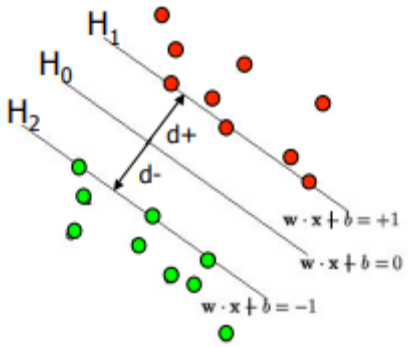



Figure 5.9. Maximum-margin hyperplane and margins for an SVM trained with samples from two classes [Berwick and Idiot, 2016].

Define the hyperplanes H such that:

$$\begin{cases} w \bullet xi + b \geq +1 \, when \, yi = +1 \\ w \bullet xi + b \leq -1 \, when \, yi = -1 \end{cases} \tag{5.37}$$

where w represents the weight vector, and b is the intercept.

The points on the hyperplane $H_1$ and $H_2$ are known as support vectors. We are looking for a classifier (linear separator) with as big a margin as possible. In a 2 dimensional space, the distance from a point $(x_0,y_0)$ to a line: Ax+By+c = 0 is: $|Ax_0 + By_0 + c|/\text{sqrt}(A^2+B^2)$. The distance between $H_1$ and $H_2$ is then: $2\times|w\bullet x+b|/||w||=2/||w||$. Maximizing the margin is obtained by minimizing ||w|| subject to the constraints in equation (5.37) (to make sure that there are no data points between $H_1$ and $H_2$) [Berwick and Idiot, 2016].

Generalizing this concept to account for nonlinear boundaries, SVM aim to solve the following optimization problem:

$$\min \frac{1}{2} w^T w \tag{5.38}$$

$$\text{subject to: } y_i \, (w^T \varphi(x_i) + b) \geq 1$$

$\varphi$ is a function which maps (projects) the samples from the k-dimensional feature space to a larger (potentially infinite) dimensional space, where the samples are linearly separable as shown in Figure 5.10. The solution of the classification problem in a higher dimensional

space involves only a dot product between the vectors resulting from the projection. Furthermore, $K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$ is called the kernel function. The use of kernel functions permits to avoid the definition and application of the mapping functions and to apply the direction dot product on the measured vectors in the original space. Some common kernel functions are provided in Table 5.7 where r, $\gamma$, and d are kernel parameters [Misra, 2019].

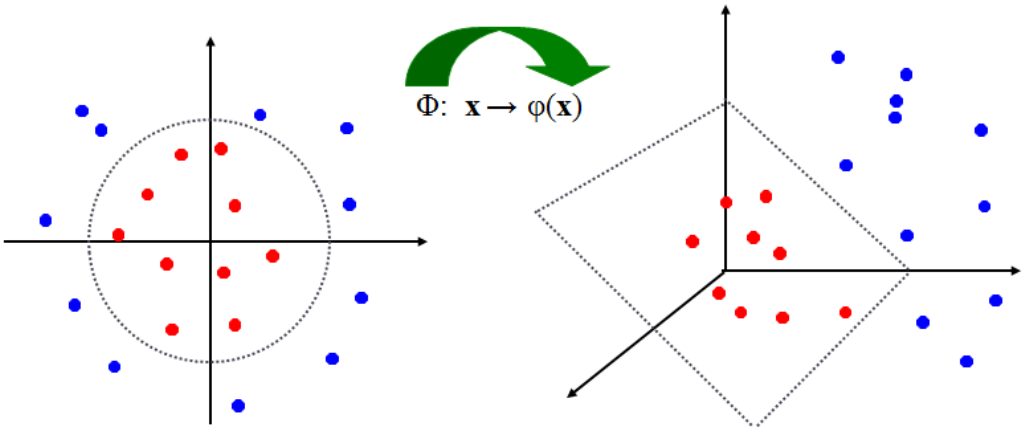



Figure 5.10. Mapping the original input space to some higher dimensional space where the samples are linearly separable using kernel function [Zafari et al., 2019].

Table 5.7. Common SVM kernels.

| **Kernel** | **$K(x_i, x_j)$** |
|---|---|
| Linear | $x_i^T x_j$ |
| Polynomial | $(\gamma x_i^T x_j + r)^d$ |
| Radial basis function (RBF) | $\exp(-\gamma \|\|x_i - x_j\|\|^2)$ |
| Sigmoid | $tanh(\gamma x_i^T x_j + r)$ |

The optimization problem defined in equation (5.38) cannot be used in many real-world problems when features derived from the data are noisy and can not be linearly separated (see Figure 5.11-a). In order to handle situations like these, soft margin formulation was proposed by some researchers [Misra, 2019]. The idea behind the soft margin classification is to allow SVM to make a certain number of mistakes and keep margin as wide as possible so that other points can still be classified correctly. This can be done simply by modifying the objective of SVM as follows:

$$\min \frac{1}{2} w^T w + C$$

where C is a hyper-parameter that decides the trade-off between maximizing the margin and minimizing the mistakes. When C is small, classification mistakes are given less importance and focus is more on maximizing the margin, whereas when C is large, the focus is more on avoiding misclassification at the expense of keeping the margin small [Misra, 2019].

It should be noted that not all mistakes are equal. Data points that are far away on the wrong side of the decision boundary should incur higher penalty as compared to the ones that are closer. The idea is: for every data point$x_i$, a slack variable $\xi_i$ is introduced. The value of $\xi_i$ is the distance of $x_i$ from the corresponding class's margin if $x_i$is on the wrong side of the margin, otherwise zero. Thus the points that are far away from the margin on the wrong side would get more penalty (see Figure 5.11-b) [Misra, 2019].

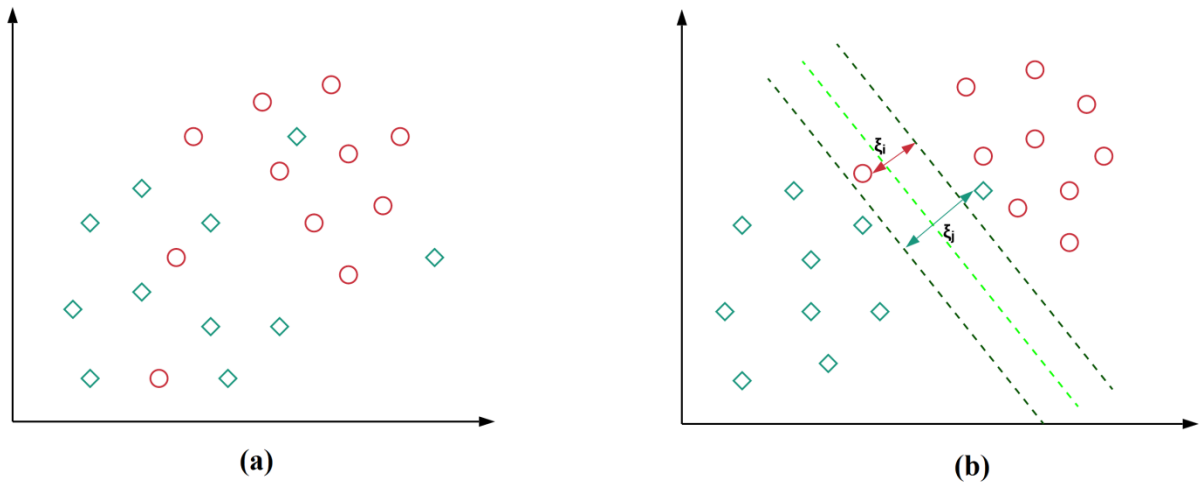



Figure 5.11. (a) Data representation where the two classes are not linearly separable, (b) The penalty incurred by data points for being on the wrong side of the decision boundary [Misra, 2019].

With this idea, the new objective is to minimize the following function, where each data point $x_i$ needs to satisfy the following constraint:

$$\min \frac{1}{2} \mathrm{w}^{\mathrm{T}}\mathrm{w} + C\Sigma\xi_i \tag{5.39}$$

$$\text{subject to } \begin{cases} y_i\,(w^T\varphi(x_i) + b) \geq 1 - \xi i \\ \xi i \geq 0 \; for \; all \; i \end{cases}$$

After selecting the classifier, it is important to know how to select the kernel. If the number of features is large, one may not need to map data to a higher dimensional space. In this case, the nonlinear mapping does not improve the performance and using the linear kernel is good enough. Since the number of selected features in our case study is small, a non-linear kernel was chosen [Hsu et al., 2003]. The RBF kernel was selected for 3 reasons: the first reason is that other non-linear kernels have more hyper-parameters than RBF as shown in Table 5.7. The second reason is that the kernel values of the polynomial kernel may go to infinity or to zero while the degree d is large. Finally, the sigmoid kernel is not valid under some parameters [Hsu et al., 2003]. There are two parameters for RBF kernel: C and $\gamma$. In order to identify the best parameters, a grid search using cross-validation was applied on C and $\gamma$.

Various pairs of (C, $\gamma$) values are tried and the one with the best cross-validation accuracy is picked. Exponentially growing sequences of C and ɤ were used (C =$2^{-30}$, $2^{-29}$, ……., $2^{29}$,$2^{30}$ and $\gamma = 2^{-30}$, $2^{-29}$, ……., $2^{29}$,$2^{30}$) [Hsu et al., 2003]. Several advanced methods for kernel parameters selection exist. We preferred the simple grid search method because it is more reliable to do exhaustive parameter search instead of using approximations. The only disadvantage of the grid-search is that it may be time consuming, but here the time required to select the best parameters will no defer from the time required by the advanced methods since we are only selecting 2 parameters and the data we are working on is small [Hsu et al., 2003]. The set with 32 subjects being limited; it has been divided into 4 folds, with the 75/25 % training/validation subsets proportion. Each patient is in the validation set exactly once. Sequentially, one fold was validated using the classifier trained on the remaining 3 folds. The total accuracy is obtained by calculating the mean of all the folds accuracies. In this work, we decided not to use a separate test set due to a low database size. As a result the validation set can be considered as test set.

Data is split into folds by using the stratified sampling technique in order to ensure the same data class distribution in the folds. The reason for selecting such sampling method is that since all the folds contain the same class distribution and the same number of classes, the algorithm will not tend to learn or test a class that is not represented at all in a fold. This will ensure that the model not only fits the training data well but also generalizes on

test/validation data. The scenario of k folds stratified cross validation with M classes is represented in Figure 5.12.

Before applying the SVM, features are scaled to the range [-1, 1] in order to avoid the dominance of features with greater numeric ranges [Hsu et al., 2003]. Equation (5.40) is applied on each feature separately for scaling, where upper and lower refer to the upper and the lower values of the scaling range (here the upper value is 1 and the lower value is -1), and minimum and maximum refer to the minimum and maximum feature values in the sample. The scaling factors (minimum and maximum) are obtained from training data and used to scale the test data.

$$Scaled\_data = lower + \frac{(\text{upper} - \text{lower}) \times (\text{data} - \text{minimum})}{\text{maximum} - \text{minimum}} \qquad (5.40)$$

## 5.1.4 Experiments and numerical results

The prediction performance is evaluated in term of the accuracy, sensitivity, specificity, and area under curve (AUC) of receiving operator characteristic (ROC). The accuracy, sensitivity and specificity are defined as follows:

$$\text{Accuracy}=\frac{TP+TN}{TP+TN+FP+FN}\times 100$$

$$\text{Sensitivity}=\frac{TP}{TP+FN}\times 100$$

$$\text{Specificity}=\frac{TN}{TN+FP}\times 100$$

where true positive (TP) represents the number of correctly classified PD subjects, true negative (TN) represents the total number of correctly classified HC subjects, false positive (FP) represents the number of actually healthy subjects diagnosed as PD, and false negative (FN) represents the total number of PD patients incorrectly classified as HC. Sensitivity measures the ability of the system in correctly classifying PD subjects, where specificity measures the ability of the system in correctly classifying Healthy subjects.

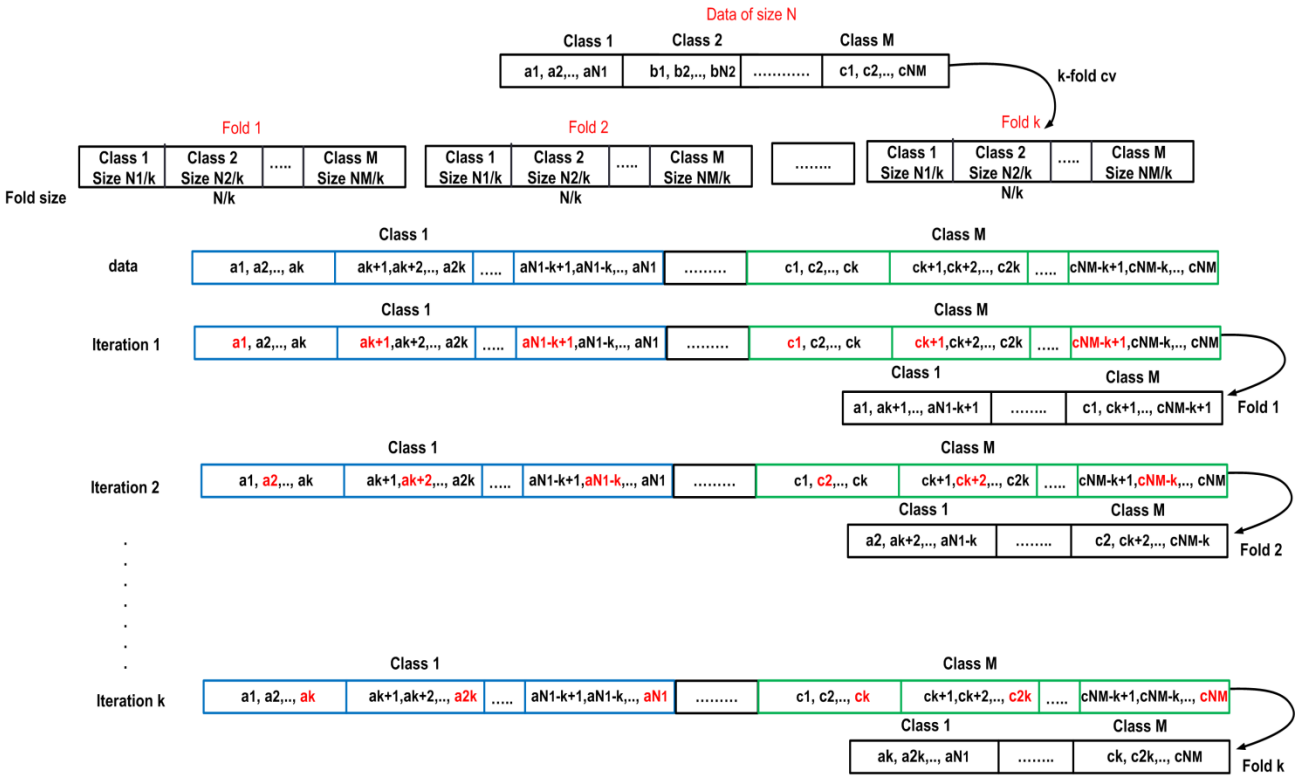



Figure 5.12. k folds stratified cross validation with M classes.

The ROC curve is a two-dimensional measure of classification performance. It can be understood as a plot of the probability of classifying correctly the positive examples against the rate of incorrectly classifying true negative examples at different classification threshold values [Rakotomamonjy, 2004]. The true classes and predicted probabilities obtained by the algorithm are needed to plot the ROC curve. The predicted probability is compared to a given threshold. If it is greater than or equal to the threshold, the sample is labeled as positive class and if it is smaller than that, it is labeled as negative class. The AUC is the most commonly used ROC index, where a value near 1 means that the model has a good measure of separability (a good model), and a value near 0 means it is a poor model [Rakotomamonjy, 2004].

In the first stage, prediction performance is considered for the seven handwriting tasks and the “All-tasks” separately where only statistical tests are applied for feature selection. Statistical tests reduce the total number of features (for the seven tasks and the “All-tasks”) from 1512 to 376 features. The distribution of selected features for different tasks is shown in Figure 5.13. The significance level with the best validation accuracy, and the numerical

results achieved by the SVM classifier with 4 folds cross-validation using only statistical tests for feature selection are presented in Table 5.8.

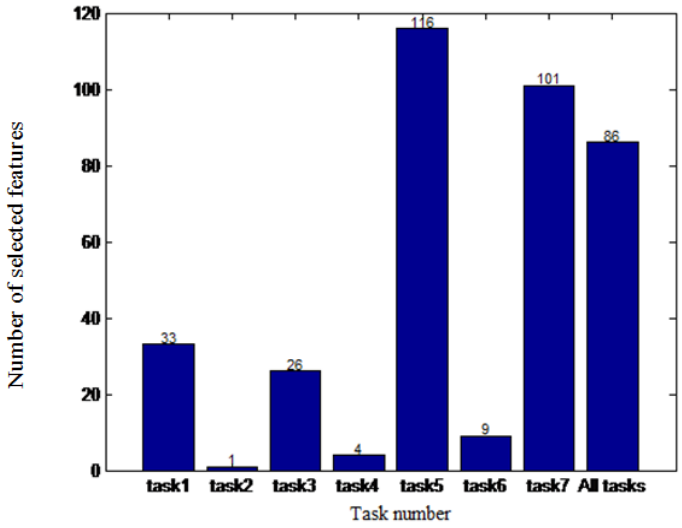



Figure 5.13. The number of selected features per task using statistical tests.

Table 5.8. Performance of each task in PD classification using statistical tests for feature selection.

| Task # | Accuracy (%) | Sensitivity (%) | Specificity (%) | AUC | Significance Level |
|---|---|---|---|---|---|
| 1 | 87.5 | 87.5 | 87.5 | 0.48 | 0.0111 |
| 2 | 93.75 | 93.75 | 93.75 | 0.68 | 0.0015 |
| 3 | 90.63 | 93.75 | 87.5 | 0.59 | 0.0218 |
| 4 | 87.5 | 87.5 | 87.5 | 0.30 | 0.0008 |
| 5 | 87.5 | 75 | 100 | 0.39 | 0.1306 |
| 6 | 84.38 | 75 | 93.75 | 0.51 | 0.0083 |
| 7 | 84.38 | 81.25 | 87.5 | 0.47 | 0.103 |
| All | 90.63 | 81.25 | 100 | 0.47 | 0.126 |

According to Table 5.8, features that are selected for “All-tasks” by the statistical test returned 90.63 % accuracy with 86 features selected. The number of selected features is still large; the suboptimal incremental approach described in section 5.1.2.2 is applied. Every feature of the 86 selected is used separately as an input to the SVM classifier in order to evaluate its classification accuracy. The classification accuracy for each feature selected in the first stage is shown in Figure 5.14. The highest classification accuracy of a single feature is 87.50 %. One possible approach to determine the most relevant features consists in a progressive increment in the size of the feature vector by adding, at each iteration, the feature that would maximally increase the classification performance. Validation set is used here for selecting the most relevant features. This is a suboptimal approach that provides a kind of

benchmark of the relevance of the features in the desired task. Starting from the feature with the highest validation accuracy, the validation accuracy is computed as the number of features is increased. The highest validation accuracy obtained at every iteration of the suboptimal approach is shown in Figure 5.15. In order to get these results, first of all feature with the highest accuracy (total accuracy from 4-folds cross validation) separately was selected. Then features are added one by one incrementally till we get n=86 features. The highest classification accuracy was 96.87 % for N=12 features. Since the validation set is used here for feature selection as for model performance estimation, this may result in an overoptimistic bias in our estimate of the model's generalization performance that needs to be checked on a different data set.

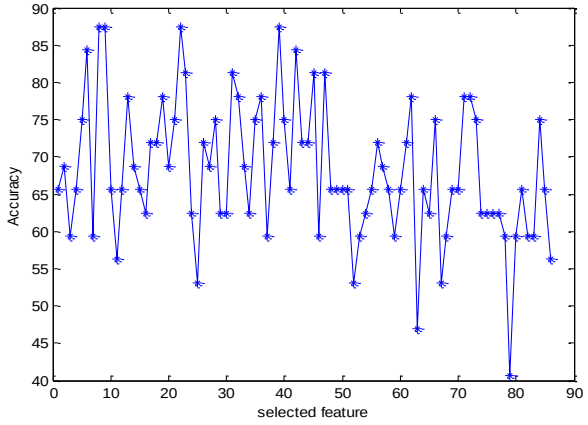



Figure 5.14. The classification accuracy for each feature selected in the first stage for "All-tasks", where it is used separately as input to SVM classifier.

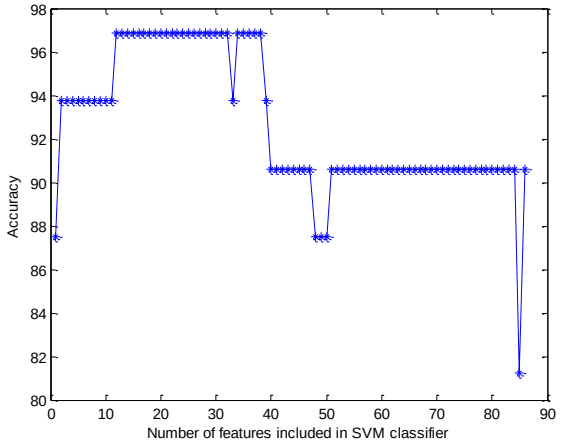



Figure 5.15. The highest classification accuracy obtained at every iteration of the suboptimal approach.

The "All-tasks" performance obtained with one and two stage feature selection methods are shown in Table 5.9. It is clear that a smaller subset of features gave better classification accuracy. This indicates clearly that some of the features disturb the performance of the prediction system, mainly because of the curse of dimensionality, a critical factor in this particular case where a limited amount of training data is available. The 12 selected features providing the best "All-tasks" performance are represented in Table 5.10.

Table 5.9. Table of comparison between the "All-tasks" performance obtained with one and two stage feature selection methods.

| **Performance** | **1 stage Feature selection** | **2 stages Feature selection** |
|---|---|---|
| Accuracy (%) | 90.62 | 96.87 |
| Sensitivity (%) | 81.25 | 93.75 |
| Specificity (%) | 100 | 100 |
| AUC | 0.4727 | 0.5938 |

Table 5.10. The two stages selected features providing the best "All-tasks" performance.

| **Feature** | **Statistic** |
|---|---|
| Rising edge: number of changes in velocity pressure (NCP) | Median |
| Rising edge: correlation between pressure and vertical velocity | STD |
| Main part: NCP | STD |
| Main part: NCP | 1$^{st}$ percentile |
| Falling edge: correlation between pressure and horizontal velocity | 1$^{st}$ percentile |
| Falling edge: correlation between pressure and horizontal acceleration | 1$^{st}$ percentile |
| Falling edge: correlation between pressure and vertical acceleration | 99$^{th}$ percentile |
| Horizontal velocity | 1$^{st}$ percentile |
| Stroke Time | 1$^{st}$ percentile |
| NCV/stroke | Mean |
| NCV/stroke | Median |
| NCV/stroke | 1$^{st}$ percentile |

The selected features providing the best performance are a combination of pressure, kinematic, and correlation between the pressure and kinematic features. The features with the highest relevance come mainly from two tasks: task 2 and task 3. These two tasks are considered long and somehow complex. Copying these cursive tasks needs higher cognitive force and explains the effect of disease on handwriting.

A comparison of the performance of the developed method in this study and the performance of the previous works is presented in Table 5.11. Drotar et al [Drotar et al., 2016], [Drotar et al., 2015a], and [Drotar et al., 2015b] extracted features from handwriting samples taken from the PaHaW database, distinct from the one used in this study, containing samples from 37 PD patients (on-state) and 38 HC subjects. According to the results

presented in Table 5.11, our developed SVM model reports highest performance across all the mentioned works although results are not always measured on the same database. It is difficult to compare results obtained by different works using different datasets. Future tests on PaHaw database would offer an additional validation of the results and conclusions.

Table 5.11. Comparison table between the developed method and the previous studies.

| **Reference** | **Database** | **Significance level** | **Feature selection method** | **Classifier** | **Accuracy (%)** | **Sensitivity (%)** | **Specificity (%)** |
|---|---|---|---|---|---|---|---|
| Developed method [Taleb et al., 2017] | HandPDMultiMC | variable between 0&1 | described in section 5.1.2 | SVM | 96.87 | 93.75 | 100 |
| [Drotar et al., 2016] | PaHaW | 0.05 | Mann-Whitney U-test | SVM | 81.3 | 87.4 | 80.9 |
| [Drotar et al., 2015b] | PaHaW | 0.05 | Mann-Whitney U-test | SVM | 88.13 | 89.47 | 91.89 |
| [Drotar et al., 2015a] | PaHaW | 0.05 | Mann-Whitney U-test | SVM | 89.09 | N/A | N/A |

## 5.1.5 Conclusions

The main goal of this part of the thesis is to build a language-independent SVM model for assessment the motor disorders in PD patients based on handwriting features. Seven different handwriting tasks taken from HandPDMuliMC subset are used for this purpose, where this subset includes samples in three languages: Arabic, French and English. In order to assess different motor symptoms, advanced handwriting mark0ers based on kinematic, stroke, pressure, entropy, and intrinsic features are extracted from the “on-paper” periods only. To avoid the risk of falling in a curse of dimensionality, a two-stage feature selection approach is applied to remove the irrelevant features while keeping features that are necessary and sufficient to describe the target concept. Due to the small data we worked on, a 4-fold cross validation SVM classifier with RBF model was used for binary classification. We have succeed to build a language-independent model for PD diagnosis using handwriting analysis with 96.87 % accuracy, 93.75 % sensitivity, and 100 % specificity when a combination of kinematic, pressure, and correlation between kinematic and pressure features is used. From a clinical point of view, the acceleration and stroke size are regulated by the motor control mechanism of the wrist and finger movement, a mechanism that is inaccurate or absent in PD. Moreover, further detailed information that can not be obtained from the kinematic features might be offered by the pressure features, hence, the significance to show the relationship between kinematic and pressure features [Drotar et al., 2016].

The proposed diagnosis model is considered relevant even though the size of the database is small, because a 4-fold cross validation SVM classifier was used to validate the stability of the model through generating different combinations of the data, where stratified sampling technique is applied to ensure the same class distribution in all the folds. This improves the reliability and guarantees the effectiveness of the results.

Our perspective is to build a model which not only fits the training data well but also generalizes well on test/validation data. The best way to achieve this is by training our model on larger training set. Our database is perspective in size and can be easily expanded for this purpose. Our next work consisted of collecting more samples to enlarge our database, and building a model to predict PD stage and progression using the combination of features selected in this part.

## 5.2 Predicting Parkinson's disease progression using engineered features and support vector machines

Monitoring the progression of the disease over time requires repeated clinical visits by the patient. Hence, there is a need to define a reliable system that assists in the decision-making process for the diagnosis of PD and the prediction of the modified H&Y stage (included stages 0 through 5 with the addition of stages 1.5 and 2.5 to account the intermediate course of Parkinson disease). Details about modified H&Y stages can be found in chapter 2. The goal here is to build a reliable system that predicts PD stage and progression based on the selected features found and described in section 5.1 [Taleb et al., 2018].

The H&Y stage is a standard clinical measure of the progression of PD. It show the presence and severity of PD symptoms and can be used for assessing Parkinsonism and for quantifying the degree of impairment caused by Parkinsonian symptoms.

The total number of subjects studied here is 32; where 16 are PD (in their “on-state”) and 16 HC; where the H&Y stage of HC subjects is equal to 0. H&Y stage of each of the 32 subjects are shown in Table 5.12, where samples distribution between classes for each of the H&Y stages is shown in Figure 5.16. Based on Figure 5.16, we can see that we are working

with a multiclass classification where the database is imbalanced and the number of samples within each class is very limited.

Table 5.12. H&Y stages of the 32 Subjects.

| Subject | H&Y stage on-state |
|---|---|
| PD1 | 1 |
| PD2 | 3 |
| PD3 | 2 |
| PD4 | 2 |
| PD5 | 1 |
| PD6 | 2 |
| PD7 | 1 |
| PD8 | 2 |
| PD9 | 2 |
| PD10 | 1.5 |
| PD11 | 2 |
| PD12 | 1.5 |
| PD13 | 1 |
| PD14 | 3 |
| PD15 | 4 |
| PD16 | 1.5 |
| C1-16 | 0 |
| Nb. Of classes | 6 |

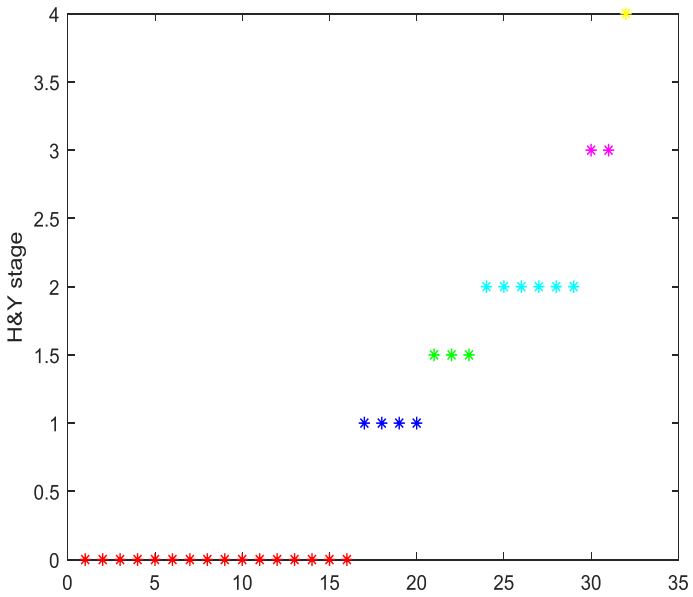



Figure 5.16. Samples distribution between classes for each of the H&Y stage.

The completed task sheets for PD patients under medication in different stages of the disease, and HC subject are shown in Figure 5.17. It is obvious how the height and width of letters decrease with the stage of the disease. From the other side, we can see that tremor oscillations have larger amplitude in higher stages of the disease.

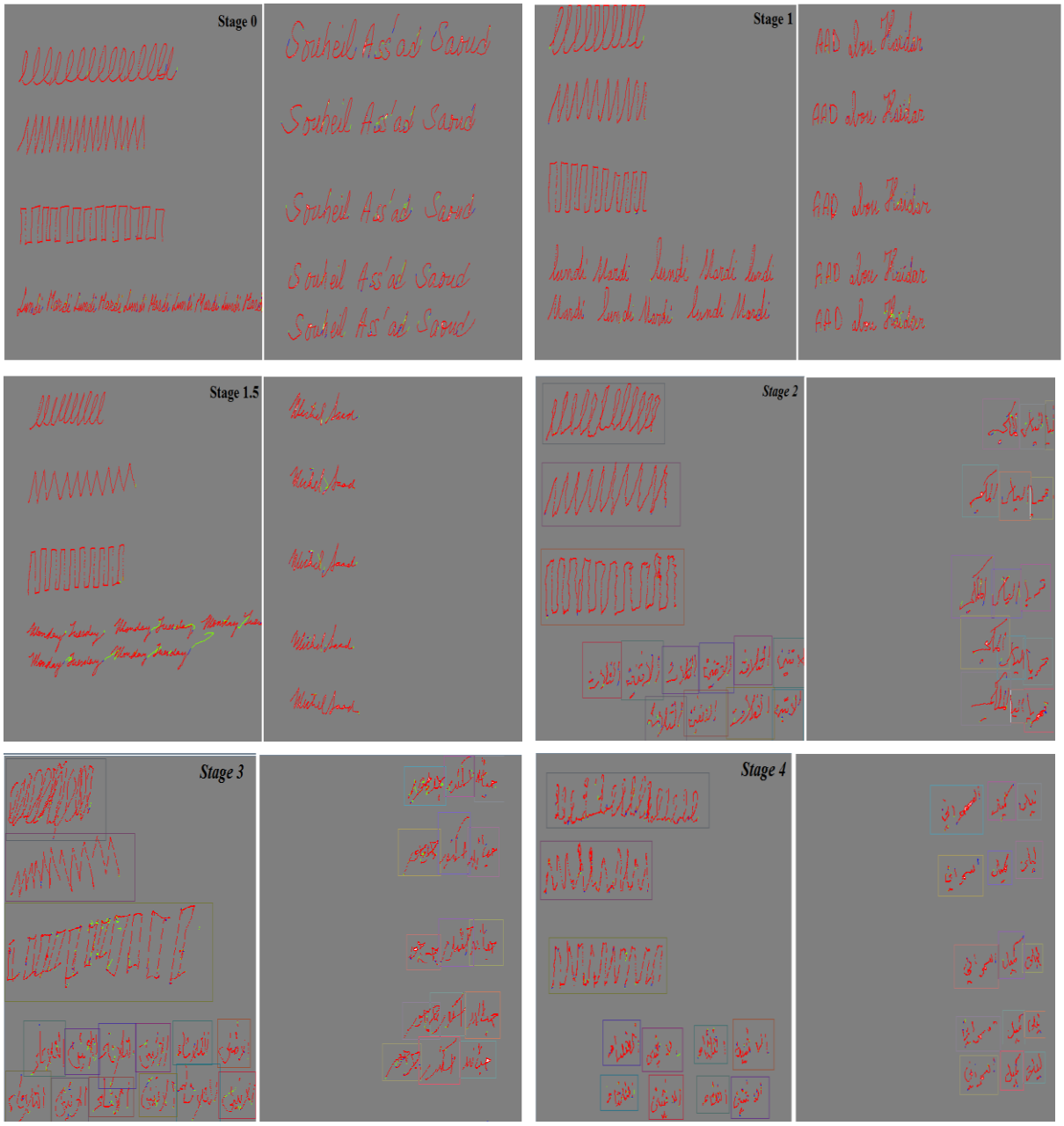



Figure 5.17. Completed task sheets for PD patients under medication in different stages of the disease and a HC subject.

## 5.2.1 Segmentation and feature extraction

In this part we are dealing with multi-class classification, where each class contains certain samples due to the small dataset that we are working with. The use of one global feature vector can be penalizing for this purpose since some details are smoothed in the extracted features. Therefore, it has been decided to extract several vectors from each task. For each handwriting task, each word/pattern is divided into overlapping segments of length

L/5 with a shift of L/50 where L is the whole word/pattern length. The total number of segments per word/pattern is calculated as follows:

$$Number\ of\ segments = \frac{total\ length - overlap}{segment\ size - overlap} = \frac{L - \frac{9L}{50}}{\frac{L}{5} - \frac{9L}{50}} = 41$$

For each word/pattern within a task, we extract the set of 12 features selected in section 5.1 for each of the 41 segments, and then we calculate the average across the different words/patterns. Once the feature vectors are extracted for each of the 41 segment and each task, the mean feature vector per segment across the different tasks is calculated and applied to build and test a multi-class SVM classifier. A system overview is shown in Figure 5.18.

## 5.2.2 Multi-class SVM classifier

SVM originally separates the binary classes (k=2) with a maximized margin criterion. However, real-world problems often require the discrimination for more than two categories. The multi-class classification problems (k>2) are commonly decomposed into a series of binary problems such that the standard SVM can be directly applied [Wang et al., 2014]. Two representative ensemble schemes exist: “one *versus* rest” (OVR) and “one *versus* one” (OVO).

### 5.2.2.1 One-versus-rest approach

The OVR approach constructs k separate binary classifiers for k-class classification. The $m^{th}$ binary classifier is trained using the data from the $m^{th}$ class as positive examples and the remaining k-1 classes as negative examples. During testing, the class label is determined by the binary classifier that generates maximum output value [Wang et al., 2014].

### 5.2.2.2 One-versus-one approach

The OVO approach considers each binary pair of classes and trains classifier on subset of data containing those classes. So it trains total n×(n-1)/2 classes. During the classification phases each classifier predicts one class, and the class which has been predicted most is the answer [Wang et al., 2014].

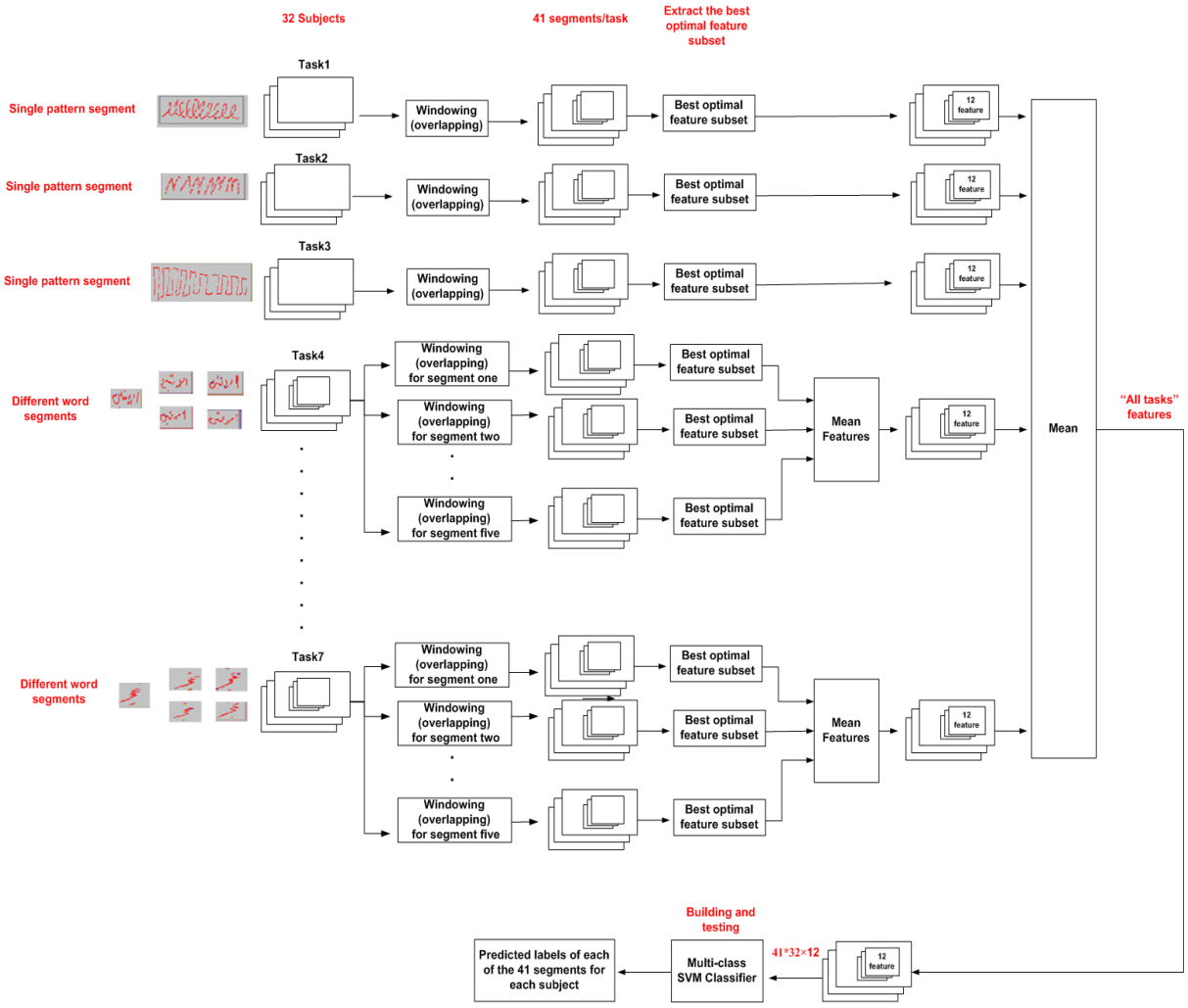



Figure 5.18. Windowed segmentation applied on each handwriting task.

The OVR approach was used in this work since it is less computationally expensive than the one-versus-one. A 4 fold cross-validation multi-class SVM classifier with RBF kernel is applied, of the 4 subsamples one subsample is retained as the validation data, one other subsample is retained as the test data, and 2 subsamples are used as training data. Cross validation is done by participant; this means that there are no windows in training, validation and testing referring to the same participant. The kernel parameters are selected via grid search method, stratified sampling technique is used, and features are scaled to the range [-1, 1] before applying the SVM (see section 5.1.3).

## 5.2.3 Score level fusion

The Score level fusion is applied to combine the scores of all segments from one subject in order to acquire the final scores for classification decision. In each fold of cross validation (CV) and for each subject in the test set, the multi-class SVM classifier computes the scores for each segment of the 41. The 41 scores shall be combined to produce a final score that will be the basis of the decision.

The scores of each classifier for each segment *j* of subject *i* are represented by the following vector:

$$S_{ij}=[S_{ij1},S_{ij2},\ldots,S_{ijk},\ldots,S_{ijC}] \tag{5.41}$$

where *C* is the number of classes in the training data, and $S_{ijk}$ denotes the score for the $k^{th}$ class for the $j^{th}$ segment of subject *i*.

The differences in classification performance among the segments of each class for each subject should be considered. Therefore, a score-level fusion based on learned weights of each segment for each class of each subject is used. The weights terms of each segment for each class are learned by a three layers Multilayer Perceptron (MLP) that is trained with backpropagation algorithm; where cross-entropy cost function and stochastic gradient descent (SGD) optimizer [Kim, 2017] are applied here.

The final score vector for each subject *i*, which is defined as follows:

$$O_i=[O_{i1}, O_{i2},\ldots\ldots, O_{ik}, \ldots, O_{iC}] \tag{5.42}$$

The decision function of the multiclass classification problem is defined as follows:

$$l_i=\text{argmax}_{k=1,\ldots\ldots,C}\, O_{ik} \tag{5.43}$$

where $l_i$ is the classification result (class label) of the $i^{th}$ subject, and $O_{ik}$ is the $k^{th}$ value of the final score vector of the $i^{th}$ subject.

To do this, in each fold and for each subject, the score values provided by the SVM classifier for each of the N segments (N is a hyper-parameter that vary between 1 and 41) were used as a neural network MLP input features (N×C -dimensional feature) where C represents the number of classes corresponding to the H&Y stages. The hyper-parameter N is selected using the validation set.

For each subject, M observations were included; where M is the number of possible combinations of N adjacent segments as shown in Figure 5.19. In this case, M is defined as follows:

$$M=\frac{totallength-overlap}{segmentsize-overlap}=\frac{41-(N-1)}{N-(N-1)}=41-N+1$$

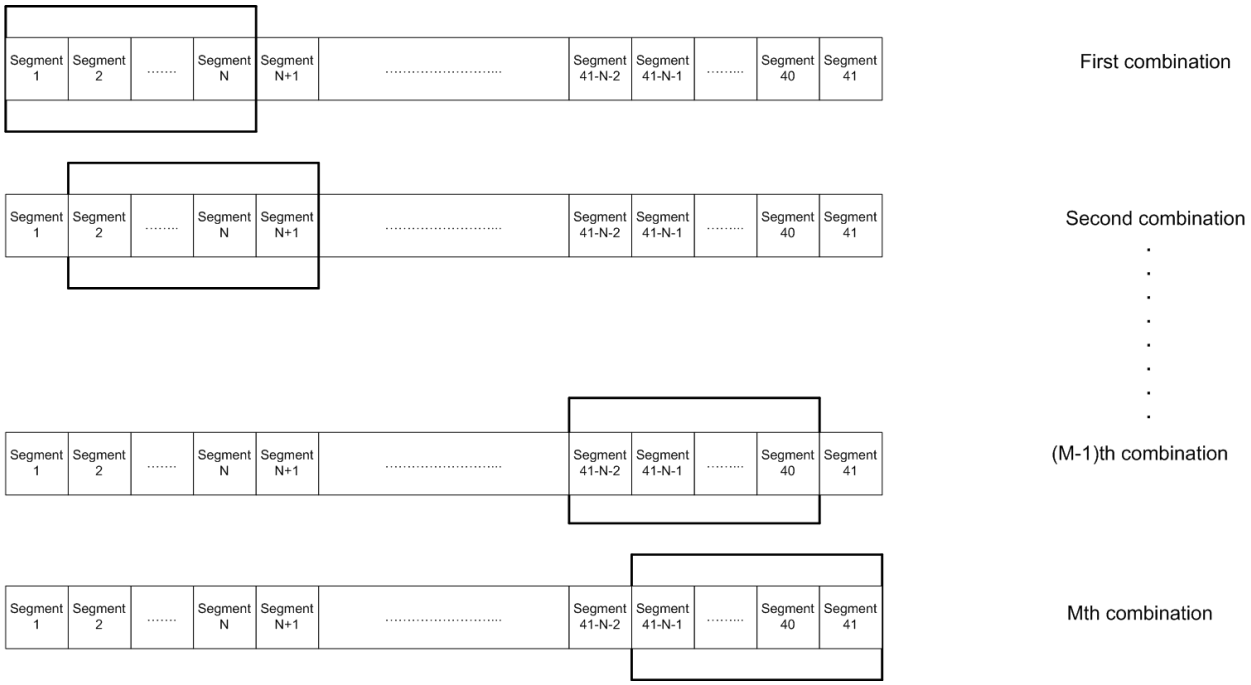



Figure 5.19. The M possible combinations of N adjacent segments.

### 5.2.3.1 Multilayer perceptron

MLP networks are a class of feedforward artificial neural network (ANN). They consist of perceptrons that are organized in layers: an input layer, one or more hidden layers, and the output layer [Du and Swamy, 2013]. Every perceptron in one particular layer is usually connected to every perceptron in the layer above and below. These connections carry

weights $w_i$. Each perceptron calculates the sum of the weighted inputs, and feeds it into its activation function. The result is then passed on to the perceptrons of the next layer. The output layer has the same number of perceptrons as there are classes, and the perceptron with the highest activation provides the classification result of the input sample. Training is achieved by successively feeding all training samples into the network, and comparing the output with the true class label. Backpropagation learning is the most popular learning rule for performing supervised learning tasks. It is not only used to train feedforward networks such as MLP, but also is adapted to "recursive neural networks" (RNNs). It uses a gradient-search technique to minimize a cost function between the desired and actual network outputs. Due to the backpropagation algorithm, MLP can be extended to many layers. The backpropagation algorithm propagates backward the error between the desired result and the network output through the network. After providing an input pattern, the output of the network is then compared with a given target pattern and the error of each output unit calculated [Du and Swamy, 2013]. This error signal is propagated backward, and a closed-loop control system is thus established. The weights can be adjusted by a gradient-descent based algorithm. The amount of that the weights are updated during training is referred to the learning rate. The learning rate is a configurable hyper-parameter used in the training of neural networks that has a small positive value, often in the range between 0 and 1.

Activation functions would introduce non-linear properties to the ANN. An ANN without activation function would not be able to learn and model non-linear input-output mapping functions (like in image, speech, and video processing etc.). The most popular activation functions are as follows:

- **Binary step function**: it is a threshold-based activation function. If the input value is above or below a certain threshold, the neuron is activated and sends exactly the same signal to the next layer (see Figure 5.20) [Gupta and Dishashree, 2020].

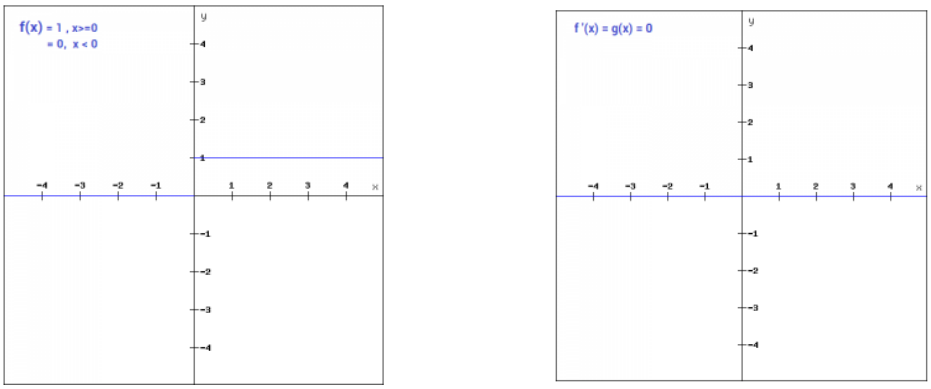



Figure 5.20. Binary step function and its derivative [Gupta and Dishashree, 2020].

- **Linear function:** takes the inputs, multiplied by the weights for each neuron, and creates an output signal proportional to the input (see Figure 5.21) [Gupta and Dishashree, 2020].

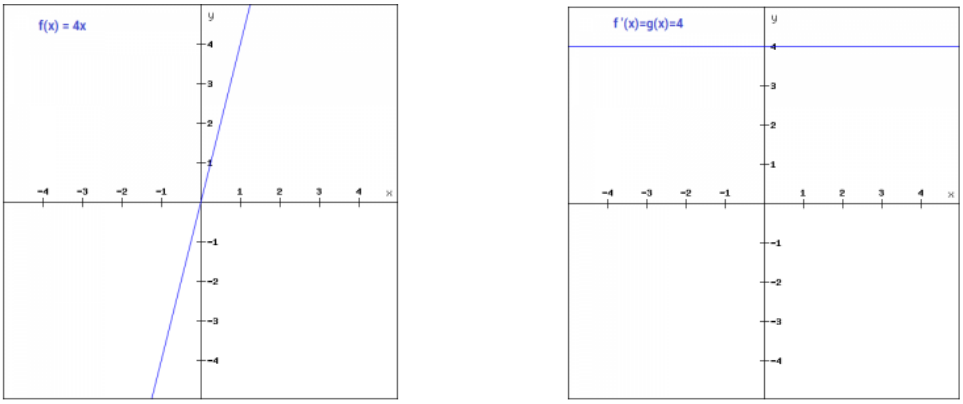



Figure 5.21. Linear activation function and its derivative [Gupta and Dishashree, 2020].

- **Sigmoid:** transforms the values between the range 0 and 1. The mathematical expression for sigmoid is represented in Figure 5.22.

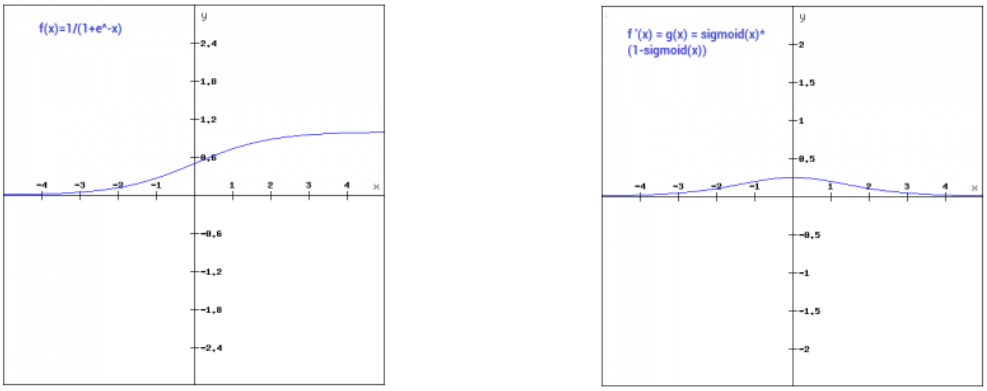



Figure 5.22. Sigmoid activation function and its derivative [Gupta and Dishashree, 2020].

- **Tanh:** is very similar to the sigmoid function. The only difference is that it is symmetric around the origin. The range of values in this case is from -1 to 1. Thus the inputs to the next layers will not always be of the same sign (Figure 5.23) [Gupta and Dishashree, 2020].

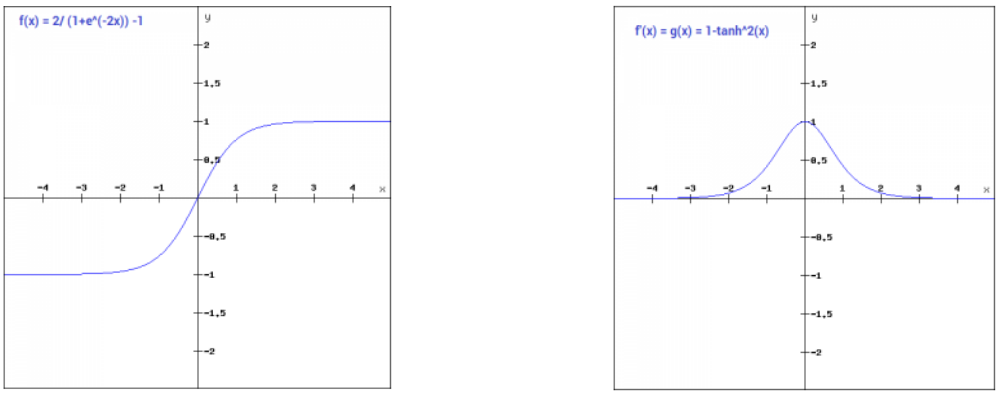



Figure 5.23. Tanh activation function and its derivative [Gupta and Dishashree, 2020].

- **ReLU:** stands for rectified linear unit. The mathematical representation of this function is shown in Figure 5.24.

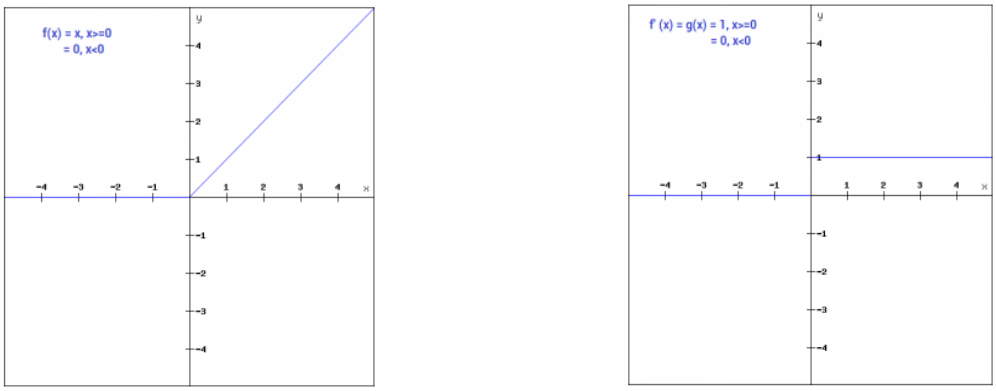



Figure 5.24. ReLU activation function and its derivative [Gupta and Dishashree, 2020].

- **Leaky ReLU (LReLU):** is an improved version of the ReLU function. For the ReLU function, the gradient is 0 for x<0, which would deactivate the neurons in that region. Leaky ReLU is defined to address this problem. Instead of defining the ReLU function as 0 for negative values of x, it is defined as an extremely small linear component of x [Gupta and Dishashree, 2020]. The mathematical expression is shown in Figure 5.25.

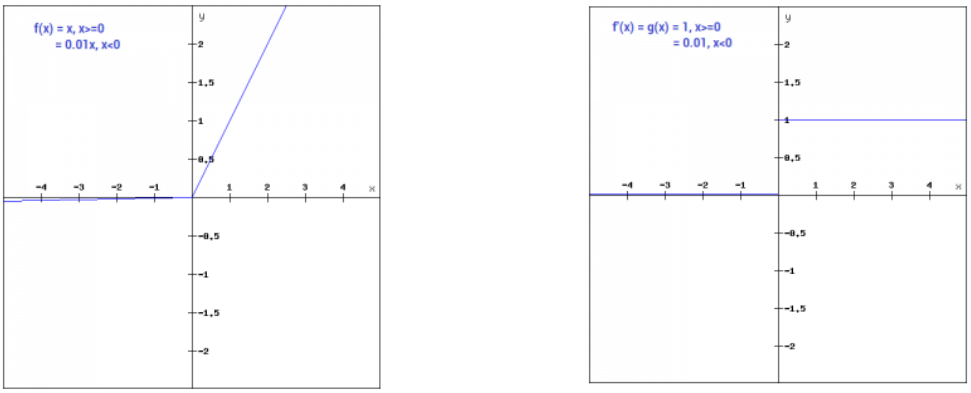



Figure 5.25. Leaky ReLU activation function and its derivatve [Gupta and Dishashree, 2020].

- **Parameterized ReLU (PReLU):** is another variant of ReLU that aims to solve the problem of getting zero gradient for the left half of the axis. The parameterised ReLU introduces a new parameter as a slope of the negative part of the function as shown in Figure 5.26 [Gupta and Dishashree, 2020].

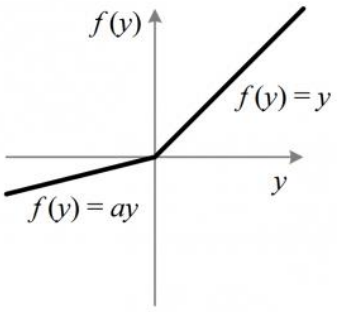



Figure 5.26. Parameterized ReLU activation function [Gupta and Dishashree, 2020].

- **Softmax:** is often described as a combination of multiple sigmoids. Thus sigmoid is widely used for binary classification problems, where the softmax function can be used for multiclass classification problems. This function returns the probability for a data point belonging to each individual class [Gupta and Dishashree, 2020]. The mathematical expression is defined in equation (5.44).

$$\sigma(x_i) = \frac{e^{x_i}}{\sum_{k=1}^{K} e^{x_k}}, i = 1, \dots \dots, K \tag{5.44}$$

The MLP can be categorized either as a fully connected, or as a partially connected network. A fully connected network is connected by all nodes between layers, unlike a partially connected network which does not require connecting all nodes between former and next layers. The partially connected network uses less weights memory; however, it was proved that less connection reduced the performance through the simple experiment.

The MLP Network implemented here for the score level fusion purpose is shown in Figure 5.27. It is composed of a single hidden layer with PReLU activation function, and an output layer (C nodes) with softmax activation. For the hidden layer, ReLU is a preferred choice because its derivative is 1 as long as x is positive and 0 when x is negative, where both sigmoid and tanh functions are not suitable for hidden layers because if x is very large or very small, the slope of the function becomes very small which slows down the gradient descent [Rizwan, 2018]. Here we applied PReLU just to avoid exact zero derivatives. The activation function for the output layer would depend on whether we are performing classification or

regression. For binary classification (i.e. problems of two classes), the Sigmoid function can be used. For multiclass classification (i.e. problems with more than two classes), the softmax function is used. For regression problems (i.e. real-value outputs), the linear/identity function is used [Rizwan, 2018]. Since we are performing multiclass classification, softmax function is selected of the output layer.

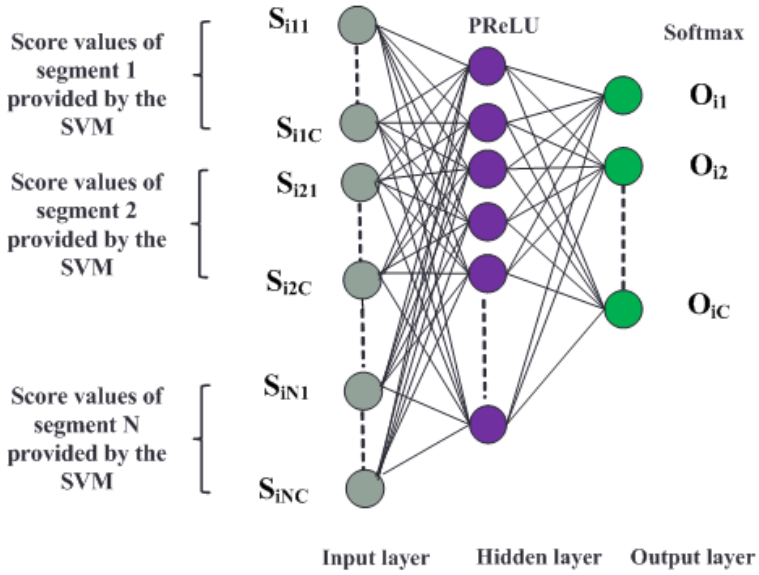



Figure 5.27. The MLP model used for score level fusion of N segments.

Once the MLP architecture is chosen, three different hyper-parameters should be determined:

- **Hidden layer nodes:** the number of hidden layer nodes is calculated in a way to avoid overfitting. The number of independent parameters in our network must be smaller than the number of data points available. In this work, we assume there are no biases. The empirical rule of thumb is to select a form for the fit such that the number of data points is 10 times the number of coefficients, calculated as follow:

$$N_h=\frac{N_s\times N_i}{a\times(N_i+N_o)} \tag{5.45}$$

where $N_s$, $N_i$, and $N_o$ refer to the number of samples in training set (number of subjects in training multiplied by M), the number of input nodes (N×C), and number of output nodes (C) respectively, and *a* represents the scaling factor (in our case it is set to 10).

- **Learning rate:** a high value can cause undesirable divergent loss function, and a low value means that the calculation reaches the solution too slowly. For this reason, step decay for the learning rate is used, where we start with a high learning rate value and then gradually reduce the value by some percentage after a set number of training epochs [Jeremy, 2018-b]. The step decay learning rate used here is shown in Figure 5.28.

- **Initial weights:** weight initialization can actually have a profound impact on both the convergence rate and final quality of a network. Usually, random weight drawn from Gaussian distributions with fixed STD is used. The problem with such method is that in case weights are initially high, the slope of gradient changes slowly and learning takes a lot of time. In case weights are initially low, then the variance of the input signal starts diminishing as it passes through each layer in the network.

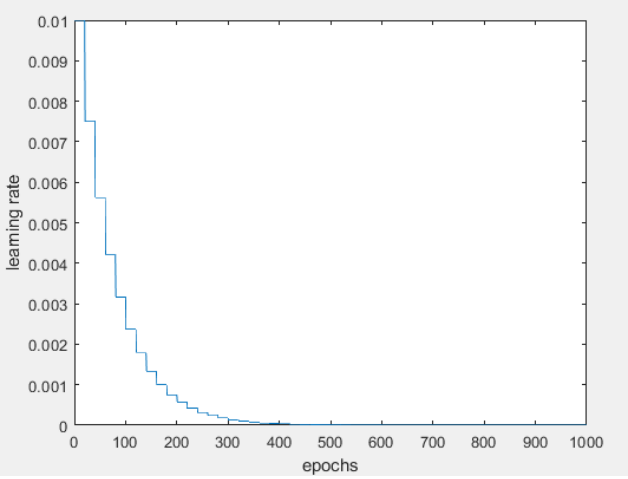



Figure 5.28. The step decay learning rate used in this study.

The input eventually drops to a really low value and can no longer be useful [Yadav, 2020]. This problem is known as vanishing gradient. To keep the

signal from exploding to a high value or vanishing to zero, the variance should remain the same with each forward passing layer. This initialization process is known as Xavier standard initialization [Joshi et al., 2016]. However, Xavier standard initialization is based on the assumption that the activation functions are linear. It can not be applied in our case since we are using PReLU activation function. A new initialization technique was proposed by [He et al., 2015] that is similar to Xavier concept, but it is specifically designed for ReLU/PReLU activation functions. Based on this, the initial weights used here are described in equation (5.46), where dim1 and dim2 refer to the number of neurons in layers 1 and 2 respectively, and n is the number of inputs to the MLP.

$$W_i = \text{randn}(\text{dim2},\text{dim1}) * \text{sqrt}(2/n) \tag{5.46}$$

### 5.2.3.2 Combination approach

As described in section 5.2.3, each subject has M possible observations. When doing training using the MLP model represented in Figure 5.27, each observation in the training is considered an independent training instance. The trained MLP model predicts the label of each of the M observations. The output probability vector fusion was applied to combine the probabilities of all the M observations from one subject in order to acquire the final probability vector for classification decision. The final output probability vector is computed by getting the mean of all the M outputs, and the class label is identified by determining the class with highest value in the final output vector. An overview of the proposed model for H&Y stage detection is described in Figure 5.29.

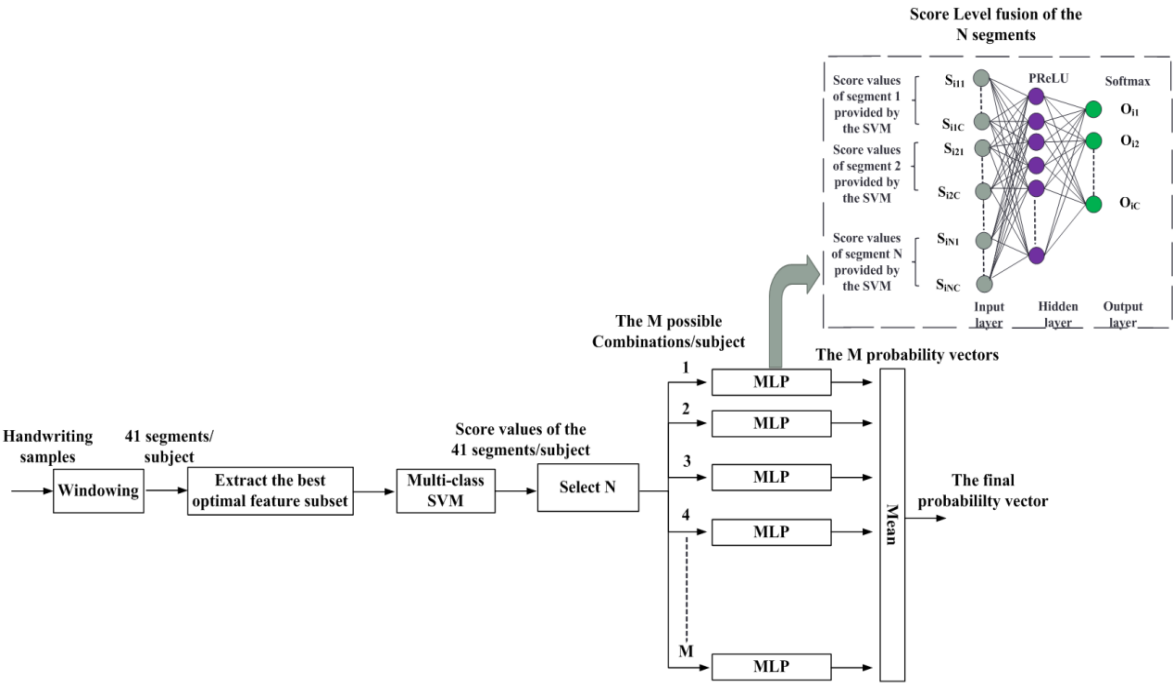



Figure 5.29. Overview of the proposed model for PD stage detection.

## 5.2.4 Imbalanced data

As shown in Figure 5.16, the data set exhibits an unequal distribution between its classes; which is considered imbalanced. The problem of learning from imbalanced data sets is twofold. First, from the perspective of classifier training, imbalance in training data distribution often causes learning algorithms to perform poorly on the minority classes [Jeni et al., 2013]. In addition, the imbalance between majority and minority would lead machine learning to be biased and to produce poor performances if the imbalanced data is used directly [Goel et al., 2013]. A common solution is to re-sample the data prior to training to re-balance the class distribution. The right way to re-sample imbalanced dataset and to avoid overfitting, is by dividing it into training and testing sets and applying a re-sample technique to the training set only. The testing set which contains only original samples is used to evaluate the classification performance [Dubey et al., 2014]. Re-sampling methods can be divided into oversampling and under-sampling techniques. The oversampling technique add copies of instances from the under-represented class, and the under-sampling technique delete instances from the over-represented class. The under-sampling techniques are used in case the

number of samples is large, which is not our case [Brownlee, 2016]. Advanced oversampling methods for imbalanced data exist and are summarized as follows:

- **SMOTE**: in synthetic minority oversampling technique (SMOTE) algorithm minority class is oversampled by generating synthetic examples rather than by oversampling with replacement. The SMOTE algorithm creates artificial examples based on the feature space, rather than data space, similarities between existing minority examples. These synthetic examples are generated along the line segments joining a portion or all of the K nearest neighbors of the minority class. Depending on the amount of the oversampling required, neighbors from the K nearest neighbors are randomly chosen [Sudarsen et al., 2017]. SMOTE algorithm for binary classification is represented in Figure 5.30, where β is a parameter used to specify the desired balance level after generation of the synthetic data. For each minority data example $x_i$, the same number of synthetic data examples are generated according to the following steps [He et al., 2008]:

  Do the **loop** from 1 to N:

  ✓ Randomly choose one minority data example $x_{zi}$ belong to the minority class from the K nearest neighbors for data $x_i$

  ✓ Generate the synthetic data example:

  $s_i = x_i + (xz_i - x_i) \times \lambda$, λ is a random number ∈ [0, 1].

End **loop**

Since SMOTE generates the same number of synthetic data samples for each original minority examples without consideration to neighboring examples belonging to the majority class, the occurrence of overlapping between classes is increased. This reduces the capability to discriminate between classes [Sudarsen et al., 2017].

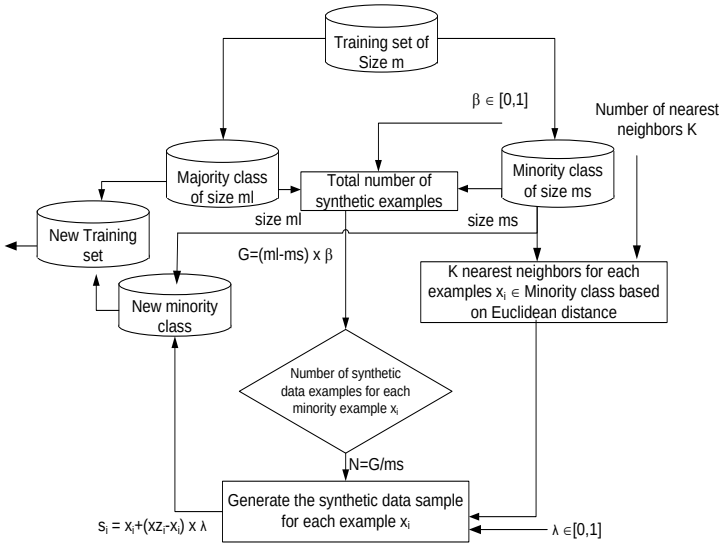



Figure 5.30. SMOTE flowchart.

- **ADASYN:** adaptive synthetic sampling approach (ADASYN) method has been proposed to overcome the limitation existing in SMOTE approach, i.e. avoid overlapping between classes due to the synthetic data added [Sudarsen et al., 2017]. ADASYN is based on the idea of adaptively generating minority data samples according to their distributions; the more minority class samples are harder to learn, the more synthetic data is generated [He et al., 2008]. ADASYN algorithm for the two-class classification problem is summarized in the flow chart in Figure 5.31. For each minority data example $x_i$, different $g_i$ synthetic data examples are generated according to the steps mentioned in SMOTE algorithm. The number of synthetic examples $g_i$ is related to the number of neighboring examples belonging to the majority class. ADASYN/SMOTE can be generalized to multiple-class imbalanced learning problems as well. To extend the ADASYN/SMOTE idea to multi-class problems, one needs first to calculate and sort the degree of class imbalance for each class with respect to the most significant class [He et al., 2008].

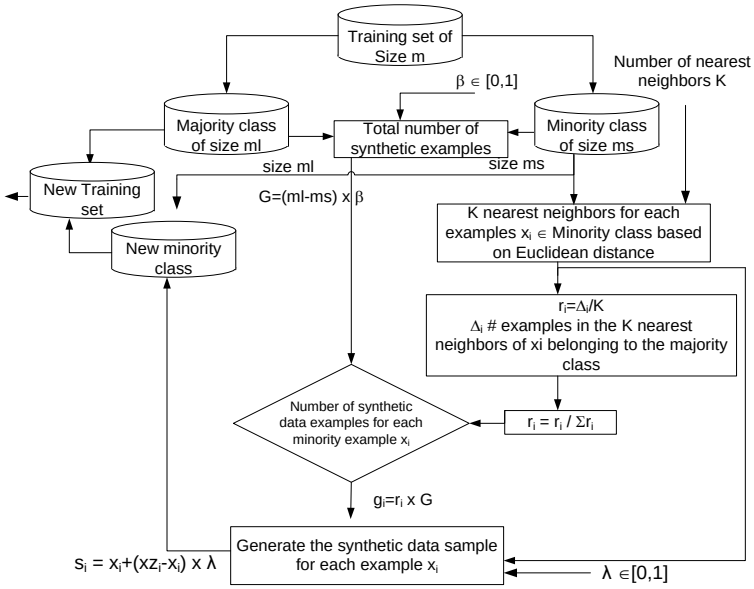



Figure 5.31. ADASYN flowchart.

## 5.2.5 Data sampling approaches studied

For the task of predicting the disease progression, re-sampling only the training dataset prior to classification is done. Two basic data sampling approaches in addition to the no sampling approach were examined. In each fold of the cross validation, the training set is re-sampled and the SVM model is tested on the new re-sampled training set and the original test set. The data sampling approaches studied in this work are:

1. **No Sampling:** all of the data points from majority and minority training sets are used.

2. **ADASYN and SMOTE oversampling:** these techniques for multiclass are described above. β is set to 1 for a fully balanced data set, and the number of nearest neighbors is set to 5.

These sampling techniques are applied to generate synthetic data per participant level and not segment level, where the 41 feature segments are concatenated together forming a single feature vector. The new generated sample is later divided into 41 equal segments. The concept used is summarized in Figure 5.32.

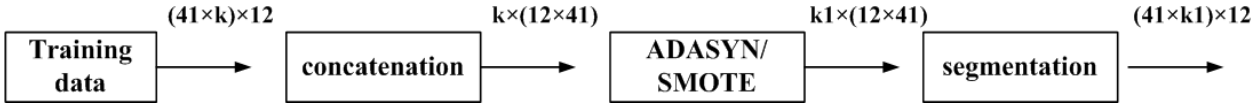



Figure 5.32. The concept used in ADASYN and SMOTE sampling. k refers to the number of subjects in the training data, and k1 represents the number of subjects in the re-sampled training data.

## 5.2.6 Experiments and results

Many obstacles are faced in this work. First of all, PD stage classification is a very challenging task due to the large variability over patients (2 patients may have the same stage of disease but with different symptoms, or 2 patients with adjacent stages may have the same symptoms), irrelevant motion interference (some irrelevant motions may appear, causing certain uncertainty), and the limited availability of data (especially in this case where we are working with multiclass classification). In addition, imbalanced data and classes distribution in each fold is another problem. Imbalanced data may lead to biased machine learning and to poor performances, where also the unequal class distribution in the folds may increase the overfitting. To overcome such obstacles, many experiments were studied and compared, where each time the number of subjects or the selected features varies. The prediction performance is evaluated in terms of the accuracy, sensitivity, and specificity.

| | | Predicted Number | | | |
|---|---|---|---|---|---|
| | | Class 1 | Class 2 | … | Class *n* |
| Actual Number | Class 1 | $x_{11}$ | $x_{12}$ | … | $x_{1n}$ |
| | Class 2 | $x_{21}$ | $x_{22}$ | … | $x_{2n}$ |
| | . . . | . . . | . . . | . . . | . . . |
| | Class *n* | $x_{n1}$ | $x_{n2}$ | … | $x_{nn}$ |

Figure 5.33. Multiclass classification confusion matrix.

Given the confusion matrix defined in Figure 5.33, TP, TN, FP, and FN are calculated for each class *i* using the following equations [Manliguez, 2016]:

$$TP_i = X_{ii} \quad (5.47) \qquad FP_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} X_{ji} \quad (5.48)$$

$$TN_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} \sum_{\substack{k=1 \\ k \neq i}}^{n} X_{jk} \quad (5.49) \qquad FN_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} X_{ij} \quad (5.50)$$

Accuracy, sensitivity, and specificity are obtained as follows, where $TP_{all}, TN_{all}, FP_{all}$, and $FN_{all}$ are defined in equations (5.51), (5.52), (5.53), and (5.54).

Accuracy= $sum(TP_{all}./(TP_{all} + TN_{all} + FP_{all} + FN_{all})) \times 100$

Sensitivity= $mean(TP_{all}./(TP_{all} + FN_{all})) \times 100$

Specificity= $mean(TN_{all}./(TN_{all} + FP_{all})) \times 100$

$$TP_{all} = [TP_1, TP_2, \ldots\ldots, TP_n] \quad (5.51) \qquad FP_{all} = [FP_1, FP_2, \ldots\ldots, FP_n] \quad (5.52)$$

$$TN_{all} = [TN_1, TN_2, \ldots\ldots, TN_n] \quad (5.53) \qquad FN_{all} = [FN_1, FN_2, \ldots\ldots, FN_n] \quad (5.54)$$

In the first experiment, all the 32 subjects are studied where the set of features selected in section 5.1 are used. According to Figure 5.16, the number of samples in both stages 3 and 4 is very small; which means that these classes will not be represented in all the folds. To avoid the possibility of testing a class that is not represented in the training data, both stages 3 and 4 samples are eliminated in the second experiment. The results obtained in both experiments 1 and 2 are shown in Table 5.13. In order to check whether this bad performance is caused by the insufficiency of the training data or a bug in the implementation, another experiment is conducted; where the dataset is distributed into 2 classes (one class for PD and another for HC). According to Table 5.13, we can say that the low performance is related to the small training data.

The first three experiments described above deal with the set of features selected in section 5.1. The feature selection process is done where the dataset is distributed into 2 classes (class 1 for HC and class 2 for PD) each task is considered as 1 segment. Since here

we are applying window segmentation of each task, it would be better to reselect features using the two-stage feature selection approach described in section 5.1.2, where window segmentation is applied on each task. The performance obtained in this experiment is shown in Table 5.13, where the new selected features are represented in Table 5.15. The new set of features also is a combination of kinematic, pressure, and correlation between kinematic and pressure features; which confirm our finding in our previous work. Based on these features, experiments 1 and 2 are repeated. We can see how this new feature set improves the accuracy of PD stage detection by 6.25 % when the 6 stages are included, and by 3.44 % when only 4 stages are studied.

The main contribution is to ensure that the model not only fits the training data well but also generalizes on test/validation data. To do that, samples with stages 1.5, 3, and 4 are eliminated to ensure the same class distribution in training, validation, and test data. According to Table 5.13, an improvement in PD stage detection accuracy is shown.

Since we are working with imbalanced data, this would lead our model to be biased and to produce poor performances. Before applying the sampling techniques described before to generate new synthetic samples, we have decided to study only classes that are approximately balanced (stages 0 and 2) to see how balanced data can play an important role in classification performance. The improvement obtained can be explained by the fact that classifiers attempt to reduce global quantities such as the error rate, and since the classes are somehow balanced this means that the number of misclassified samples will decrease.

According to the description of the modified H&Y stages in chapter 2, the signs of adjacent H&Y stages are somehow similar. To reduce the variability over patients, we decided to apply stage group classification instead of exact stage classification by grouping samples with adjacent H&Y stages into single class. Stages 1 and 1.5 are grouped into single class, the same for stages 3 and 4. It is clear how stage group classification performance is better than the exact stage classification (the accuracy is improved from 56.25 % to 62.5 %).

Finally, due to the limited data used, and since our target is to get a good performance without eliminating or merging any classes, a new experiment is performed where all the stages samples are included. The difference between this experiment and experiment 5 is that

once the hyper-parameter N is selected via the validation data, the model is retrained with the combination of both training and validation data. An improvement has been detected in the classification accuracy (from 56.25 % to 62.5 %) as shown in Table 5.13.

The exact overall performances obtained for all the experiments defined above are presented in Table 5.13, where the accuracies per class are shown in Table 5.14.

For each experiment where we have imbalanced data, the 2 sampling techniques (SMOTE and ADASYN) were also applied to rebalance the training data. In experiment 10 even though we are working with imbalanced data, we did not study any of the sampling techniques since the combination of the rebalanced training and the imbalanced validation after hyper-parameter selection will yield to new imbalanced training data. The results obtained with the resampling techniques are presented in Table 5.13. The results show that PD stage classification with artificially balanced data is worse than classification with imbalanced data. Modifying the dataset with resampling-like methods is changing the reality, so it can lead to a relevant approach or can also be of nonsense to just rebalance the classes. Imbalance data problem depends on many factors: the real purpose of the classifier, the number and distribution of samples, and how easy classes are separable. So, it is not always a good idea to resample the training dataset, especially when classes are not separable. When applying SMOTE or ADASYN resampling techniques, the synthetic examples are generated along the line segments joining a portion or all of the K nearest neighbors of the minority class as shown in Figure 5.34. In case a majority sample is nearby, the probability of overlapping between classes will increase and the capability to discriminate between classes will decrease.

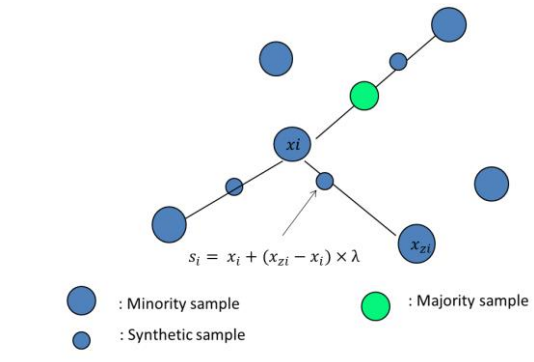



Figure 5.34. SMOTE or ADASYN resampling techniques limitation in case of non-separable classes.

Another way to reduce the variability over patients is to calculate the 1-off accuracy, where the 1-off accuracy represents the accuracy when the result is off by 1-adjacent stage label left or right [Qawaqneh et al., 2017]. The results obtained shown how the 1-off accuracy is higher than the exact one, and how the 1-off left accuracy (when the predicted stage label is off by 1-adjacent stage label left) is higher than the 1-off right accuracy (when the predicted stage label is off by 1-adjacent stage label right). This finding can be explained by the fact that PD patients are studied in their "on-state", and in some cases, patients on dopaminergic medication tend to look like in an earlier stage of the disease. The 1-off accuracy is only calculated in the experiments where the stages of the samples are adjacent.

Table 5.13. Comparison of different sampling techniques in predicting PD H&Y stage.

| Exp. | Sampling method | N | Exact Perf (%) | 1-off left Perf (%) | 1-off right Perf (%) | 1-off left or right Perf (%) |
|---|---|---|---|---|---|---|
| 1 | **The 6 H&Y stages samples are studied with the set of features found in section 5.1** | | | | | |
| | No sampling | 1 | Acc: 50<br>Sens: 16.67<br>Spec: 83.76 | Acc: 59.37<br>Sens: 29.17<br>Spec: 87.84 | Acc: 50<br>Sens: 16.67<br>Spec: 83.76 | Acc: 59.37<br>Sens: 29.17<br>Spec: 87.84 |
| | SMOTE | 21 | Acc: 50<br>Sens: 16.67<br>Spec: 84.68 | Acc:59.37<br>Sens: 29.17<br>Spec: 87.78 | Acc: 50<br>Sens: 16.67<br>Spec: 84.68 | Acc: 59.37<br>Sens: 29.17<br>Spec: 87.78 |
| | ADASYN | 36 | Acc: 46.87<br>Sens: 18.75<br>Spec: 83.89 | Acc: 53.13<br>Sens: 27.08<br>Spec: 85.98 | Acc: 50<br>Sens: 24.31<br>Spec: 85.33 | Acc: 56.25<br>Sens: 32.64<br>Spec: 86.62 |
| 2 | **Only the first 4 stages samples are studied with the set of features found in section 5.1** | | | | | |
| | No sampling | 1 | Acc: 58.63<br>Sens: 31.25<br>Spec: 77.76 | Acc: 65.51<br>Sens: 43.75<br>Spec: 81.61 | Acc: 58.63<br>Sens: 31.25<br>Spec: 77.76 | Acc: 65.51<br>Sens: 43.75<br>Spec: 81.61 |
| | SMOTE | 12 | Acc: 55.17<br>Sens: 25<br>Spec: 75.84 | Acc: 65.52<br>Sens: 43.75<br>Spec: 81.61 | Acc: 55.17<br>Sens: 25<br>Spec: 75.84 | Acc: 65.52<br>Sens: 43.75<br>Spec: 81.61 |
| | ADASYN | 5 | Acc: 55.17<br>Sens: 25<br>Spec: 76.67 | Acc: 62.07<br>Sens: 37.5<br>Spec: 80.69 | Acc: 55.17<br>Sens: 25<br>Spec: 76.67 | Acc: 62.07<br>Sens: 37.5<br>Spec: 80.69 |
| 3 | **Binary classification where 2 classes are studied (HC or PD) with the set of features found in section 5.1** | | | | | |
| | No sampling | 40 | Acc: 78.12<br>Sens: 56.25<br>Spec: 100 | | | |
| 4 | **Binary classification where 2 classes are studied (HC or PD) with the new set of features described in Table 5.15** | | | | | |
| | No sampling | 40 | Acc: 78.12<br>Sens: 68.75<br>Spec: 87.5 | | | |
| 5 | **The 6 H&Y stages samples are studied with the set of features described in Table 5.15** | | | | | |
| | No sampling | 38 | Acc: 56.25<br>Sens: 23.61<br>Spec:88.04 | Acc: 68.75<br>Sens: 44.44<br>Spec: 90.9 | Acc: 59.37<br>Sens: 27.78<br>Spec: 88.62 | Acc: 71.87<br>Sens: 48.61<br>Spec: 91.47 |
| | SMOTE | 39 | Acc: 53.12<br>Sens: 22.57<br>Spec:89.29 | Acc: 56.25<br>Sens: 28.13<br>Spec:89.89 | Acc: 59.38<br>Sens: 27.78<br>Spec:90.46 | Acc: 62.5<br>Sens: 33.33<br>Spec:91.06 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | ADASYN | 5 | Acc: 46.87<br>Sens: 19.1<br>Spec:87.69 | Acc: 50<br>Sens: 24.65<br>Spec:88.28 | Acc: 50<br>Sens: 23.26<br>Spec:88.26 | Acc: 53.13<br>Sens: 28.82<br>Spec:88.86 |
| 6 | **Only the first 4 stages samples are studied with the set of features described in Table 5.15** | | | | | |
| | No sampling | 6 | Acc: 62.07<br>Sens: 33.33<br>Spec: 84.25 | Acc: 68.96<br>Sens: 47.92<br>Spec: 87.17 | Acc: 65.51<br>Sens: 39.58<br>Spec: 85.21 | Acc: 72.41<br>Sens: 54.17<br>Spec: 88.13 |
| | SMOTE | 33 | Acc: 55.17<br>Sens: 30.21<br>Spec: 83.96 | Acc: 62.07<br>Sens: 46.87<br>Spec: 85.96 | Acc: 58.62<br>Sens: 36.46<br>Spec: 84.92 | Acc: 65.52<br>Sens: 53.13<br>Spec: 86.92 |
| | ADASYN | 37 | Acc: 58.62<br>Sens: 34.37<br>Spec: 85.75 | Acc: 65.52<br>Sens: 51.04<br>Spec: 87.76 | Acc: 62.07<br>Sens: 40.63<br>Spec: 86.72 | Acc: 68.97<br>Sens: 57.29<br>Spec: 88.72 |
| 7 | **Samples with stages 1.5, 3 and 4 are eliminated with the set of features described in Table 5.15** | | | | | |
| | No sampling | 6 | Acc: 73.08<br>Sens: 52.78<br>Spec: 85.3 | | | |
| | SMOTE | 37 | Acc: 73.08<br>Sens: 56.25<br>Spec: 85.15 | | | |
| | ADASYN | 26 | Acc: 57.69<br>Sens: 57.64<br>Spec: 80.15 | | | |
| 8 | **Samples that belong to balanced classes are studied with the new selected features** | | | | | |
| | No sampling | 38 | Acc: 90.91<br>Sens: 66.67<br>Spec: 100 | | | |
| | SMOTE | 7 | Acc: 81.82<br>Sens: 81.25<br>Spec: 83.33 | | | |
| | ADASYN | 22 | Acc:72.73<br>Sens: 68.75<br>Spec: 83.33 | | | |
| 9 | **Stage group classification studied instead of exact stage classification with the new features selected** | | | | | |
| | No sampling | 28 | Acc: 62.5<br>Sens: 39.88<br>Spec: 84.1 | | | |
| | SMOTE | 27 | Acc: 56.25<br>Sens: 38.76<br>Spec: 83.54 | | | |
| | ADASYN | 11 | Acc: 56.25<br>Sens: 39.51<br>Spec: 83.54 | | | |
| 10 | **The 6 H&Y stages samples are included where the training is augmented by combining training and validation after hyper-parameter selection** | | | | | |
| | No sampling | 38 | Acc: 62.5<br>Sens: 32.29<br>Spec: 90.44 | Acc: 68.75<br>Sens: 46.18<br>Spec: 91.68 | Acc: 65.63<br>Sens: 33.33<br>Spec: 91.04 | Acc: 71.87<br>Sens: 47.22<br>Spec: 92.27 |

Table 5.14. Overall and per class classification accuracies.

| Exp. | Class 1 (Stage 0) | Class 2 (Stage 1) | Class 3 (Stage 1.5) | Class 4 (Stage 2) | Class 5 (Stage 3) | Class 6 (Stage 4) | Exact Overall Accuracy (%) |
|---|---|---|---|---|---|---|---|
| 1 | 100 | 0 | 0 | 0 | 0 | 0 | 50 |
| 2 | 100 | 25 | 0 | 0 | | | 58.62 |
| 3 | 100 | 56.25 | | | | | 78.12 |
| 4 | 87.5 | 68.75 | | | | | 78.12 |
| 5 | 100 | 25 | 0 | 16.67 | 0 | 0 | 56.25 |
| 6 | 100 | 0 | 0 | 33.33 | | | 62.07 |
| 7 | 100 | 25 | | 33.33 | | | 73.08 |
| 8 | 100 | | | 66.67 | | | 90.91 |
| 9 | 100 | 42.86 | | 16.67 | 0 | | 62.5 |
| 10 | 93.75 | 50 | 0 | 50 | 0 | 0 | 62.5 |

Table 5.15. The new selected features set obtained with windowed segmentation.

| **Feature** | **Statistic** |
|---|---|
| Rising edge: correlation between pressure and vertical velocity | STD |
| Rising edge: correlation between pressure and horizontal acceleration | STD |
| Main part: correlation between pressure and vertical velocity | STD |
| Main part: correlation between pressure and vertical acceleration | Mean |
| Falling edge: NCP | STD |
| Acceleration | Mean |
| Stroke height | STD |
| NCA/stroke | STD |

## 5.2.7 Conclusions

Developing a reliable handwriting-based system that assists in the decision-making process to diagnose PD at an early stage, and to predict the H&Y stage is the objective of this work. PD is often difficult to diagnose, but even at an early stage, small handwriting differences may be machine-detectable. Due to the small size data and the multiclass distribution, windowed segmentation was applied on each task. A 4 fold cross-validation multi-class SVM classifier with RBF model was applied, where we ensured that no windows in training, validation and testing referring to the same participant. An MLP model was used to combine the scores of all segments from one subject in order to acquire the final scores for classification decision.

It is well known that classification is sensitive to features and data size, for this reason many experiments were studied and analyzed, where each time the number of samples or the set of features selected varies. Two data resampling techniques (ADASYN and SMOTE) used to rebalance the training data before classification, in addition to the no sampling approach

were examined and compared. In addition, since certain uncertainty in classification may occur due to the small features variation between adjacent stages, the exact and the 1-off accuracies are both calculated.

Many obstacles were faced in this part such as: large variability over patients in term of symptoms and stage of disease, imbalanced data and class distribution, and limited number of samples. Re-balancing training data, 1-off accuracy, stage group classification, ensuring the same class distribution in all the subsets are all studied to see how the factors mentioned before affect the classification performance. The results obtained confirm that similar class distribution in all the subsets, and balanced data play major roles in the classification performance, and that either the 1-off accuracy or stage group classification can overcome the large variability over patients' problem. In addition, combining training and validation datasets after hyper-parameters selection improves the classification performance and emphasizes the importance of the amount of data available in PD classification and especially in PD stage prediction.

Finally, applying resampling-like methods to rebalance training data is not straightforward because it depends on many factors such as the real purpose of the classifier, the number and distribution of samples, and whether classes are easily separable or not. Training a classifier with rebalanced training data may return better or worse performance than the classifier trained on the unchanged training data, depending on many factors.

As a conclusion, the performance obtained in this work is not satisfactory and not nearly compelling as the one obtained with the binary classification (section 5.1). Predicting PD stage using a limited data is a challenging task even with windowing technique, especially when the patients are studied in their on-state, since levodopa medication in most cases reduces the symptoms of PD, where it also may contribute to the development of dyskinesia or uncontrolled movements (mentioned in chapter 2). In both cases, it is difficult to separate between the stages because patient can even get nearer to the non-PD or early stages, or nearer higher stages. As a future work, it will be important to perform our modal on a large and balanced dataset in both the "on-state" and the "off-state" cases.

## 5.3 Visual representation and deep learning for Parkinson's disease early detection

Based on our previous work described in section 5.1, we have found that handwriting can be a tool for PD diagnosis with a 97 % prediction accuracy when a combination of kinematic, pressure, and correlation between kinematic and pressure features is used. However, since hand-crafted features model required expert knowledge of the field, and since we are working with small database; this motivate us to learn pen-based features by means of deep learning where short term analysis is applied to avoid losing some important information while applying global features extraction. We focus in this work on the definition of a deep computer-based PD early detection system from handwriting; where deep models should be trained on all the languages in order to obtain a language-independent feature vector. Machine learning and now deep learning are popular approaches for signal and image classification. They have been applied to the classification of sequences such as the speech signal, handwriting textline images (seen as a sequence using a sliding window approach), and text (for translation). According to the state of art chapter (chapter 3), [Pereira et al., 2018], [Moetesum et al., 2019], [Gallicchio et al., 2018], and [Khatamino et al., 2018] have applied deep learning to the classification of subjects into PD and HC.

In contrast to classical machine learning approaches, deep learning approaches automatically extract features from raw signals/records. This is a clear advantage since the DNN model is able to extract the best features associated to each task, in an automatic way using end-to-end training.

When dealing with online handwriting time series, the variation over the time axis defines a challenge for models requiring a fixed dimension input. One approach would consist in normalizing the time series leading to a fixed dimension visual representation as suggested in [Pereira et al., 2018]. An alternative approach explicitly considers the time dimension, especially that the variation over this dimension is nonlinear. In this work, we propose two deep learning classes of models in order to tackle the time variation in time series classification. In the first one, 2D representations of time series are fed in the 2D CNN. The

second one integrates the time variation within a Long-Short-Term Memory Networks model which is directly applied on the time series [Taleb et al., 2019-a].

Since each sensor outputs the whole signal acquired during the handwriting task, we can represent such data as a Time series, as depicted in Figure 5.35 that represents the output of task1 from a healthy subject and a PD patient. The differences between drawings are not intuitively recognizable, where the signals extracted from PD patient are noisier than those of the HC subject. As a conclusion, online handwriting dynamic signals can provide more detailed and complex information for PD detection. Here comes the importance of studying the extracted signals instead of the drawings. In this work, we decided to study the whole handwriting dynamic signals so we can extract both in-air and on-surface features.

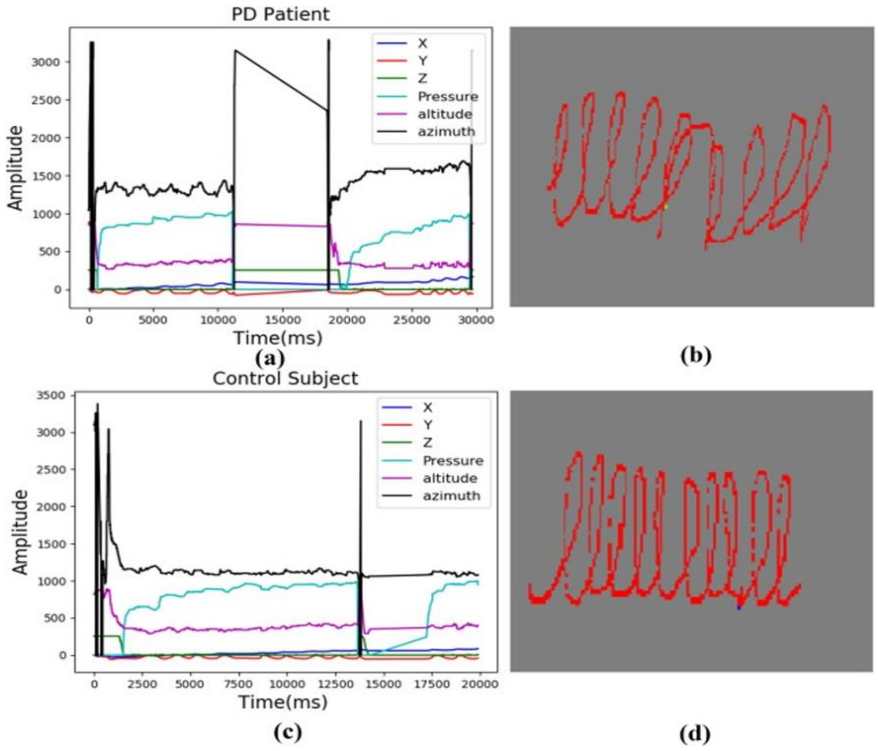



Figure 5.35. Cursive repetitive 'l' samples from PD patient (b) and HC subject (d), and their respective signals in (a) and (c)

## 5.3.1 Deep Learning for time series classification

In recent years, CNNs have shown excellent performance on image classification tasks [Gamboa and Borges, 2017]. In order to benefit from this, Pereira et al. [Pereira et al., 2018] proposed to transform a Time series into an image and use it as an input to 2D CNN,

which will be able to learn features that are used to distinguish healthy individuals from PD patients. From the other side, Long-Short-Term Memory Networks (LSTM) is a family of neural networks that excels in learning from sequential data and can cope with variable length time series [Atienza, 2017]. LSTMs are quite popular in dealing with text based data, and have been quite successful in language translation and text generation [Burakhimmetoglu, 2017]. Since LSTMs can store information during long time intervals, they are thus also appropriate for processing time series representing handwriting signals [Burakhimmetoglu, 2017].

In this work, a comparison between 1D CNN-BLSTM and 2D CNN models is done; where for the 2D CNN two new approaches are proposed to encode the raw time series into 2D images and compared to the one proposed by [Pereira et al., 2018].

#### 5.3.1.1 Pre-processing

As mentioned in chapter 4, HandPDMultiMC is a multilingual dataset with Arabic, French, and English samples. In this work, HandPDMultiMC size was enlarged to 42 subjects (21 PD and 21 HC). In order to have the same writing direction, the X coordinates of the Arabic samples are flipped as shown in Figure 5.36. After that, to achieve a uniform range across all subjects, each X and Y in the input dataset is normalized by subtracting the minimum and the mean, respectively (the minimum and the mean are calculated for each sequence separately) as shown in equations (5.55) and (5.56). In addition, for all our models, 2D CNN and the 1D CNN-LSTM, all images and raw time series are normalized to the range (0, 1) to achieve a uniform contrast and intensity range, where the scaling factors are obtained from training data and used to scale the test data.

$$X_{new} = X_{flipped} - \min(X_{flipped}) \tag{5.55}$$

$$Y_{new} = Y - \text{mean}(Y) \tag{5.56}$$

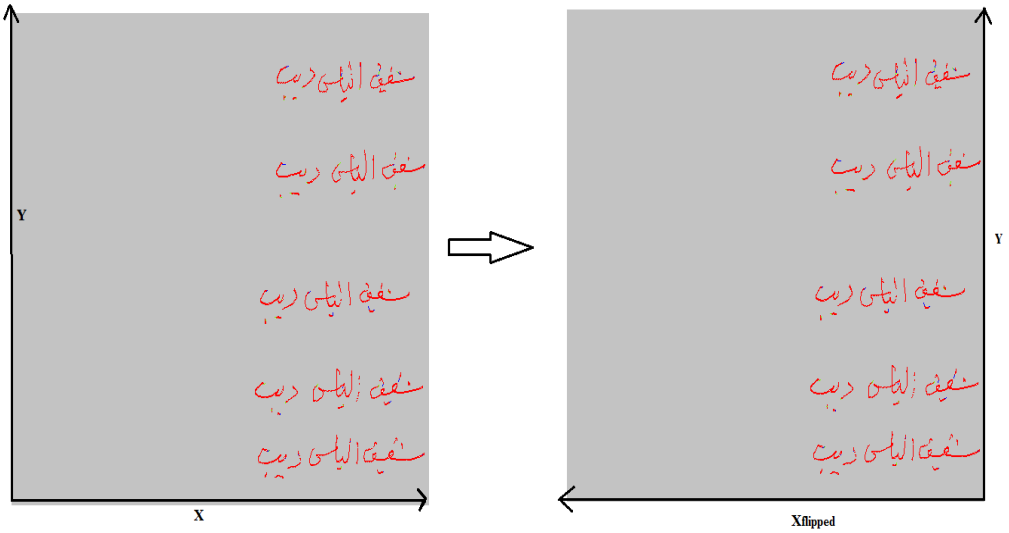



Figure 5.36. Flipping the X coordinate of the Arabic samples.

### 5.3.1.2 2D representations of time series

In the following, we describe how to obtain images from the time series related to a single task. Each handwriting task is composed of n rows (time in ms) and 7 columns (stand for X, Y, Z, pressure, altitude, azimuth, and time stamp). In order to get the best signals combination, the number of time series features used is a hyper-parameter varying between 1 and 7 (total number of extracted signals) and defined by k.

Three different approaches for encoding time series as images are applied; where one approach transforms the whole data (n×k matrix) into a single image, and the two others convert each feature signal into a separate image. Usually the converted image size depends on n; when the length of time series is n. The size of time series n differs from one person to another, because a person may take longer than another to perform the exam, and from one experiment to another for the same person. To keep the number of input feature maps identical for the 2D CNN, images sizes should be same for all the subjects. In this work, grayscale images of size 64×64 are considered and time series are normalized in order to be represented as such images.

#### 5.3.1.2.1 Concatenation approach (time series-based)

This representation is inspired from Pereira et al. [Pereira et al., 2018]. It consists in transforming k time series (corresponding to k measurements) of length n into a single image. The whole data (n×k matrix) is transformed into one image by concatenating the n rows into one vector and then reshaping it into a square matrix of size sqrt(n×k) × sqrt(n×k). This squared matrix is resized to 64×64 pixels resolution using down-sampling technique. Many down-sampling techniques exist such as Bilinear, Bicubic, Wavelet, and Lanczos. Ye et al. [Ye et al., 2005] showed that Lanczos algorithm performs better than all other techniques in image resizing because it attempts to reconstruct the image by using a series of overlapping sine waves to produce what is called a "best fit" curve. Based on this finding, Lanczos resampling technique is applied among this work for image resizing, where this method uses a convolution kernel to interpolate the pixels of the input image in order to calculate the pixel values of the output image. The Lanczos convolution kernel K(x) is defined as follows:

$$K(x) = \begin{cases} sinc(x).\, sinc\left(\frac{x}{a}\right); & if\ |x| < a \\ 0\ otherwise \end{cases}$$

where a is a positive integer which determines the size of the kernel and sinc(x) function is defined as *sin(x)/x*.

An overview of this approach is shown in Figure 5.37. In contrast to Pereira et al. [Pereira et al., 2018] we search for the best k measurements to include in the 2D representation. The time series-based image of a patient and a healthy subject concerning the 7 tasks are shown in Figure 5.38. It is clear that there is a difference between the HC subject and the PD patient, and a difference between tasks; which means that each task seems to capture distinct information.

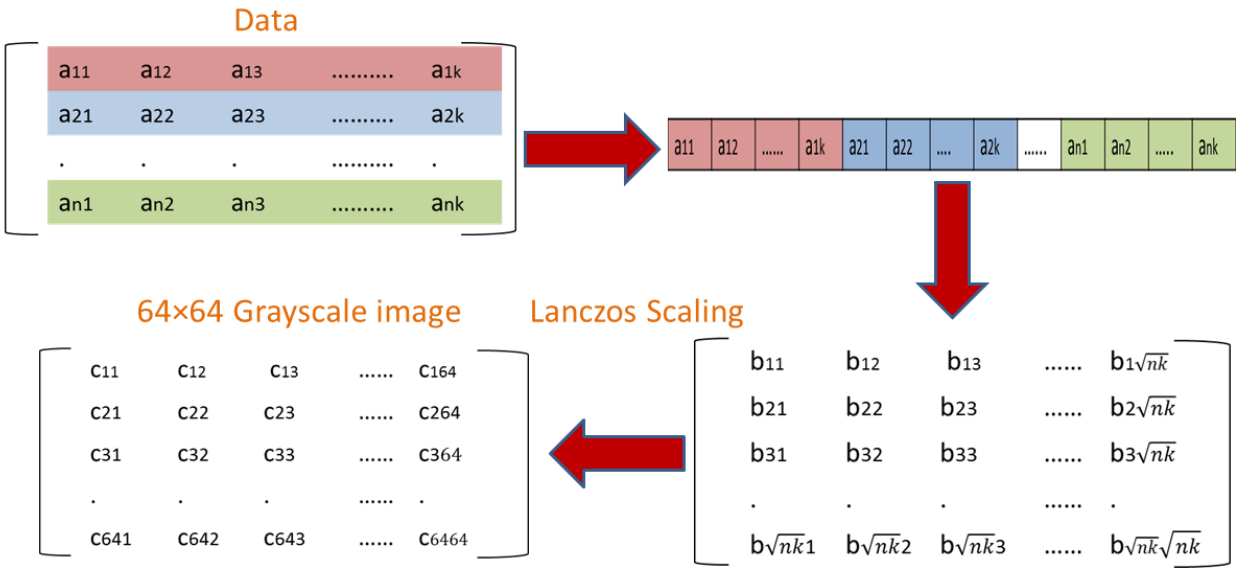



Figure 5.37. Time-series based image representation (k time series from a single task).

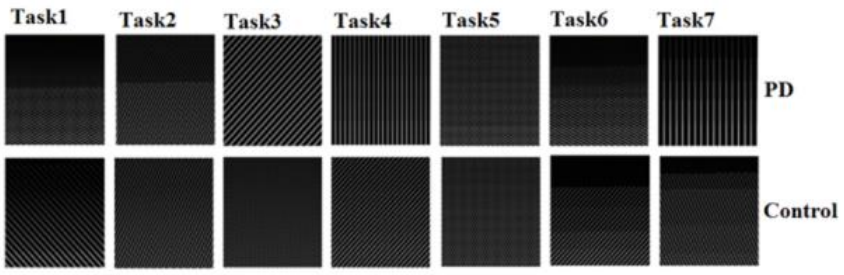



Figure 5.38. The Time series-based images of a PD patient (first row) and a HC subject (second row) concerning the 7 tasks when k=7.

#### 5.3.1.2.2 Modified gramian angular field

Gramian angular field (GAF) is applied here to encode each feature time series as image. The purpose here is to represent the time series in a polar coordinate system instead of the Cartesian coordinates [Wang et al., 2015]. Given a time series S={$s_1$, $s_2$….., $s_n$} of n observations, S is first rescaled to $\tilde{S}$ so all the values fall in the interval [-1,1] then transformed to polar coordinate using these equations:

$$\tilde{S}=\frac{2S-\min(S)-\max(S)}{\max(S)-\min(S)} \tag{5.57}$$

$$\varphi_i = \arccos(\tilde{s}_i), -1 \leq \tilde{s}_i \leq 1, \tilde{s}_i \in \tilde{S} \tag{5.58}$$

$$r_i = \frac{t_i}{C}, t_i \in \mathbb{N} \tag{5.59}$$

where $t_i$is the time stamp and C is a constant factor used to regularize the span of the polar coordinate system. After transforming the rescaled time series into the polar coordinate system, the angular perspective is exploited by considering the trigonometric sum/difference between each couple of points to identify the temporal correlation within different time intervals. The gramian Summation Angular Field (GASF) and gramian Difference Angular Field (GADF) are defined as follows:

$$GASF_{i,j} = \left[\cos(\varphi_i + \varphi_j)\right] \tag{5.60}$$

$$GADF_{i,j} = \left[\sin(\varphi_i - \varphi_j)\right] \tag{5.61}$$

The size of gramian matrix is n×n. In order to get the same image size (64×64) for all samples, piecewise aggregation approximation (PAA) is used to smooth the time series while keeping trends, prior transforming into polar coordinate, by dividing the time series of length n into 64 equi-sized "frames". The mean value of the time series falling within a frame is calculated and a vector of these values becomes the data reduced representation [Wang et al., 2015]. An overview of gramian angular field approach is represented in Figure 5.39. GADF and GASF are applied to each time series separately; this means that for each exam we will get 7 different images. GADF and GASF of a PD patient and a HC subject concerning task1 for each of the 7 signals are shown in Figure 5.40.

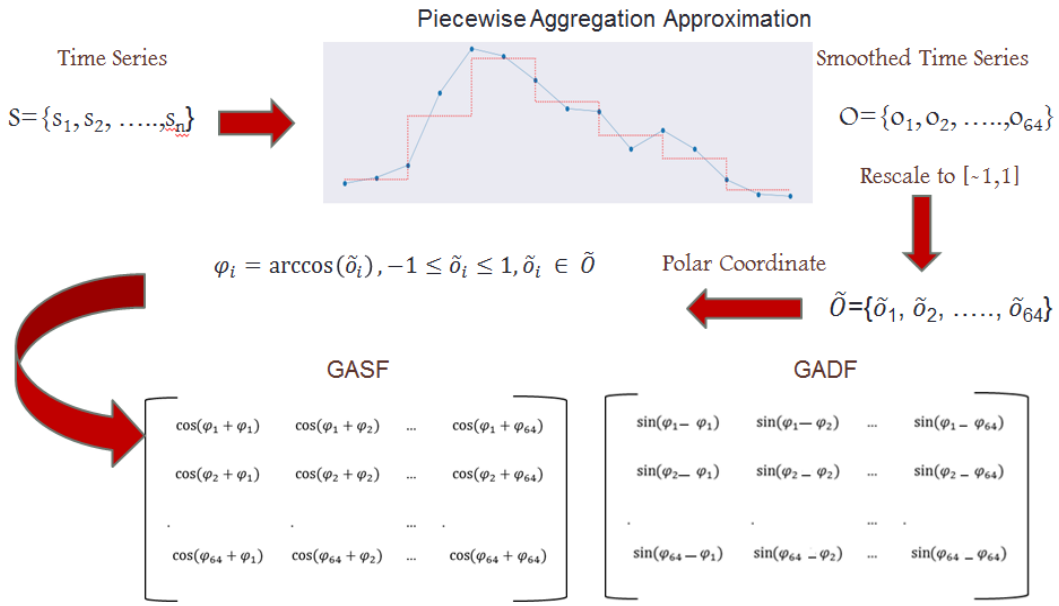



Figure 5.39. An overview of Gramian Angular Field approach.

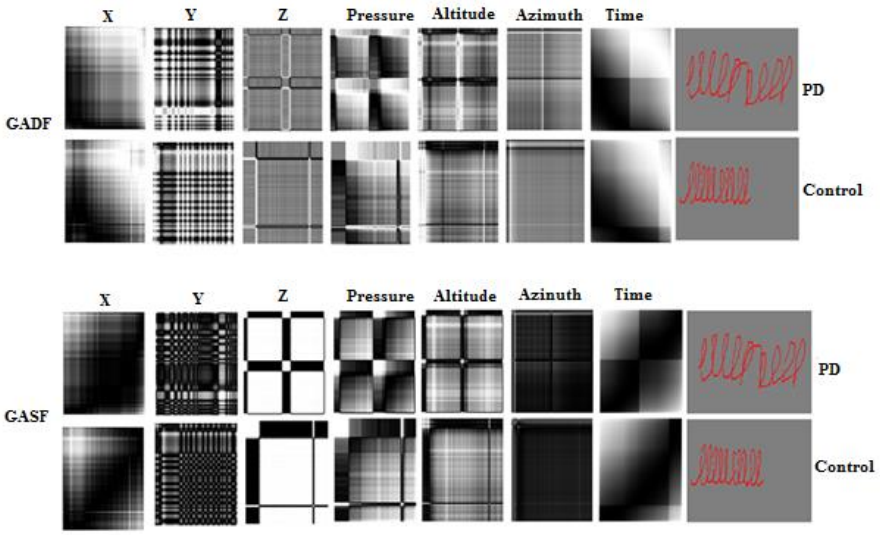



Figure 5.40. GADF and GASF of a PD patient and a HC subject in task1 for each of the 7 signals.

The idea of PAA is to replace each segment by its mean. However, this method can remove by smoothing some important information (such as tremor) that can play a role in PD detection. In this work we have decided to apply a modified gramian angular field approach instead of the original one described above, where each time series is divided into M segments of length 64. Each segment is converted into image using GAF method. The

number of segments M depends on the original time series length; which means that for each subject we will get different number of images. Only GADF images are studied in this work.

#### 5.3.1.2.3 Spectrogram

Handwritten dynamics signals are considered as nonstationary signals that are difficult to separately analyze in time or in frequency domain. Time-frequency representations method is generally used to analyze such signals [Khan et al., 2011]. STFT is the second approach proposed to encode each feature time series as image.

To obtain the time varying frequency spectrum of a non-stationary signal *v[l]*, a sequence of discrete fourier transforms (DFT) of a windowed signal is done. In other word, the STFT is defined as follows:

$$V[k, l] = DFT\{[v[l]w[0], \ldots\ldots\ldots\ldots, v[l+N-1]w[N-1]]\}, k=0, \ldots\ldots, N-1$$

where k is the index denoting frequency, l denotes time, w is the window and N represents the window length. The index k is associated to the frequency $kFs / N$, with $Fs$ is the sampling frequency. $V[k, l]$ represents the magnitude and phase of frequency $F = kFs / N$ at time $t = l \times Ts$ , with $Ts = 1/ Fs$ being the sampling interval [Hadeel, 2016]. $|V[k, l]|$ is a function of two variables (time and frequency indices), its plot is three dimensional and often it is represented as an image by associating the value to an intensity level or a color. This plot is defined by spectrogram. Given a signal of 2 sinusoids of frequencies $\omega_1$ and $\omega_2$ respectively as shown in Figure 5.41-a, where the ideal time frequency plot is shown in Figure 5.41-b. This would yield perfect resolution in frequency, since we see only the exact frequency, and perfect resolution in time, since we see exactly when the frequency changes. However, what we get in reality is the plot shown in Figure 5.42 since it is basically an application of the DFT, it presents the same issues associated to the artifacts due to the window function, such as the main lobe and sidelobes [Hadeel, 2016].

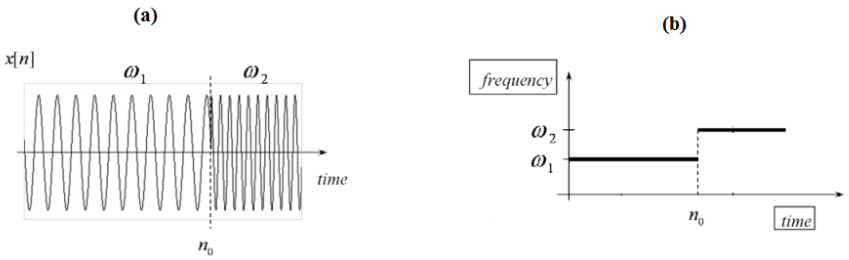



Figure 5.41. Ideal Time-Frequency plot.

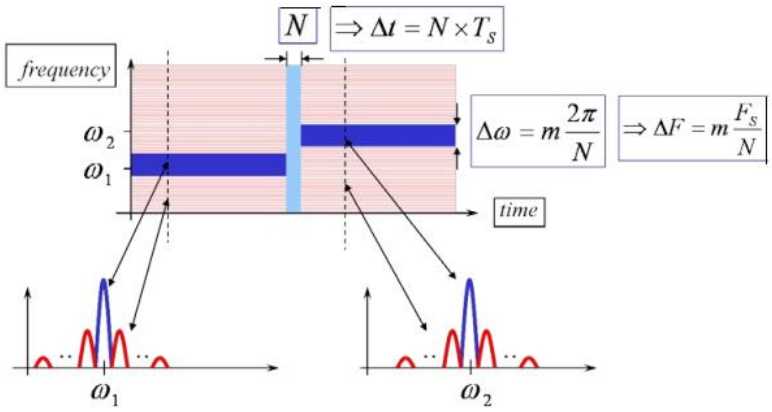



Figure 5.42. Actual Time-Frequency plot [Hadeel, 2016].

The resolutions in frequency and time are defined in equations (5.62) and (5.63) respectively, where m is a coefficient depending on the window used. Wide window size returns good frequency resolution and poor time resolution. It is difficult to maintain good time and frequency resolutions at the same time.

$$\Delta\omega = m\frac{2\pi}{N} \tag{5.62}$$

$$\Delta T = N \times T_s \tag{5.63}$$

The most common used windows are represented in Figure 5.43. According to the frequency spectra, it is clear how some window types have lower sidelobes than others, but higher frequency resolution.

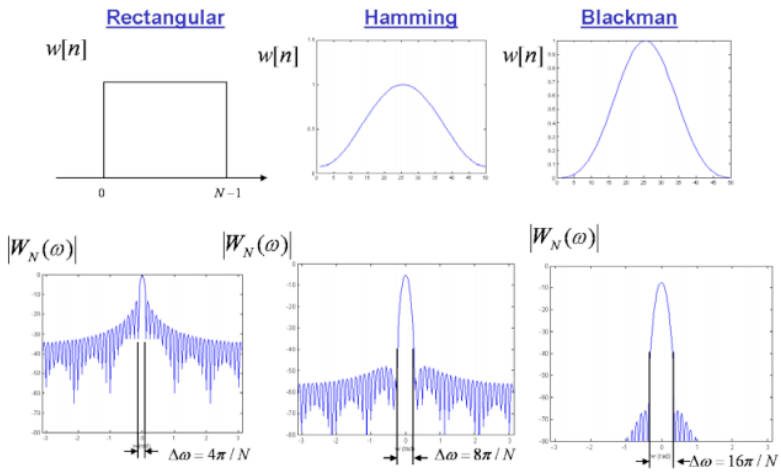



Figure 5.43. The three most common used windows and their frequency spectra [Hadeel, 2016].

In this work, different window sizes and types were studied, and the ones returning the best spectrogram resolution were selected. The Y coordinate signal for a given PD patient is plotted in Figure 5.44. According to this plot, we can say that this signal has a frequency of approximately 0.7 Hz.

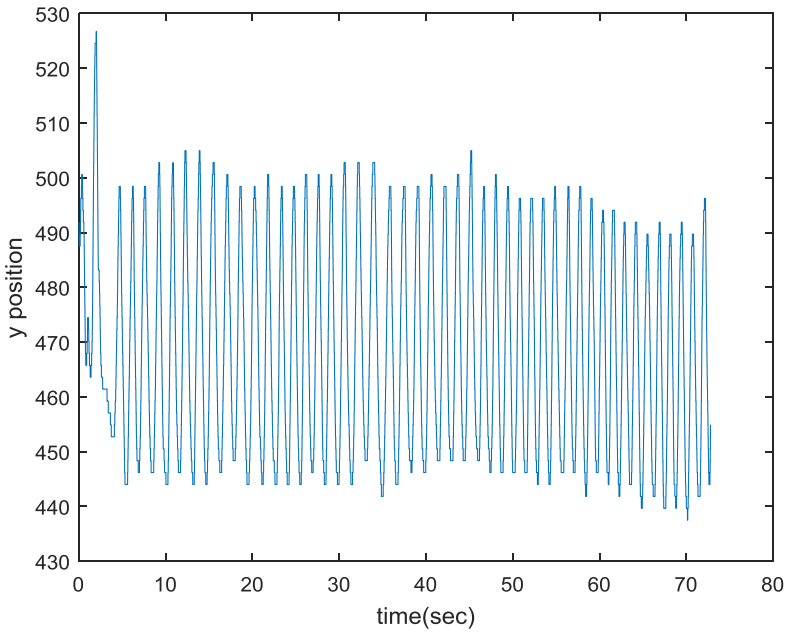



Figure 5.44. Y coordinate plot for a given PD patient concerning Task 1.

The "one-sided" spectrograms of the signal represented in Figure 5.44 for different window types and sizes are plotted in Figure 5.45, where the sampling frequency fs is equal to the digitizing tablet sampling frequency (197 samples/second see chapter 4).

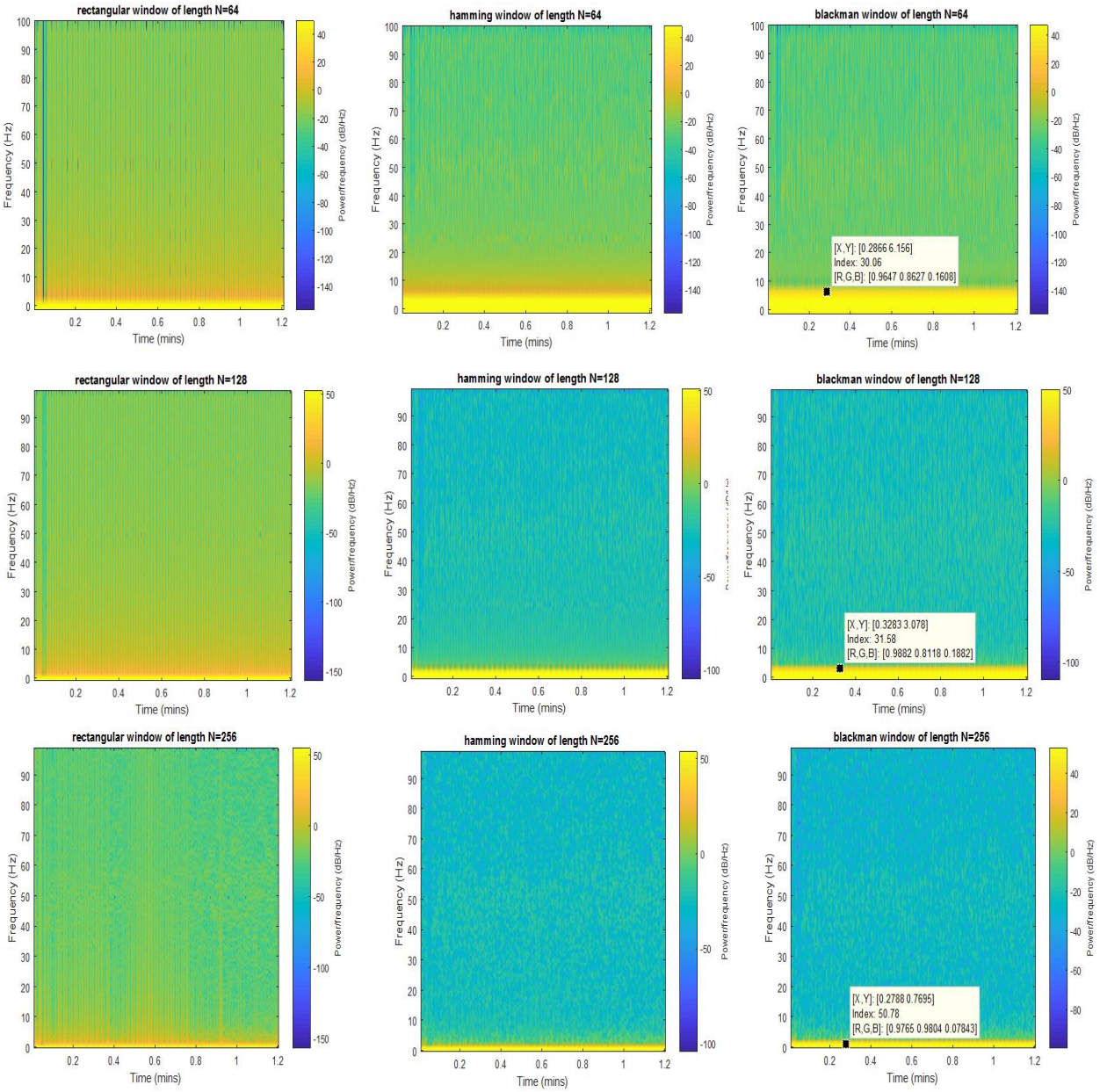



Figure 5.45. The effect of the type and size of the window on spectrogram resolution.

Looking to Figure 5.45, Blackman window of size 256 seems to yield a better solution of the artifacts, and a better frequency resolution. Based on this finding, a Blackman windowing function is used where both the window length and the number of fast fourier transform (FFT) points are set to 256 and the overlapping rate is 50 %.

In line with the idea of treating a spectrogram like an image, the number of frequencies and the number of time bins in the spectrogram representation refer to the height and the width of the output image in pixels. In addition, the numerical "brightness" value of each pixel of this two dimensional image is then equal to the output value of the spectrogram at the particular time and frequency corresponding to that pixel. These values should be converted to a logarithmic scale (decibels) to get a better view into the most important parts of the spectrogram, then normalized to [0, 1] generating a grayscale image shown in Figure 5.46. The width of the image depends on the length of the signal. To keep the number of input feature maps identical, the area of spectrogram should be the same for all subjects. While we believe that the variation over the time axis is non-linear, we assume here that this variation is linear and we apply Lanczos technique to resize the spectrogram images to 64×64 pixels resolution. The spectrograms of a PD patient and a HC subject concerning task1 for each of the 7 signals are shown in Figure 5.47.


Figure 5.46. Spectrograms of the same signal: (a) non logarithmic scale and (b) logarithmic scale.

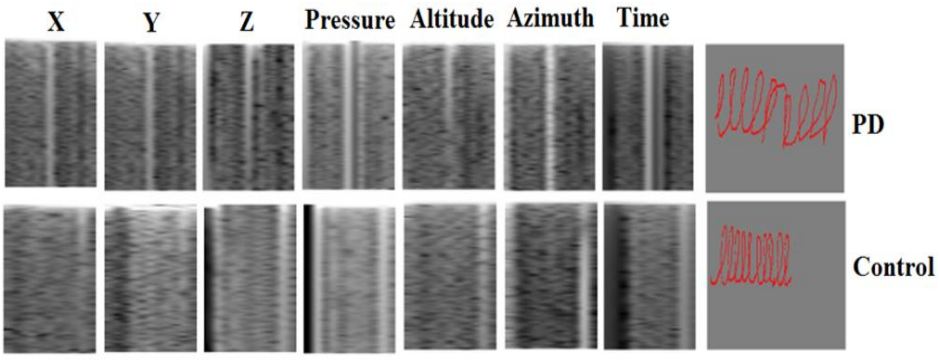



Figure 5.47. The normalized spectrograms of a PD patient and a HC subject in task1 for each of the 7 signals.

### 5.3.1.3 Deep learning architectures

In the following the 2D CNN architecture as well as the 1D CNN-BLSTM used for detection from direct time series will be described.

#### 5.3.1.3.1 2D CNN architecture

A 2D CNN consists of a neural network that extracts features of the input image and another neural network that classifies the feature image as shown in Figure 5.48. The feature extraction neural network consists of piles of convolutional and pooling layer pairs. The convolution layer converts the image using the convolution operation. It can be thought of as a collection of digital filters. The pooling layer combines the neighboring pixels into a single pixel. Therefore, the pooling layer reduces the dimension of the image [Kim, 2017].

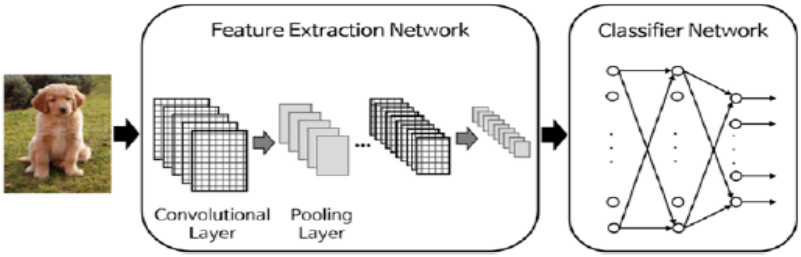



Figure 5.48. Typical architecture of 2D CNN [Kim, 2017].

The convolution layer generates new images called *feature maps*. The feature map accentuates the unique features of the original image. The convolution layer operates in a very different way compared to the other neural network layers. This layer does not employ connection weights and a weighted sum. Instead, it contains filters that convert images. The process of the inputting the image through the convolution filters yields the feature map. The number of generated feature maps by the convolution layer is equal to the number of filters. The values of the filter matrix are determined through the training process. Therefore, these values are continuously trained throughout the training process. This aspect is similar to the updating process of the connection weights of the ordinary neural network (described in section 5.2.3.1). The outputs if the convolution layer passes through an activation function. The pooling layer reduces the size of the image, as it combines neighboring pixels of a

certain area of the image into a single representative value. The representative value is usually set as the mean or maximum of the selected pixels.

The 2D CNN model implemented in this work is summarized in Figure 5.49. The feature extractor layers consist of two convolution layers, each followed by ReLU activation function, and two pooling layers. Starting with a 64×64 pixel image with one channel (Grayscale), all the convolutional layers employ kernels of size 5×5 with stride of 1 pixel, and the maxpooling operations are applied on regions of size 2x2, with stride 2. The convolutional layers convert the input image into 64 feature maps of size 16×16. It can be noted that different convolution (ReLU unit) and pooling operations are applied on each image in order to learn different features independently [Taleb et al., 2019-a].

After using convolution layers to extract the spatial features of an image, a fully connected layer is applied for the final classification. The output of the convolution layers is flattened, and then a hidden layer with 23 nodes and ReLU activation function is used before performing the final classification. The output layer is composed of 2 nodes (binary classification PD or HC) with softmax activation function. The initial weights are drawn from uniform distributions within [-limit, limit] where limit is sqrt(6/(fan_in+fan_out)), fan_in is the number of input units in the weight tensor and fan_out is the number of output units in the weight tensor (equivalent to gaussian distributions with 0 mean and STD of sqrt(2/(fan_in+fan_out))). This initialization is also known by Xavier normalized initialization, where its role is to preserve constant variance in both forward and backward passes [Glorot and Bengio, 2010]. The initial biases are zeros initialized. The number of input images k is a hyper-parameter varying between one and seven (the number of signals). This 2D CNN model can be used for classification from a single image including k measurements (time-series based), or classification from k measurements, where each measurement is encoded into an image (modified GAF and spectrogram). The number of hidden nodes is chosen based on the empirical rule described in equation (5.64):

$$N_h = \frac{N_s \times N_i}{a \times (N_i + N_o + 1)} - \frac{N_o}{N_i + N_o + 1} \tag{5.64}$$

where $N_{s,}$ $N_{i,}$ and $N_o$ refer to training size, input and output nodes respectively, and the scaling factor a is set to 5 here.

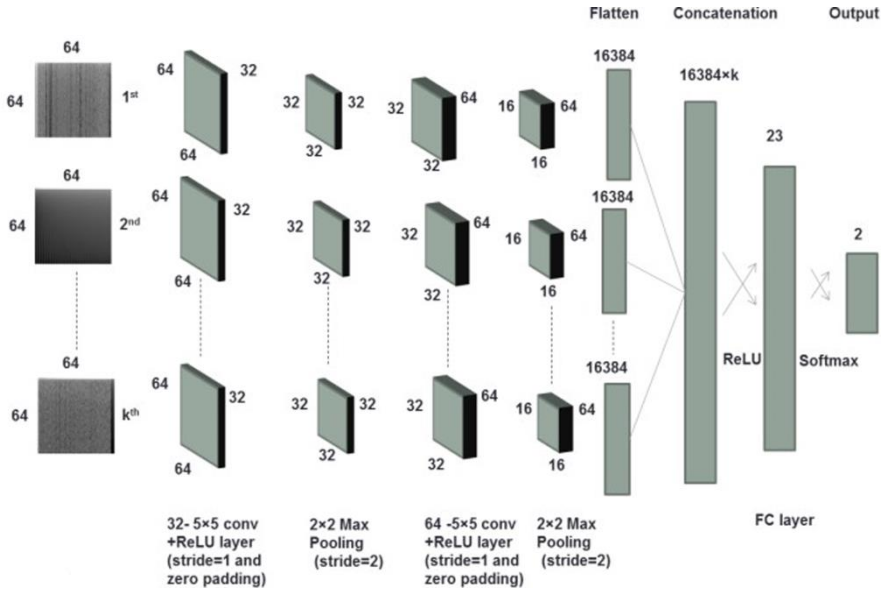



Figure 5.49. Single-task 2D CNN architecture. This architecture takes as input k two dimensional representations of 1D time series.

#### 5.3.1.3.2 Slicing and combination approach

As mentioned before, each subject has M GAF images per task. When doing training using the 2D CNN model represented in Figure 5.49, all the training window slices images are considered independent training instances. Window slicing is also applied when predicting the label of a testing time series. The trained 2D CNN model predicts the label of each of the window slices. No window slices referring to the same participant exist in training and test. To make the final prediction for each subject in the test set, 3 different operations are used:

- Using a majority vote among all these slices, where majority voting consists on choosing the class label which has the maximum number of vote by each classifier.

- Getting the product of the M probability vectors outputs of the 2D CNN models. The final class label will be the one with the highest probability.

- Using bidirectional-LSTM (BLSTM) summarized in Figure 5.50; where the M probability vectors outputs of the 2D CNN models are considered as a Multivariate sequence (which have two or more variables observed at each time) of length M, and are used as input to a dynamic BLSTM to decide the final prediction.

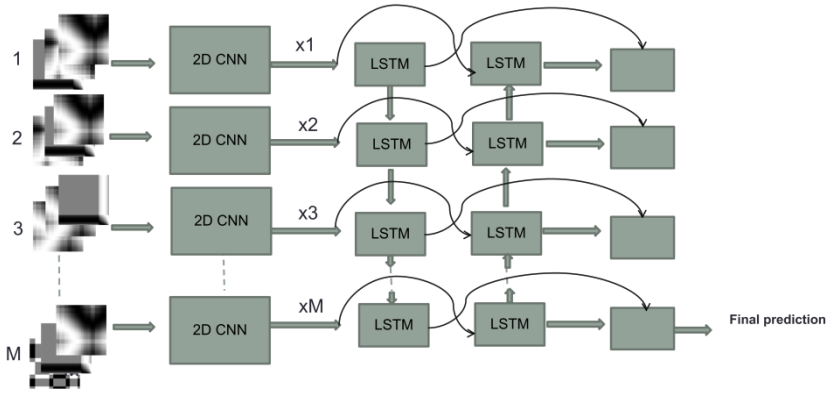



Figure 5.50. Slicing combination approach. k time series are cut into M slices. In each slice, k input images are built. The decisions (probability distribution for each class) provided by each slice images are input as a sequence of length M to a BLSTM.

Recurrent neural network (RNN) is a neural network that has recurrent connections, so it can store information inside the network and also accept sequences of different lengths as input. Generally, at any time step of a sequence, RNNs compute some memory based on its computations thus far; i.e., prior memory and the current input. This computed memory, stored by the hidden states, is used to make predictions for the current time step and is passed on to the next step as an input. The basic architecture of a recurrent neural network is illustrated in Figure 5.51, where $x_t$, $h_t$, and $o_t$ represent the input, hidden states, and output respectively at time step t. $W_{hh}$ represents the weights matrix from the memory states $h_t$ at time t to the memory states $h_{t+1}$at time (t +1). $W_{xh}$ represents the weight matrix from the input $x_t$ to the hidden states $h_t$, whereas $W_{ho}$ represents the weight matrix from the memory states $h_t$ to $o_t$ [Pattanayak, 2017].

RNNs fail to store information during long time intervals because the gradients in instances of long sequences have a high chance of either going to zero or going to infinity very quickly. When the gradient of the loss function appropriates zero, the network will be hard to be trained. This problem is defined by vanishing gradient. The LSTM is a special version of the RNN, which can store information during longer time interval. The architecture

of LSTMs is quite a bit different than that of traditional RNNs: the hidden cells are replaced by LSTM cells (long-short term memory) as represented in Figure 5.52. The new element in LSTMs is the introduction of the cell state Ct, which is regulated by three gates. The gates are composed of sigmoid functions so that they output values between 0 and 1. At sequence step t the input $x_t$ and the previous step's hidden states $h_{t-1}$ decide what information to forget from cell state $C_{t-1}$ through the forget gate layer. The forget gate looks at both the input and the hidden state and assign a number of between 0 and 1 for each element in the cell state vector $C_{t-1}$ using sigmoid activation function as shown in Figure 5.52. Next, like the forget gate, the input gate role is to decide which cell units should be updated with new information. Last, the output gate determines which cell state to output [Pattanayak, 2017].

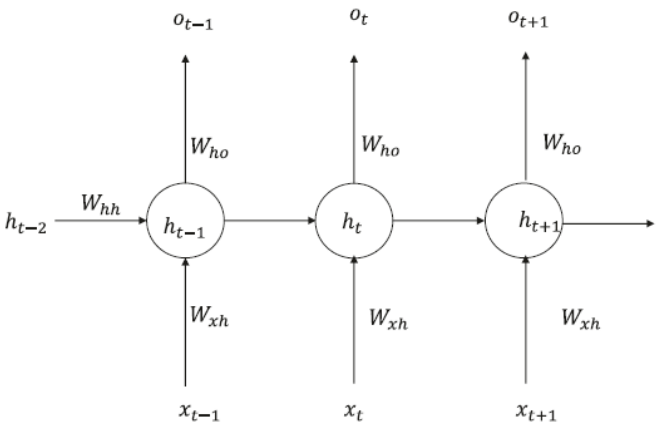



Figure 5.51. Structure of RNN [Pattanayak, 2017].

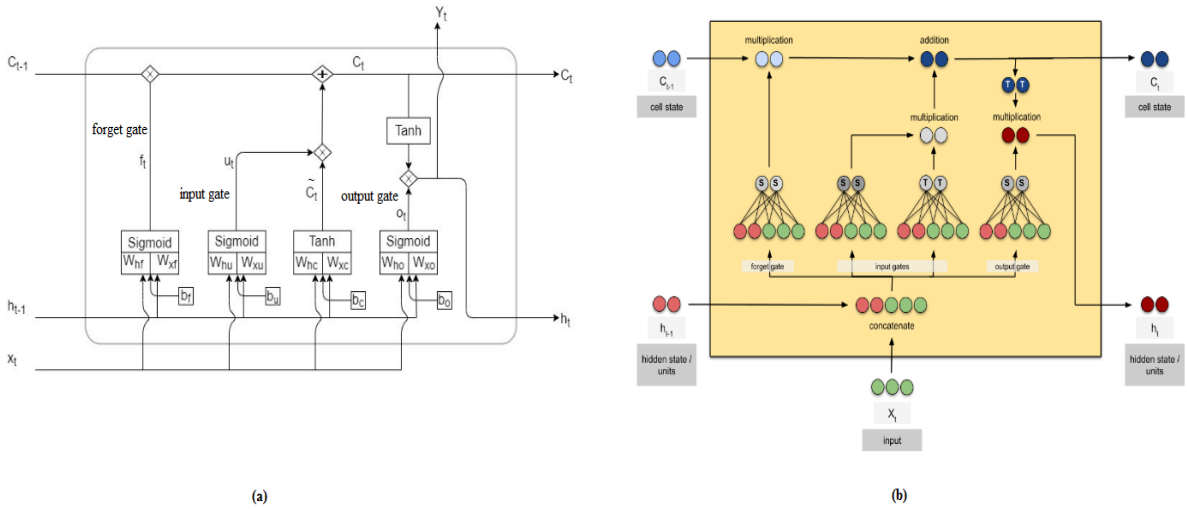



Figure 5.52. LSTM cell architecture: (a) block architecture [Singh, 2020] and (b) animated architecture [Karim, 2020].

The LSTM can throw away information that it thinks is not useful via the forget gate, and only the relevant information is exposed to the rest of the network via the output gate. This will reduce exploding and vanishing gradient problems.

BLSTMs are a special type of LSTM that makes use of both the past and future states to predict the output label at the current state. A BLSTM combines two LSTMs, one of which runs forward from left to right and the other of which runs backward from right to left as shown in Figure 5.53.

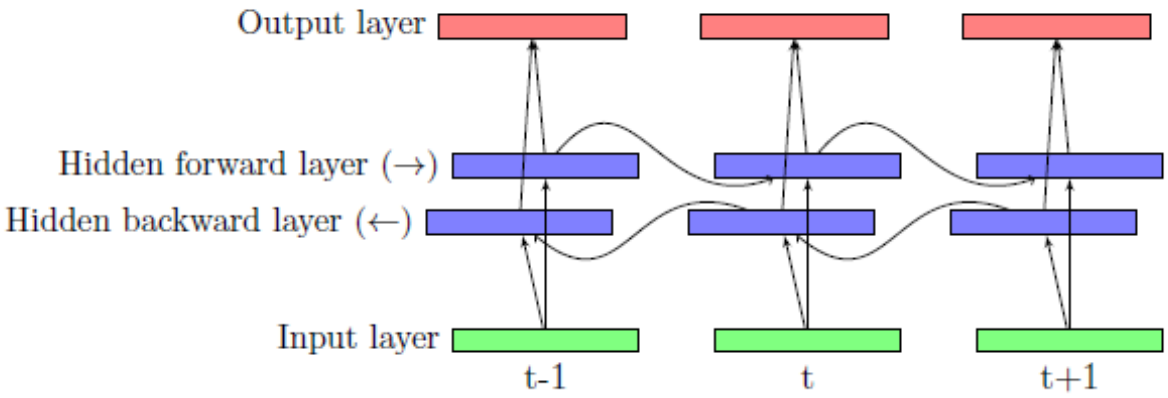



Figure 5.53. Bidirectional LSTM.

In this work, the basic BLSTM model is implemented to combine the results obtained by the M classifiers where there would be no output at time step, but only one output at the last time step. Only the final sequence step would contribute to the cost function; there would be no cost function associated with the intermediate sequence steps. The number of hidden nodes is selected in a way of having more data than the parameters. An empirical rule defined in equation (5.65) is applied for hidden nodes selection.

$$2\times a\times[g\times[N_h\,(N_h\ +N_i)+N_h\,]]=N_s\ \times N_i \tag{5.65}$$

where g refers to the number of feed forward neural networks (FFNNs) in the LSTM cell, $N_h$ is the hidden nodes, $N_i$ is the size of input (here it is equal to 2), $N_s$ is the training size, and a is the scaling factor. For LSTM, $g$ is equal to 4 (see Figure 5.52-b), and the scaling factor $a$ is set to 5 in this work.

In the LSTM cell there are 8 sets of weight parameters (4 associated with the hidden state and 4 associated with the input vector), and 4 different bias parameters as shown in Figure 5.52-a. For the BLSTM model in Figure 5.50, we initialize the 16 weights and the 8 biases as tensors full of zeros since we are going to learn weights and biases, so it does not matter very much what they initially are. The output of the BLSTM is multiplied by a weight matrix, and added to a bias vector to obtain the final prediction value, where both the weight and the bias are drawn from random normal distributions.

#### 5.3.1.3.3 1D CNN-BLSTM architecture

The extracted signals represented in Figure 5.35-a and c are noisy, therefore, the analysis of such kind of data without feature extraction can be challenging since we need a noise-robust approach to perform a correct classification. For the sake of comparison, we also suggest to apply a 1D CNN-BLSTM directly on the time series without visualizing them as images, where the architecture involves using 1D CNN layers for feature extraction on input data combined with BLSTMs to support sequence prediction. The LSTM reduce the vanishing gradient problem by very large margin but still can face such problems when the sequences are so long. For this reason, the CNN layers are not only used for feature extraction here, but also to reduce the length of the sequence. Instead of converting the time series into images, the entire raw time series are used here as input to the model. The convolutional layers are constructed using one-dimensional kernels that move through the sequence. A 1D CNN-BLSTM model on multivariate time series is represented in Figure 5.54. The CNN layers consist of two convolution layers, each followed by ReLU activation function, and two pooling layers. All the convolutional layers employ 1D kernels of size 1×2 with stride of 1 pixel, and the maxpooling operations are applied on regions of size 1x2, with stride 2.The output of the 1D CNN is a sequence of length n/4 of vectors of size 32; where n represents the time series length. This sequence is then used as input to a dynamic BLSTM, since sequences are of variable length. The number of time series k is a hyper-parameter, and it varies between 1 and 7. For the BLSTM, the number of hidden nodes also follows the empirical rule described in (5.65), where $N_i$ is equal to 32 in this case. The initial weights and biases are chosen similarly to the ones mentioned in the previous section.

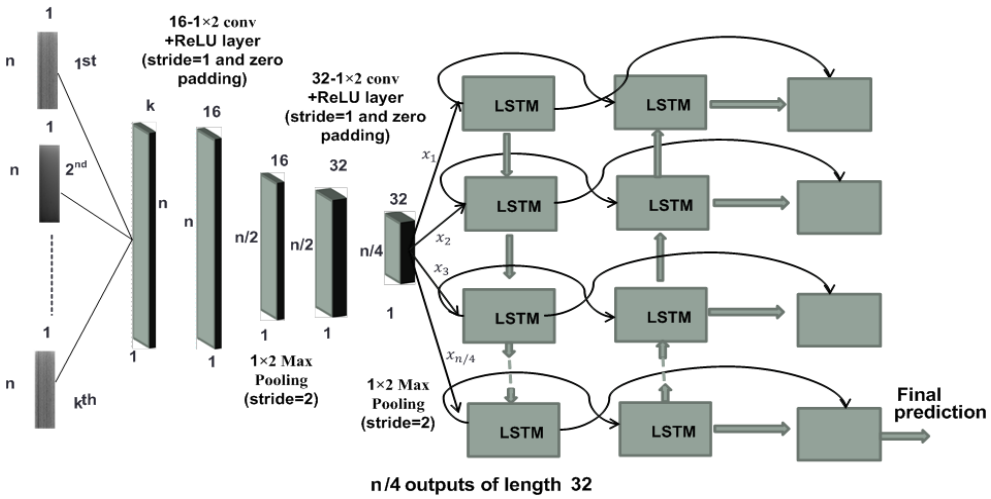



Figure 5.54. Single-task 1D CNN-BLSTM architecture on multivariate time series. The output of the 1D CNN is fed to a BLSTM as a sequence.

#### 5.3.1.3.4 All tasks combination approach and hyper-parameter k selection

To enhance recognition, seven single-task systems can be combined into an all-task system. The combination approach considered in this work for obtaining the overall accuracy is the majority voting. In order to get the best time series features combination (setting up the hyper-parameter k) a suboptimal incremental approach has been used. The feature providing the highest overall validation accuracy (the accuracy obtained by combining the outputs of the seven tasks using majority voting) is first selected. Then, features are added incrementally by selecting, at every iteration, the one yielding the highest overall validation accuracy. The iterations stop when no more increase in performance is observed.

## 5.3.2 Experiments and numerical results

The performance was figured out in term of accuracy, sensitivity, and specificity defined in section 5.1. Due to the small data we worked on, the set with 42 subjects is divided into 3 folds, with the 66.66/33.33 % (training/validation) proportion using stratified sampling method. Sequentially, one fold is validated using the classifier trained on the remaining 2 folds. The total accuracy is obtained by calculating the mean of all the folds accuracies. The same criteria applied in section 5.1, is also applied here; where the validation set is considered here as test set. The models described above are tested, where cross-entropy cost function and Adam optimizer are applied here [Kim, 2017]. Adam is a replacement optimization algorithm for SGD and can handle sparse gradients on noisy problems [Kingma and Ba, 2015]. To

avoid overfitting 2 techniques are applied: the first one is defined by early stopping [Prechelt, 1998] where we evaluate the loss rate on the validation set, and if it does not improve for 20 epochs, network training is stopped. The second approach employed to prevent the network from overfitting is dropout technique (with 0.4 dropout rate) applied, where a dropout layer is added right after the hidden layer [Srivastava et al., 2014]. To get faster training convergence, mini-batch training combined with shuffling after each epoch is applied [Bengio, 2012]. Since dropout was applied, we decided not to add batch normalization because it fulfills some of the same goals as dropout [Loffe and Szegedy, 2015]. The performance measures in Table 5.17 represent the average of 3 runs considering the majority voting combination approach defined in section 5.3.1.3.4, where the best results are in bold. The number of parameters learned in each model per fold and per task is obtained based on equations (5.66), (5.67), (5.68) and (5.69) and presented in Table 5.16; where k refers to the number of signals. The total number of learned parameters for each model is presented in Table 5.17.

| | | | |
|---|---|---|---|
| MLP with one Hidden layer | $[i \times h + h \times o] + h + o$ | *i: input size*<br>*h: hidden layer nodes*<br>*o: output size* | (5.66) |
| 2D CNN one layer | $\mathbf{[}c \times \mathbf{(}f \times f\mathbf{)} \times p\mathbf{]} + p$ | *c: input channels*<br>*f: filter size*<br>*p: number of filters* | (5.67) |
| 1D CNN one layer | $\mathbf{[}c \times \mathbf{(}1 \times f\mathbf{)} \times p\mathbf{]} + p$ | *c: input channels*<br>*f: filter size*<br>*p: number of filters* | (5.68) |
| BLSTM | $2g \times \mathbf{[}h\mathbf{(}h+i\mathbf{)} + h\mathbf{]}$ | *g: number of FFNNs in a unit (g=4 for LSTM)*<br>*h: size of hidden units*<br>*i: input size* | (5.69) |

Table 5.16. The total number of parameters trained in each model per fold and per task.

| Model | Layer | Characteristics | Nb. of parameters per layer | Nb. Of Parameters |
|---|---|---|---|---|
| 2D CNN (time series based image) | Conv1 | *c: 1, f: 5×5, p: 32* | 832 | 428,999 |
| | Conv2 | *c: 32, f: 5×5, p: 64* | 51,264 | |
| | MLP | *i: 16384, h: 23, o: 2* | 376,903 | |
| 2D CNN (GADF with segmentation and BLSTM) | k-Conv1 | *c: 1, f: 5×5, p: 32* | 832×*k* | 16,829,312×*k+3,416* |
| | k-Conv2 | *c: 32, f: 5×5, p: 64* | 51,264×*k* | |
| | MLP | *i: 16384×k, h: 1024, o: 2* | 16,777,216×*k+3,074* | |
| | BLSTM | *g: 4, h: 5, i: 2* | 320 | |
| | | *Output weight (2×h,o)*<br>*Output bias (o)* | 22 | |
| 2D CNN (GADF with segmentation and Majority votes) | k-Conv1 | *c: 1, f: 5×5, p: 32* | 832×*k* | 16,829,312×*k+3,074* |
| | k-Conv2 | *c: 32, f: 5×5, p: 64* | 51,264×*k* | |
| | MLP | *i: 16384×k, h: 1024, o: 2* | 16,777,216×*k+3,074* | |
| 2D CNN (GADF with segmentation and Probability products) | k-Conv1 | *c: 1, f: 5×5, p: 32* | 832×*k* | 16,829,312×*k+3,074* |
| | k-Conv2 | *c: 32, f: 5×5, p: 64* | 51,264×*k* | |
| | MLP | *i: 16384×k, h: 1024, o: 2* | 16,777,216×*k+3,074* | |
| 2D CNN (with spectrogram) | k-Conv1 | *c: 1, f: 5×5, p: 32* | 832×*k* | 428,928×*k+71* |
| | k-Conv2 | *c: 32, f: 5×5, p: 64* | 51,264×*k* | |
| | MLP | *i: 16384×k, h: 23, o: 2* | 376,832×*k+71* | |
| 1D CNN-BLSTM | Conv1 | *c: k, f: 1×2, p: 16* | 32×*k+16* | *32×k+2,614* |
| | Conv2 | *c: 16, f: 1×2, p: 32* | 1,056 | |
| | BLSTM | *g: 4, h: 5, i: 32* | 1,520 | |
| | | *Output weight (2×h,o)*<br>*Output bias (o)* | 22 | |

According to these results, we can see that both the 2D CNN model with spectrogram images as inputs, and the 1D CNN-BLSTM model with raw time series as input and lower number of parameters return the best PD detection accuracy; where best features combination was chosen in a way of returning the highest 3-folds CV overall validation accuracy. The best feature sets selected for both models do not include the time stamp feature. This finding is obvious since the multivariate time series are generated with a fixed synchronized sampling along all dimensions.

Based on these features, it is important to compare and rank the performance of each task separately to select tasks with the highest features relevance. Task-wise system accuracies for different model are represented in Table 5.18; where D1, D2, D3, D4, D5, and D6 models refer to the ones used in entries 1, 2, 3, 4, 5, and 6 in Table 5.17. It can be observed from that “all tasks” reports highest accuracies across all the 6 models. Additionally, we can observe that Task 2 (triangular wave), and Task 3 (rectangular wave) report highest accuracies across all the 6 models. The same conclusion is found in our previous work [Taleb et al., 2017] described in section 5.1.

Two main findings are found from the direct comparison of our 1D CNN-BLSTM and 2D CNN models. The first one is that our 2D CNN model fed with 2D spectrograms, and the 1D CNN-BLSTM fed with time series (with lowest complexity) perform better than the 2D CNN model fed with time series-based images, where this 2D representation is inspired by Pereira et al. [Pereira et al., 2018]. The second finding is that both 2D CNN with time series-based images and 2D CNN with GADF images (and BLSTM applied as window slices combination approach) perform the same. These findings can be explained by the fact that the signals generated by the pen are multi- frequency, non-periodic, and arbitrary [Alsheikh et al., 2016]. In addition, the BLSTM-based model takes advantage of learning the temporal feature activation dynamics, which the CNN model is not capable to model [Ordóñez et al., 2016].

Table 5.17. 3-fold CV performance measures of all-task system considering the majority voting combination approach.

| Exp. | Model | Data Input | Window slicing combination approach | Perf. (%) | Best Features Combination | # of learned parameters |
|---|---|---|---|---|---|---|
| 1. | 2D CNN | Time series-based images | | Acc :80.95<br>Sens:85.71<br>Spec:76.19 | Pressure | 9,008,979 |
| 2. | 2D CNN | GADF images using window slicing | BLSTM | Acc :80.95<br>Sens:71.43<br>Spec:90.48 | X+Y+Z+Pressure+Altitude+ Azimuth | 2,120,565,048 |
| 3. | 2D CNN | GADF images using window slicing | Majority Votes | Acc :78.57<br>Sens:80.95<br>Spec:76.19 | X+Y+Z+Pressure | 1,413,726,762 |
| 4. | 2D CNN | GADF images using window slicing | Probability vectors product | Acc: 73.81<br>Sens:61.90<br>Spec:85.71 | Pressure | 353,480,106 |
| 5. | 2D CNN | Spectrogram images | | **Acc :83.33**<br>Sens:85.71<br>Spec:80.95 | X+Y+Z+Pressure+Altitude | 45,038,931 |
| 6. | 1D CNN-BLSTM | Raw time series | | **Acc :83.33**<br>Sens:71.43<br>Spec:95.24 | X+Y+Z+Pressure+Altitude+ Azimuth | 58,926 |

Table 5.18. Task-wise system and "All-tasks" system accuracies (in %) for various models. D1:2D CNN/time series-based images, D2: 2D CNN/GADF images/BLSTM, D3: 2D CNN /GADF images/ majority votes, D4: 2D CNN /GADF images/ probability vectors product, D5: 2D CNN/ spectrogram images, and D6: 1D CNN-BLSTM/raw time series.

| Task | D1 | D2 | D3 | D4 | D5 | D6 |
|---|---|---|---|---|---|---|
| Repetitive cursive letter 'l' | **71.43** | 64.29 | 61.90 | 47.62 | 54.76 | 57.14 |
| Triangular wave | 69.05 | 69.05 | 64.29 | 59.52 | 50.00 | **76.19** |
| Rectangular wave | 33.33 | **76.19** | 73.81 | 57.14 | 64.29 | 73.81 |
| Repetitive "Monday" | 61.90 | 54.76 | 61.90 | 57.14 | 61.90 | **64.29** |
| Repetitive "Tuesday" | 50.00 | 59.52 | **66.67** | **66.67** | 64.29 | 45.24 |
| Repetitive "Name" | 69.05 | 54.76 | 35.71 | **73.81** | 64.29 | 47.62 |
| Repetitive "Family Name" | 66.67 | 61.90 | 57.14 | 66.67 | 71.43 | **73.81** |
| All tasks | 80.95 | 80.95 | 78.57 | 73.81 | **83.33** | **83.33** |

### 5.3.3 Conclusions

One main contribution of this thesis is to employ deep learning approach to aid in PD early detection. Two based learning models for end to-end time series classification are proposed: the 2D CNN and the 1D CNN-BLSTM. For the 2D CNN model, two different frameworks were proposed to encode time series into images: gramian angular field images, and spectrogram images. These two frameworks are compared with the one proposed by Pereira et al. [Pereira et al., 2018]: the direct encoding of time series into images. The advantage of using spectrogram images consists in computing local short term information that exists in the non-stationary online handwriting signals before normalization, while the other two approaches normalize the time series into a fixed dimension image without extracting local information. For the 1D CNN-BLSTM model, the raw time series are directly used with no need to convert them into images. This approach has been experimented to validate the importance of considering the local information before integrating on the time scale. We have demonstrated the importance of both: a deep architecture based on the combination of 1D CNN and BLSTM recurrent layers, and a 2D CNN model with spectrograms as input in PD detection. Our results clearly show that when explicitly considering the local short term information on the time axis of the non-stationary online handwriting signals the deep learning models provides the best performance.

Compared with these deep learning models, the SVM model trained on pre-engineering features and described in section 5.1 shows better accuracy because the SVM can be trained with a small training data, in contrast to deep learning models which require a large number of training samples to work well. This means that PD classification using deep learning is a challenging task due to the limited data availability.

This has motivated us to investigate transfer learning and data augmentation approaches based on these models to perform PD detection on large-scale data that will be described in the following section.

## 5.4 Improving deep learning Parkinson's disease early detection through data augmentation

Deep learning have shown excellent performance on classification problems where large datasets are available. However, it is challenging to apply deep learning to problems where only small datasets are available like medical data [Um et al., 2017]. Training an adequately sized neural network with a small amount of data can cause the network to memorize all training examples, in turn leading to poor performance on a holdout dataset [Brownlee, 2019]. This phenomenon, also known as overfitting, can be solved using different techniques such as collecting more labeled data (which is in our case hard to obtain), using transfer learning method, or using data augmentation. Transfer learning is a machine learning technique where a model trained on one task (a source domain) is re-purposed on a second related task [Sadouk, 2019]. Data augmentation is the process of generating artificial data from the original ones. A developed description of transfer learning and data augmentation will be presented in the following sections. Such approaches have been applied in different domains including handwritten recognition of manuscripts [Chammas et al., 2018]. In our case here, the key challenge is to maintain the correct label. It is important to find the proper data augmentation method that will preserve the correct label. In this work, we are working with pen-based dynamic signals. However, unlike in image recognition problems, data augmentation techniques have not been completely investigated for the time series classification task [Fawaz et al., 2018]. Transfer learning and data augmentation techniques for time series are proposed here to overcome the overfitting problem and to increase the recognition accuracy and robustness of the best two models found in section 5.3 and summarized in Figure 5.55 and Figure 5.56: the 2D CNN model with the combination of spectrogram images referring to X, Y, Z, pressure, and altitude features as input and the 1D CNN-BLSTM model with the combination of X, Y, Z, pressure, altitude and azimuth raw time series as input.

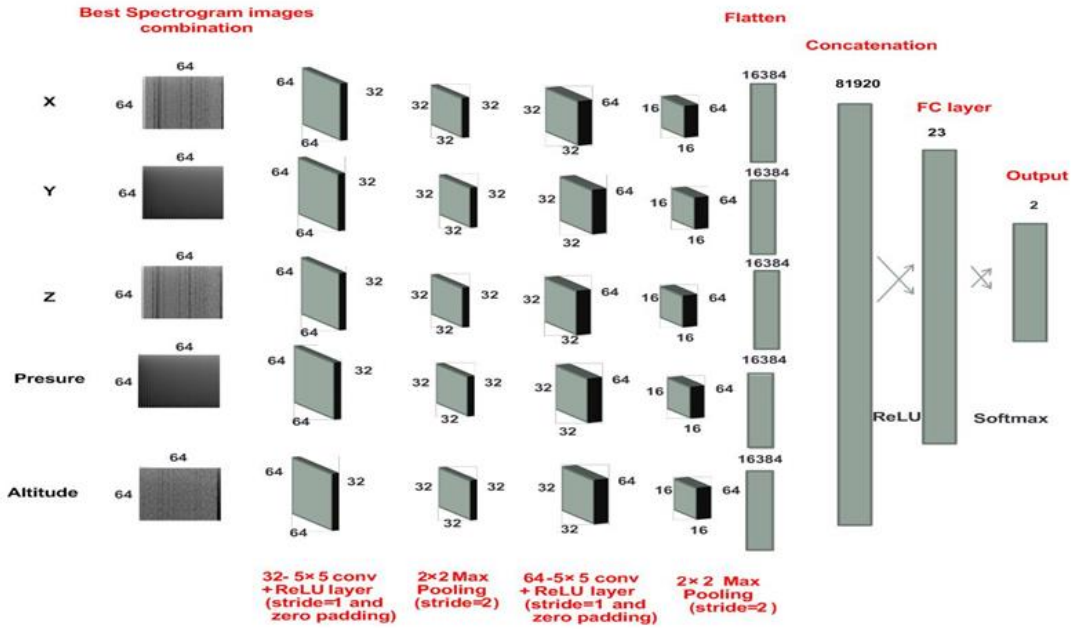



Figure 5.55. One of the best two models found in our previous work: The 2D CNN model with the combination of spectrogram images referring to X, Y, Z, pressure, and altitude features as input.

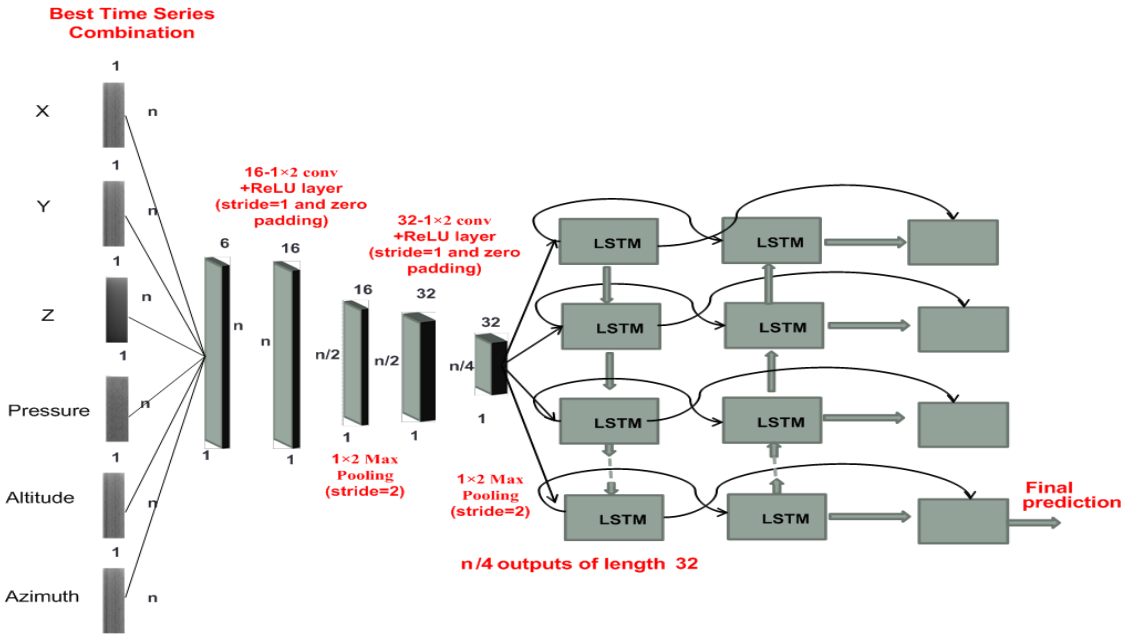



Figure 5.56. The second best model found in our previous work: The 1D CNN-BLSTM model with the combination of X, Y, Z, pressure, altitude and azimuth raw time series as input.

## 5.4.1 Transfer Learning

The first approach used to solve the overfitting problem is the transfer learning method. Some works have shown that transfer learning can be used efficiently with CNNs [Mormont et al., 2018]. For this reason, in this work we apply transfer learning process on the

2D CNN model represented in Figure 5.55. By using pre-trained models which have been previously trained on large datasets, we can directly use the weights and architecture obtained and applies the learning on our problem statement [Marcelino, 2019]. PaHaW and HandPD are two available handwriting datasets slightly larger than HandPDMultiMC. As mentioned in chapter 3, HandPD includes spiral and meander tasks collected from 92 subjects (74 PD and 18 HC) and where time series signals were captured by a BiSP. In comparison, handwriting tasks and time series signals in PaHaW are the closest to HandPDMultiMC dataset (loops and time series). Thus, PaHaW was selected here to pre-train the 2D CNN model. The PaHaW handwriting template is summarized in Figure 5.57. As mentioned in chapter 3, PaHaW dataset consists of handwriting samples collected from 37 PD patients and 38 HC subjects that are aged and gender matched. PaHaW includes eight different handwriting tasks (spiral drawing and words written in Czech) as shown in Figure 5.57. Handwritten dynamic signals such as X and Y coordinates, pen pressure, altitude, azimuth and time stamp were captured by Wacom Intuos 4 digitizing tablet. By comparing both PaHaW and HandPDMultiMC datasets, we can say that handwriting tasks in both datasets are somehow similar; they both contain loops and repetitions. The differences between the 2 sets are that the Z coordinate feature is missing in PaHaW, and the number of tasks is 8 instead of 7. To match the two datasets, only tasks with loops and repetitive words in PaHaW are studied (the first 7 tasks), and the Z coordinate feature in HandPDMultiMC is eliminated. The whole PaHaW dataset is used for pre- training in this work.

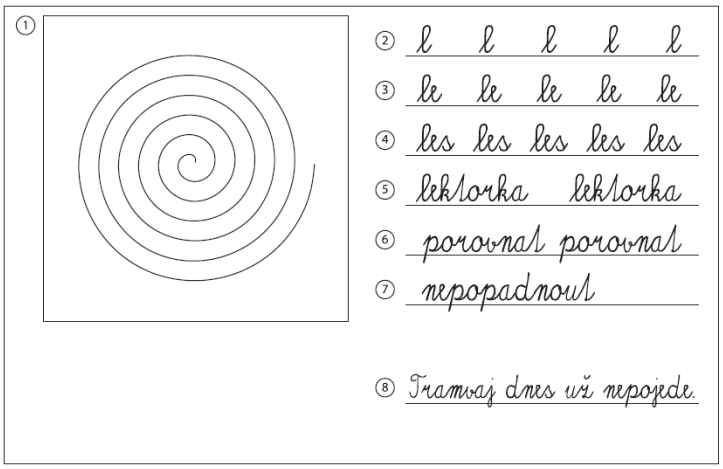



Figure 5.57. PaHaW handwriting template sample [Drotar et al., 2016].

Different transfer learning strategies are studied and compared to validate the gains of transfer learning over training our 2D CNN model from scratch. These strategies are summarized in Figure 5.58. The first transfer learning strategy freezes all the PaHaW-trained model layers and a new softmax classifier is trained using the training images of HandPDMultiMC dataset since the softmax contains relatively few parameters, it can be trained from a relatively small number of examples [Zeiler and Fergus, 2013]. The second strategy freezes only a part of the PaHaW-trained model. As we know, lower layers of the convolutional base refer to general features, while higher layers refer to specific features [Marcelino, 2019]. We studied 2 partial freeze strategies; where the first one freezes the whole convolutional base of the PaHaW-trained model and the part closer to the classifier is retrained using the training images of HandPDMultiMC dataset, and the second one freezes only the first layers of the convolutional base and retrains the rest of the model. Last, we consider full freezing of all layers of the PaHaW-trained 2D CNN model.

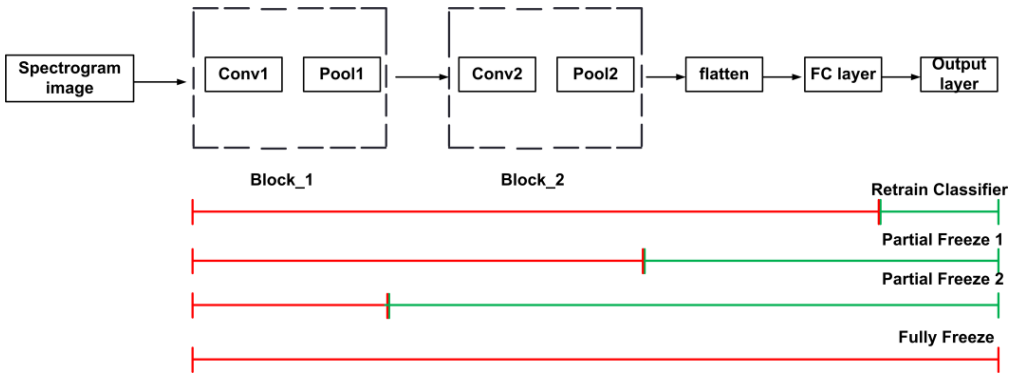



Figure 5.58. Different transfer learning strategies are studied. Green indicates that blocks are retrained, and red indicates that blocks are frozen.

## 5.4.2 Data augmentation applied to time series

Another way to reduce the errors of the classifier that are due to variance[3] is by enlarge the training data by generating synthetic (or artificial) examples. Augmenting training data mostly adds bias[4] to the classifier, and adding bias reduces variance. Time series that we are working with are collected from Wacom tablet's sensors. Augmenting the dataset by generating artificial examples changes the reality, so it may improve the performance, or can

[3] Variance is the amount that the estimate of the target function will change if different training data was used.
[4] Bias are the simplifying assumptions made by a model to make the target function easier to learn.

also be of nonsense, or even it can decrease the performance. For this reason, it is important to find the proper data augmentation method that will increase the recognition accuracy and robustness of the PD early detection system. Different data augmentation techniques can be used to generate artificial data: geometric transformation (shift, scale, rotation/reflection, Time-Wrapping, etc.), noise addition [Um et al., 2017], and deep generative models (such as Generative Adversarial Networks (GANs)). The idea behind the deep generative models is to learn the true data distribution of the training set, and generating a new data point following the same learned distribution [Panday, 2018]. From the other side, minor changes due to the geometric transformation or noise addition will not alter the data labels because they are likely to happen in real world observations (when pen sensors are not 100 % precise). Due to the absence of large databases in literature to train deep generative models, we decided to only apply geometric transformations and additive noise to augment our data. Data augmentation will be applied on the best time series combinations found in section 5.3.

#### 5.4.2.1 Data augmentation techniques used

Jittering, scaling, time-warping, and synthetic data generation techniques are used to generate new time series samples. These methods do not crop time series into shorter subsequences. This enables the network to learn discriminative properties from the whole time series in an end-to-end manner [Fawaz et al., 2018]. In the following we will go through each technique.

##### 5.4.2.1.1 Jittering

Jittering is considered as a way of simulating additive sensor noise. We focus on adding Gaussian noise to each feature time series of the original training data to obtain new training samples [Wang et al., 2018]. It can be considered as applying different noise to each sample of the time series. In order to ensure that the amplitude value of the sample will not be changed with the addition of noise, we generate the Gaussian noise with $\mu=0$ as shown in Figure 5.59. In order to explore the effect of noise intensity (STD) and the augmented multiple (m) on the work of time series data augmentation, different values of STD and m are studied. The amount of noise added is a hyper-parameter. Too little noise has no effect,

whereas too much noise makes the mapping function too challenging to learn or may alter the labels because it introduces rapid fluctuations which look similar to tremor [Um et al., 2017].

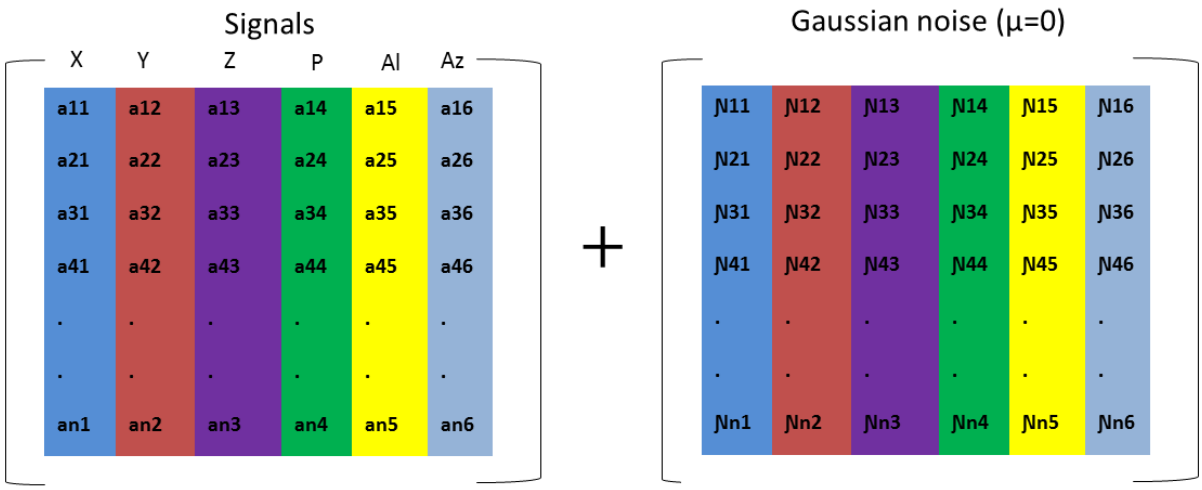



Figure 5.59. Jittering data augmentation with additive Gaussian noise of zero mean.

#### 5.4.2.1.2 Scaling

Scaling changes the magnitude of the data in a window by multiplying by a random scalar [Um et al., 2017]. It is considered as a way of simulating multiplicative sensor noise. We also focus on multiplying Gaussian noise (with a non-zero mean) to each feature time series of the original training data to obtain new training samples. It can be considered as applying constant noise to the entire samples of a time series as shown in Figure 5.60, where ".*" refers to elementwise multiplication. Different values of m and STD are studied and compared.

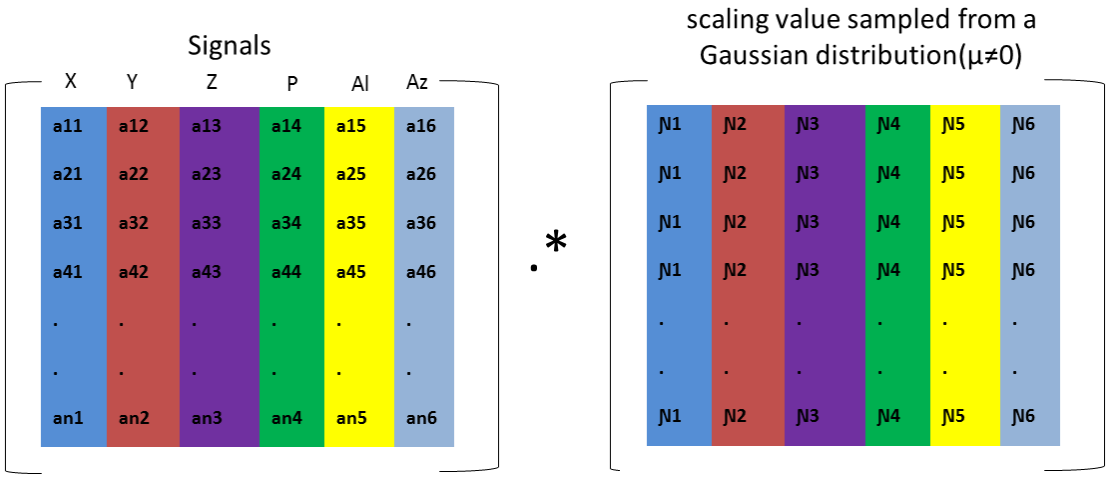



Figure 5.60. Scaling data augmentation with multiplicative Gaussian noise.

#### 5.4.2.1.3 Time-warping

Time-warping is a way to perturb the temporal location by smoothly distorting the time intervals between samples by applying smoothly varying noise to the entire samples [Um et al., 2017]. This varying noise is generated by cubic spline with a given number of knots at random magnitude (here Gaussian distribution is selected with a given mean and STD), where knots reflect the complexity of the curves.

#### 5.4.2.1.4 Generating synthetic data

To create the synthetic time series, Fawaz et al. [Fawaz et al., 2018] propose to average a set of time series and to use the averaged time series as a newly created example. Data is separated into subsets of the same class label, where the size of each subset is calculated and the maximum is selected and defined by G. The number of synthetic data per class Sl is equal to 2G-Hl; where Hl refers to the size of class l. The number of added synthetic data will rebalance classes in case they are imbalanced. Once the number of added synthetic data per class is obtained, the next step is to assign weight to each time series in the subset in a way to get generated examples that closely follow the distribution of the original samples. Two methods were developed in [Fawaz et al., 2018] for this goal. The first method is defined by averaging all the time series having the same class label by giving different weights to create diversity in the synthetized ones. These weights should follow a flat Dirichlet distribution with unit concentration parameter. This method is not an ideal solution because it has potential to fill up the complete convex hull of the original data. In the case where the classes from two interlaced U-shapes, filling the inside part of the “U” would lead to inappropriate examples. The second method was proposed to overcome the problem mentioned above, where this method is applied in this work. The intuition behind this method is to use a subset of close time series and fill their bounding boxes. The weights for each subset are assigned based on the following steps: starting with a random initial time series chosen from the subset, it is assigned with a weight equal to 0.5. Then the 5 nearest neighbors are found, and randomly 2 out of these 5 neighbors are selected and assigned with a weight equal to 0.15 each [Fawaz et al., 2018]. Therefore, in order to have a normalized sum of weights, the rest of time series in the subset will share the rest of the weight 0.2. The new generated time series length is equal to the initial time series chosen.

To get the nearest neighbors, Dynamic Time Warping (DTW) instead of the Euclidean distance is applied since we are dealing with time series of variable length, and working with noisy signals where the Euclidean distance does not handle noise. DTW gives more robustness to the similarity computation in our case. Time series of different length can be compared by replacing the one to-one point comparison, used in Euclidean distance, with a many-to-one (and vice versa) comparison. The main feature of this distance measure is that it allows recognizing similar shapes, even if they present signal transformations, such as shifting and/or scaling. To get the distance between two time series, DTW algorithm works as follows [Kyaagba, 2018]:

1. Calculate the euclidean distance between the first point in the first series and every point in the second series. Store the minimum distance calculated. (this is the ‘time warp’ stage)

2. Move to the second point and repeat 1. Move step by step along points and repeat 1 till all points are exhausted.

3. Repeat steps 1 and 2, but this time with the second series as a reference point.

4. Add up all the minimum distances that were stored and this is a true measure of similarity between the two series.

### 5.4.3 Combination approach

The experiments described in this work are divided into two rounds: single assessment and combined assessment. In the single assessment, we analyze each task separately, while in the combined assessment we combine the outputs of the 7 models (one per task) in order to find the final label and obtain what we call the overall performance of the all-task system. Each model outputs two values corresponding to the probabilities that the input time series, associated to the given task, are performed by a parkinsonian or a HC subject respectively.

Two combination schemes are considered, majority voting and MLP-based combinations. Majority voting consists on choosing the class label which has the maximum number of votes by each 2D CNN or 1D CNN-BLSTM classifier. The MLP-based

combination consists in combining the probability vectors of size 2 provided by each of the 7 models. The MLP model is composed of an input layer of 2x7=14 nodes, a single hidden layer with ReLU activation function, and 2 output nodes (corresponding to PD and HC) with softmax activation function. The number of hidden nodes is chosen according to the empirical rule defined in (5.64).

Majority voting is used as a baseline combination scheme when systems are trained using transfer learning, while neural-based combination can be seen as an enhanced combination scheme that will be used in conjunction with data augmentation.

In addition, several all-task systems can be trained from our architectures. For instance by using distinct data augmentation approaches (see section 5.4.2). For combining these all-task systems, a meta-MLP approach is used that combines the outputs of the previous MLPs, 2 per all-task system.

### 5.4.4 Experiments and results

We will investigate transfer learning and data augmentation approaches based on the 2D CNN and 1D CNN-BLSTM models selected in section 5.3 for time series classification in order to perform PD detection on large-scale data and to increase the recognition accuracy and robustness of recognition system. Also cross-entropy cost function and Adam optimizer are applied here, where early stopping procedure and dropout technique (where a dropout layer is added right after the hidden layer with 0.4 dropout rate) are applied to prevent the network from overfitting, and mini-batch training combined with shuffling at each epoch is applied for faster convergence. No batch normalization was applied.

Different parameter values (STD, knots and m) are applied and compared for data augmentation; where the number of hidden nodes in both deep models (2D CNN and 1D CNN-BLSTM) will be depending on m and the empirical rules defined previously. For jittering, a random additive scalar is sampled from a Gaussian distribution with zero mean and STD of 0.3. This value was chosen not large neither small in order to not alter the labels [Um et al., 2017]. For scaling, a random multiplicative scalar is sampled from a Gaussian distribution with mean of 1 and STD of 0.1. For time-warping, random smooth warping curve

is generated by cubic spline with five knots at random magnitudes (mean of 1 and STD of 0.2). Figure 5.61 presents the raw input time series that were selected in section 5.3 and the augmented ones using jittering, scaling, time warping, and synthetic data generation for a given subject and task (here it refers to task 1). Due to the small variation between the original and the artificial time series, a logarithmic scale is applied to get a better view of the difference between time series. The best accuracy is achieved when the training data is augmented twice. Data augmentation procedure for each of the two models (the 2D CNN and the 1D CNN-BLSTM) is resumed in Figure 5.62. For the 2D CNN model, the original raw time series combination (X, Y, Z, pressure, and altitude) are pre-processed; where X is flipped for Arabic samples and then X and Y are normalized as described in section 5.3.1.1. The correct way to apply data augmentation is by dividing data into training and testing sets and applying the augmentation techniques to the training set only, where the testing set contains only original samples. The time series in both the augmented training data and the testing data set are converted into 2D normalized grayscale spectrogram images (using the method described in section 5.3.1.2.3) and applied to train and evaluate the 2D CNN model as shown in Figure 5.62-a.

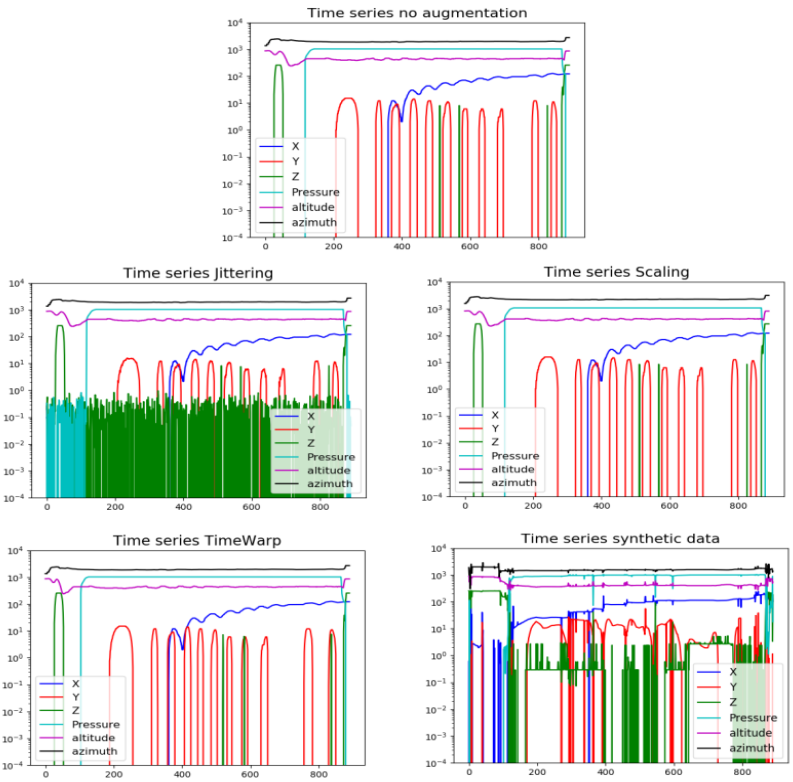



Figure 5.61. Raw input time series and time series obtained by various data augmentation approaches such as: jittering, scaling, time-warping, and synthetic data generation for Task 1.

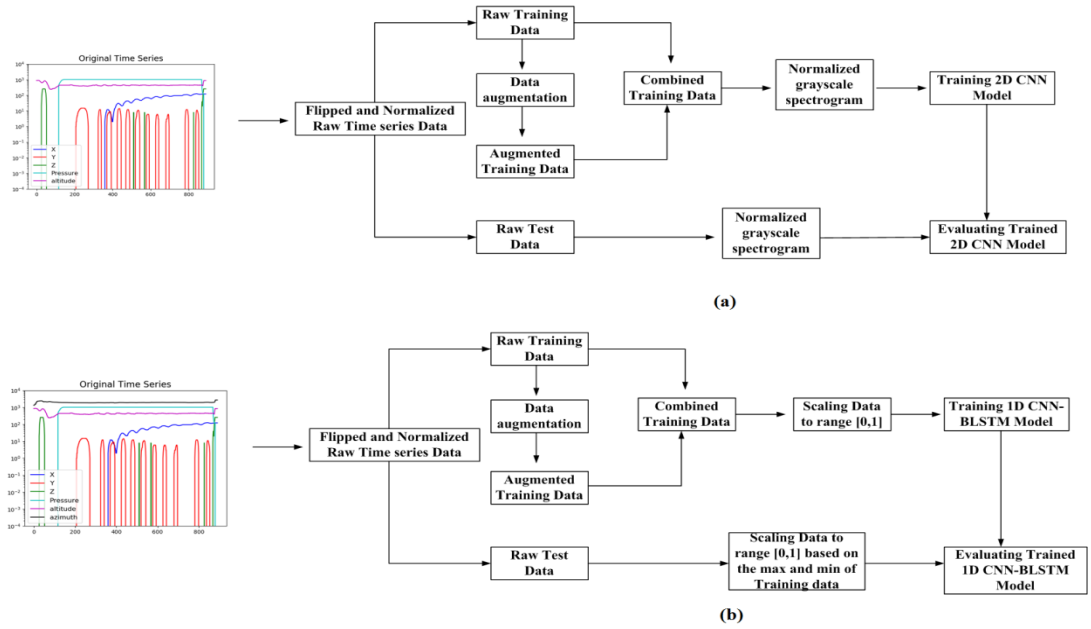



Figure 5.62. Data augmentation applied on time series for both: (a) 2D CNN and (b) 1D CNN_BLSTM models.

The same procedure described above is applied for the 1D CNN-BLSTM, where the original raw time series combination (X, Y, Z, pressure, altitude, and azimuth) are used as shown in Figure 5.62-b. The only difference is that the time series in both the augmented training and the testing data are scaled to the range (0, 1) before training and evaluating the 1D CNN-BLSTM model using the raw time series.

The different transfer learning strategies described in section 5.4.1 across the 2D CNN model are studied and analyzed; where majority voting is applied to merge the results provided by the 7 models (referring to the 7 tasks). The performance in Table 5.19 represents the average of 3 runs (3-folds CV). Comparing the transfer learning strategies, fully freezing performs worse than other strategies; where incremental performance may be observed when more convolutional layers are included in fine-tuning (partial freeze 1 and partial freeze 2). According to these results, it is clear that using PaHaW database to pre-train the 2D CNN model performs worse than training from scratch. This finding can be related to many factors: the first one is that the dataset used for pre-training the 2D CNN model (the PaHaW dataset) is also considered a small size dataset. The second factor can be related to the absence of Z coordinate feature in the PaHaW database, and it seems that this feature plays an important role in classification. Finally, PaHaW is a not a multilingual dataset as the HandPDMultiMC,

so using the pre-trained modal on unilingual dataset may be not suitable for multilingual dataset.

Table 5.19. Comparison of various transfer learning strategies across the 2D CNN model; where majority voting is used as combination approach.

| **Model** | **Data Input** | **Transfer Learning Strategy** | **Best Features Combination** | **Overall Per. (%)** |
|---|---|---|---|---|
| 2D CNN | Spectrogram images | From scratch (no transfer learning) | X+Y+Z+ Pressure+ Altitude | **Acc:83.33** Sens:85.71 Spec:80.95 |
| | Spectrogram images | Retrain classification layer | X+Y+ Pressure+ Altitude | Acc:54.76 Sens:28.57 Spec:80.95 |
| | Spectrogram images | Partial freeze 1 | X+Y+ Pressure+ Altitude | Acc:66.67 Sens:66.67 Spec:66.67 |
| | Spectrogram images | Partial freeze 2 | X+Y+ Pressure+ Altitude | Acc:66.67 Sens:66.67 Spec:66.67 |
| | Spectrogram images | Fully freeze | X+Y+ Pressure+ Altitude | Acc:45.24 Sens: 71.43 Spec: 19.05 |

Moving to data augmentation strategy, the main results of the proposed techniques are presented in Table 5.20. The 2D CNN and the 1D CNN-BLSTM all-task models are used here, and a MLP model is applied to combine the probability vectors provided by the 7 models (one model per task) with a single hidden layer of 40 hidden nodes, in order to provide the final classification decision. Scaling fails to improve the 1D CNN-BLSTM performance because changing in the intensity of the signal may alter the labels [Um et al., 2017]. On the other hand, jittering, time-warping, and creating synthetic time series by averaging a set of time series used with 1D CNN-BLSTM model improve the accuracy of PD classification by 7.15 % (3-fold CV accuracy). Data augmentation improves the 1D CNN-BLSTM performance and fails to improve the 2D CNN performance because the 1D CNN-BLSTM deals with time series directly without encoding them into spectrogram images (the case of 2D CNN), which will benefits the most from the data augmentation techniques for time series.

Training and testing accuracy curves for 1D CNN-BLSTM all-task model with Jitter, time-warping, and averaging time series data augmentation techniques, and with MLP combination approach are depicted in Figure 5.63. The accuracy plots show how data

augmentation improves the accuracy of our deep learning model and helps in reducing overfitting.

Table 5.20. 3-folds CV performance measures obtained after applying data augmentation and MLP for classification decision.

| **Model** | **Data Input** | **Augmentation Technique** | **Best Features Combination** | **Overall Per. (%)** |
|---|---|---|---|---|
| 2D CNN | Spectrogram images | jitter | X+Y+Z+Pressure+ Altitude | Acc :83.33<br>Sens:85.71<br>Spec:80.95 |
| 1D CNN-BLSTM | Raw Time series | jitter | X+Y+Z+Pressure+Altitude+Azimuth | **Acc :90.48**<br>Sens:95.24<br>Spec:85.71 |
| | | scaling | X+Y+Z+Pressure+Altitude+Azimuth | Acc :59.52<br>Sens:19.05<br>Spec:100 |
| | | time-warping | X+Y+Z+Pressure+Altitude+Azimuth | **Acc :90.48**<br>Sens:90.48<br>Spec:90.48 |
| | | synthetic data | X+Y+Z+Pressure+Altitude+Azimuth | **Acc :90.48**<br>Sens:85.71<br>Spec:95.24 |

Task-wise system accuracies for the developed models in this study and the SVM model developed in our previous work [Taleb et al., 2017] are represented in Table 5.21; where D1, D2, D3, and D4 models refer to SVM, 1D CNN-BLSTM with jittering augmentation, 1D CNN-BLSTM with time-warping augmentation, and 1D CNN-BLSTM with synthetic augmentation models *respectively.* It can be observed from Table 5.21 that "all tasks" reports highest accuracies across all the 4 models (D1, D2, D3, and D4). Additionally, Task 2 (triangular wave) and Task 3 (rectangular wave) report the highest accuracies across all the 4 models, the same conclusion found in [Taleb et al., 2017] and [Taleb et al., 2019-a].

We also carry out experiments by combining the results obtained from the 1D CNN-BLSTM all-task model with different data augmentation methods (D2, D3, and D4 in Table 5.21) using a MLP model composed of an input layer of L nodes (L is set to 4 when two data augmentation methods are combined, and 6 when three data augmentation methods are combined), a single hidden layer of H nodes with ReLU activation function (H is set to 30 when two data augmentation methods are combined, and 35 when three data augmentation methods are combined), and 2 output nodes (corresponding to PD and HC) with softmax activation function. Also the empirical rule defined in (5.64) is applied to get the number of hidden nodes. The performance measures realized in these experiments are summarized in

Table 5.22. It can be seen that combining the results of two various data augmentation methods show better performance than that of a single data augmentation. The highest accuracy is 97.62 % obtained when combining jittering and synthetic data augmentation methods. However, when we combine the results of three data augmentation methods, we found how the existence of time-warping augmentation method in the combination deteriorates the performance (a decrease from 97.62 % to 92.86 %). From a clinical point of view, inter-samples timing disturbances occurs due to the neuro-motor dysfunctions affecting wrist and finger movement of PD patients [Drotar et al., 2016], [Gómez-Vilda et al., 2017]. As mentioned before, the temporal locations of the samples are changed by time-warping; which will look similar to inter-samples time disturbances.

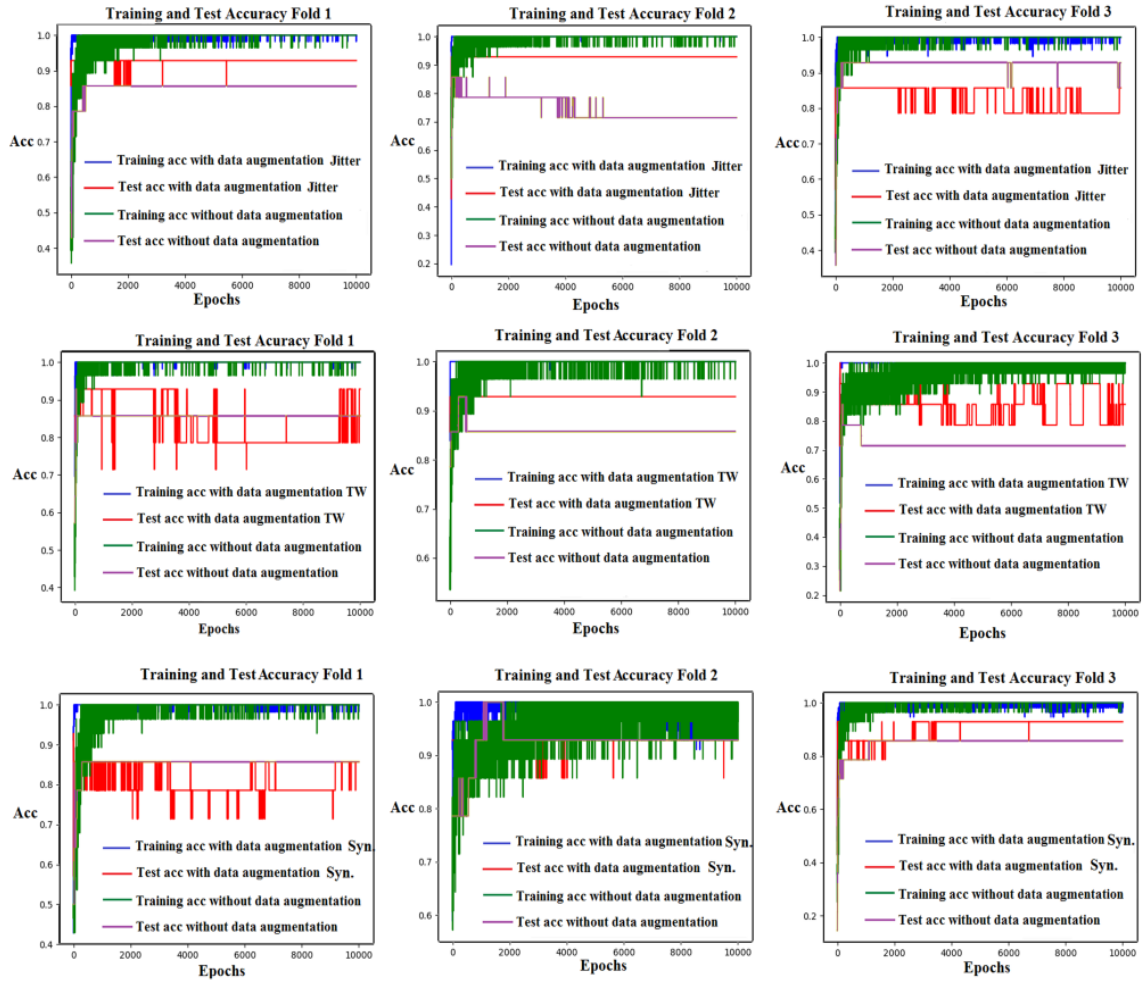



Figure 5.63. Training and testing curves for 1D CNN-BLSTM all-task model with jitter, time-warping, and averaging time series (or synthetic) data augmentation techniques and MLP combination approach.

Table 5.21. Task-wise system and "All-tasks" system accuracies (in %) for various models and training schemes. D1: SVM, D2: 1D CNN-BLSTM/jitter, D3: 1D CNN-BLSTM/time-warping, D4: 1D CNN-BLSTM/synthetic data.

| Task | D1 | D2 | D3 | D4 |
|---|---|---|---|---|
| Repetitive cursive letter 'l' | **87.5** | 59.52 | 57.14 | 47.62 |
| Triangular wave | **93.75** | 80.95 | 83.33 | 78.57 |
| Rectangular wave | **90.63** | 71.43 | 69.05 | 76.19 |
| Repetitive "Monday" | **87.5** | 78.57 | 66.67 | 76.19 |
| Repetitive "Tuesday" | **87.5** | 57.14 | 47.62 | 59.52 |
| Repetitive "Name" | **84.38** | 57.14 | 42.86 | 50 |
| Repetitive "Family Name" | **84.38** | 69.05 | 71.43 | 64.29 |
| All tasks (MLP combination) | **96.87** | **90.48** | **90.48** | **90.48** |

Table 5.22. Performance (in %) obtained by combining two or three all-task 1D CNN-BLSTM models trained with distinct data augmentation approaches. D2: 1D CNN-BLSTM/jitter, D3: 1D CNN-BLSTM/time-warping, D4: 1D CNN-BLSTM/synthetic data.

| Performance | D2+D3 | D2+D4 | D3+D4 | D2+D3+D4 |
|---|---|---|---|---|
| Accuracy | 92.86 | **97.62** | 92.86 | 92.86 |
| Sensitivity | 90.48 | **95.24** | 95.24 | 95.24 |
| Specificity | 95.24 | **100** | 90.48 | 90.48 |

The best final model that classifies people with PD and HC with an accuracy of 97.62 % is summarized in Figure 5.64, where jittering and synthetic data augmentation techniques are applied separately on 'All-tasks" system. Two MLPs models (MLP1 and MLP2) are applied, where each one is used to combine the probability vectors (each of size 2) obtained by the 7 1D CNN-BLSTM models (with the best feature set X, Y, Z, pressure, altitude, and azimuth) that are trained with distinct data augmentation approach. At a later stage, another MLP model (MLP3) is used to combine the probability vectors provided by each of MLP1 and MLP2 (each of size 2) in order to get the final prediction. Xavier normalized initialization (defined in section 5.3.1.3.1) is also applied here to initialize the weights of all the MLPs, where biases are zeros initialized. This model was trained with Nvidia GTX 1080 GPU of 8 GB memory. The number of parameters in the deep model presented in Figure 5.64 is obtained based on equations (5.66), (5.67), (5.68) and (5.69). Table 5.23 presents the total number of parameters trained in our model per fold. The time required for the training process is around 1 day, where 204,060 parameters have been learned.

An overview of the existing models applied for PD early detection through handwriting analysis is summarized in Table 5.24. In [Drotar et al., 2015-a], [Mucha et al., 2018], [Taleb et al., 2017] the authors have applied SVM models that are trained on global hand-crafted features for PD detection. Drotar et al. [Drotar et al., 2015-a] found that a

combination of kinematic, temporal, pressure, and intrinsic features return a classification accuracy of 89.09 %, where Taleb et al. [Taleb et al., 2017] report a higher accuracy of 96.87 % when a combination of kinematic, pressure, and correlation between kinematic and pressure features is used. Mucha et al., [Mucha et al., 2018] proposed another promising approach that returns an accuracy of 97.14 % when a combination of kinematic and temporal features that are extracted for both "on-paper" and "in-air" is used. In addition, Moetesum et al. [Moetesum et al., 2019], Khatamino et al. [Khatamino et al., 2018], Galicchio et al. [Galicchio et al., 2018], Pereira et al. [Pereira et al., 2018] and this study [Taleb et al., 2019-b] proposed to use deep learning to learn features from online handwriting exams.

Table 5.23. The total number of parameters trained in our model per fold.

| **Data augmentation technique** | **Task** | **Layer** | **Characteristics** | **Nb. Of parameters** |
|---|---|---|---|---|
| Jittering or synthetic data | Task 1 | Conv1 | *c: 6, f: 1×2, p: 16* | 208 |
| | | Conv2 | *c: 16, f: 1×2, p: 32* | 1,056 |
| | | BLSTM | *g: 4, h: 10, i: 32* | 3,440 |
| | | | *Output weight (2×h,o)*<br>*Output bias (o)* | 42 |
| | Task 2 | Conv1 | *c: 6, f: 1×2, p: 16* | 208 |
| | | Conv2 | *c: 16, f: 1×2, p: 32* | 1,056 |
| | | BLSTM | *g: 4, h: 10, i: 32* | 3,440 |
| | | | *Output weight (2×h,o)*<br>*Output bias (o)* | 42 |
| | Task 3 | Conv1 | *c: 6, f: 1×2, p: 16* | 208 |
| | | Conv2 | *c: 16, f: 1×2, p: 32* | 1,056 |
| | | BLSTM | *g: 4, h: 10, i: 32* | 3,440 |
| | | | *Output weight (2×h,o)*<br>*Output bias (o)* | 42 |
| | Task 4 | Conv1 | *c: 6, f: 1×2, p: 16* | 208 |
| | | Conv2 | *c: 16, f: 1×2, p: 32* | 1,056 |
| | | BLSTM | *g: 4, h: 10, i: 32* | 3,440 |
| | | | *Output weight (2×h,o)*<br>*Output bias (o)* | 42 |
| | Task 5 | Conv1 | *c: 6, f: 1×2, p: 16* | 208 |
| | | Conv2 | *c: 16, f: 1×2, p: 32* | 1,056 |
| | | BLSTM | *g: 4, h: 10, i: 32* | 3,440 |
| | | | *Output weight (2×h,o)*<br>*Output bias (o)* | 42 |
| | Task 6 | Conv1 | *c: 6, f: 1×2, p: 16* | 208 |
| | | Conv2 | *c: 16, f: 1×2, p: 32* | 1,056 |
| | | BLSTM | *g: 4, h: 10, i: 32* | 3,440 |
| | | | *Output weight (2×h,o)*<br>*Output bias (o)* | 42 |
| | Task 7 | Conv1 | *c: 6, f: 1×2, p: 16* | 208 |
| | | Conv2 | *c: 16, f: 1×2, p: 32* | 1,056 |
| | | BLSTM | *g: 4, h: 10, i: 32* | 3,440 |
| | | | *Output weight (2×h,o)*<br>*Output bias (o)* | 42 |
| | All-tasks | MLP1 or MLP2 | *i: 14, h: 40, o: 2* | 682 |
| Combination of both techniques | All-tasks | MLP3 | *i: 4, h: 30, o: 2* | 212 |
| **Total number of trained parameters per fold** | | | | **68,020** |

Moetesum et al. [Moetesum et al., 2019], exploited the static visual attributes of handwriting and fed the image into a 2D CNN model for feature extraction, where an SVM was used for prediction. This approach reaches an accuracy of 83 %. In [Pereira et al., 2018], the authors have proposed to encode the handwriting signals into images before feeding them into a 2D CNN model. This approach reaches an accuracy of 93.42 %. Similarly, Khatamino et al. [Khatamino et al., 2018], propose to study both the dynamic extracted signal-based images and visual attributes of spirals, where the extracted signal-based image is the one proposed by Pereira. A 2D CNN model was applied for feature extraction and prediction returning an accuracy of 88.89 %. However, Galicchio et al. [Galicchio et al., 2018], came with a DeepESN model for PD detection that deal with the raw signals directly and returning an accuracy of 89.33 %. Finally, in this work a multilingual 1D CNN-BLSTM model for PD detection dealing directly with the raw signals and combining two different data augmentation techniques (jittering and synthetic data) returns the best accuracy. Accuracy along with 95 % confidence intervals for this system is: 97.62 % (93.01-100 %). This model with “All-tasks” SVM yield significantly (at 95 %) better performance than all experimented models. Table 5.24 shows that our deep learning model reports highest performance across all the mentioned works (especially our SVM model described in section 5.1) although results are not always measured on the same database.

For the sake of comparison, we have conducted 2 more experiments. First, our best model has been trained and tested on the PaHaW database (where PD patients were examined in their “on-state”). In this database there is no Z coordinate so that training was performed with X, Y, pressure, altitude and azimuth time series. The second experiment consists in training and tests our best system on the HandPDMultiMC dataset using X, Y, pressure, altitude and azimuth time series, and removing the Z coordinates. The performances obtained in both experiments are shown in Table 5.24. When our best model is trained and tested using HandPDMultiMC dataset with Z coordinates, the accuracy obtained is 97.62 %, whereas the accuracy obtained when eliminating the Z coordinates feature is 90.48 %. Moreover, when the system is trained and tested on PaHaW dataset, the accuracy obtained (88.1 %) is close to the one obtained with HandPDMultiMC, without Z feature. These results confirm the importance of Z coordinates feature and the relevance of the results obtained (sine the results are consistent on different datasets).

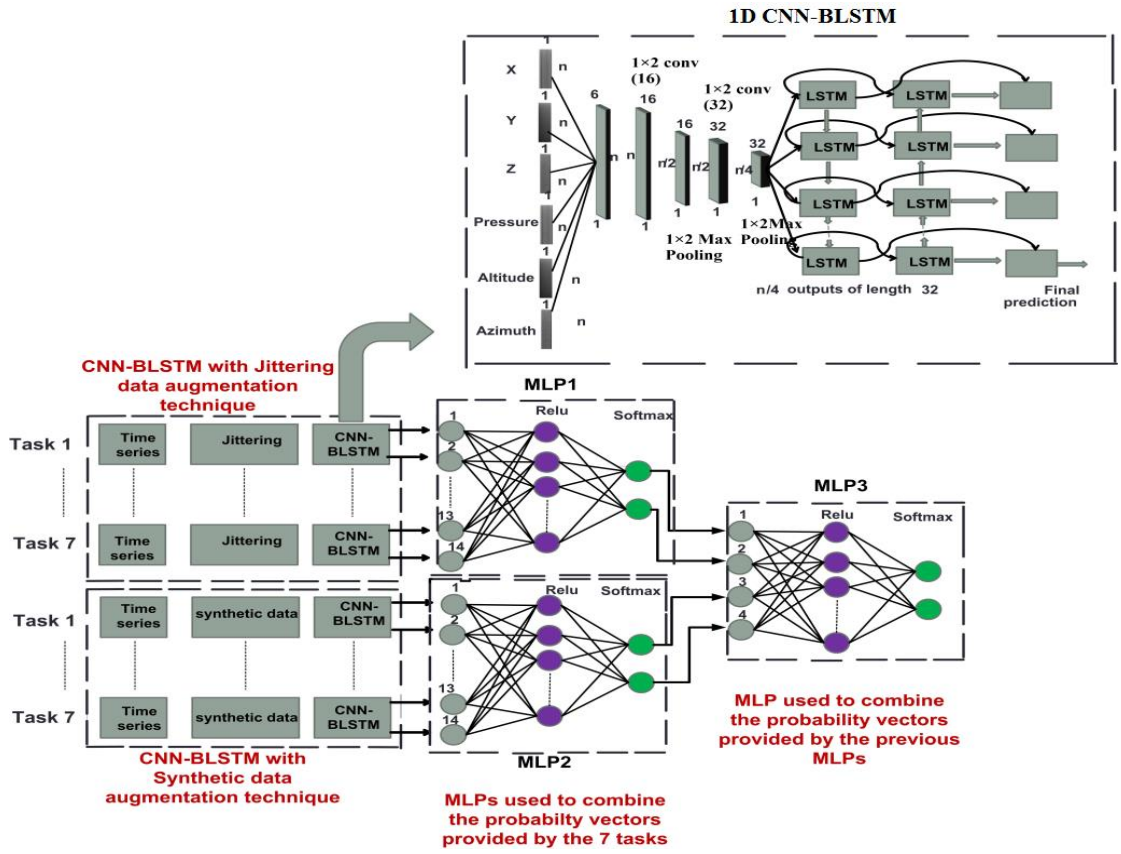



Figure 5.64. Best final model that combines through MLPs the outputs of 14 1D CNN-BLSTM systems trained with augmented data.

Table 5.24. Comparison table between our models and previous models studied.

| Reference | Database | Model | Features | Perf. (%) |
|---|---|---|---|---|
| [Drotar et al., 2015-a] | PaHaW<br>PD in on-state | SVM | kinematic, temporal, spatial, entropy, EMD, pressure (on-paper) | Acc: 89.09<br>Sens: N/A<br>Spec: N/A |
| [Taleb et al., 2017] | PDMultiMC<br>PD in on-state | SVM | Kinematic, stroke, pressure, entropy, EMD (on-paper) | Acc: **96.87**<br>Sens: 93.75<br>Spec: 100 |
| [Mucha et al., 2018] | PaHaW<br>PD in on-state | SVM | Kinematic, temporal (on-paper and in-air) | Acc: **97.14**<br>Sens: 95.50<br>Spec: 100 |
| [Moetesum et al., 2019] | PaHaW<br>PD in on-state | SVM | CNN-based visual features (on-paper) | Acc: 83<br>Sens: 84<br>Spec: 82 |
| [Khatamino et al., 2018] | ParkinsonHW<br>PD in on-state | 2D CNN | CNN-based visual features or CNN-based features (on-paper) | Acc: 88.89<br>Sens: N/A<br>Spec: N/A |
| [Gallicchio et al., 2018] | ParkinsonHW<br>PD in on-state | DeepESN | Tablet raw signals | Acc: 89.33<br>Sens: 90<br>Spec: 80 |
| [Pereira et al., 2018] | HandPD<br>PD in early-stages | 2D CNN-ImageNet | CNN-based features (on-paper and in-air) | Acc: 93.42<br>Sens: 97.84<br>Spec: 89.00 |

| | | | | |
|---|---|---|---|---|
| Proposed Model [Taleb et al., 2019-b] | HandPDMultiMC PD in on-state | 1D CNN-BLSTM | CNN-based features (on-paper and in-air) | Acc: **97.62** |
| | | | | Sens: 95.24 |
| | | | | Spec: 100 |
| | HandPDMultiMC (Z feature ex-cluded) PD in on-state | 1D CNN-BLSTM | CNN-based features (on-paper and in-air) | Acc: 90.48 |
| | | | | Sens: 90.48 |
| | | | | Spec: 90.48 |
| | PaHaW PD in on-state | 1D CNN-BLSTM | CNN-based features (on-paper and in-air) | Acc: 88.10 |
| | | | | Sens: 85.71 |
| | | | | Spec: 90.48 |

In order to approve that our deep learning model surpass the SVM model and to check if data augmentation is also effective for SVM model, we did experiments on SVM with the augmented data. Jittering data augmentation technique is applied to generate new synthetic raw signals, where the training data is augmented twice. Several noise intensities (STD) are analyzed and compared. The set of features selected in section 5.1 are extracted from the time series in both the augmented training data and the testing data set and applied to train and test the SVM model as shown in Figure 5.65. The results obtained are presented in Table 5.25. We have found that augmenting the training data did not improve the classification accuracy of SVM. Augmented data does not always guarantee an improvement in the performance, especially in the case of SVM model, since the global features may get more smoothed when introducing data augmentation with variability, so the capability of the SVM to discriminate between classes will be reduced.

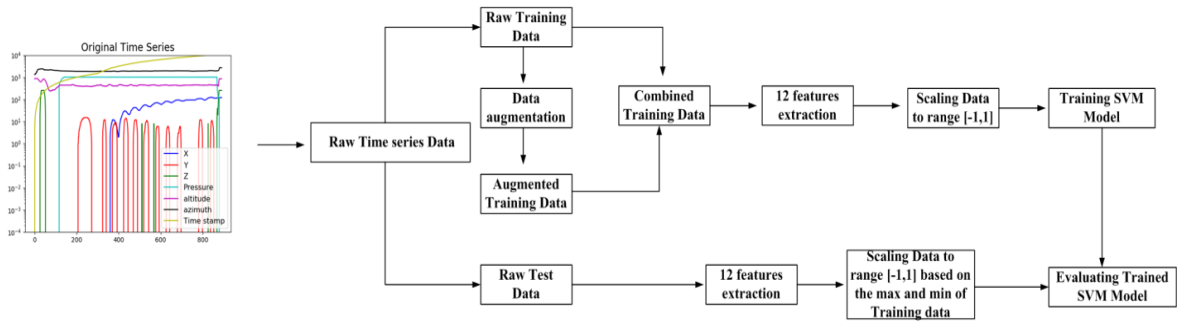



Figure 5.65. Data augmentation applied on time series for SVM model.

Table 5.25. Comparison of SVM classification performance before and after jittering data augmentation.

| Model | Accuracy (%) | Sensitivity (%) | Specificity (%) |
|---|---|---|---|
| SVM (Baseline) | **96.87** | **93.75** | **100** |
| SVM + jittering data augmentation ($\nabla = 0.1$) | 84.37 | 81.25 | 87.5 |
| SVM + jittering data augmentation ($\nabla = 0.2$) | 87.5 | 75 | 100 |
| SVM + jittering data augmentation ($\nabla = 0.3$) | 84.37 | 81.25 | 87.5 |

## 5.4.5 Conclusions

In this part of the thesis, an automatic multilingual classification system for PD early detection is developed based on online handwriting. In our previous work described in section 5.3, two based learning models for end to-end time series classification were proposed, namely the 2D CNN and the 1D CNN-BLSTM. We have demonstrated the importance of both: a deep architecture based on the combination of 1D CNN and BLSTM recurrent layers, and a 2D CNN model with spectrograms as input in PD detection. However, deep learning models require a large number of training samples to work well alike the SVM that is less sensitive to the number of training samples. This means that PD classification using deep learning is a challenging task due to the limited data availability. To cope with the limited data, two main classes of approaches were reported in this section. Firstly, multiple transfer learning strategies across the 2D CNN model for time series classification were investigated and compared. It was found that the more convolutional layers included in the fine-tuning, the better performance we get. However, there are no gains of transfer learning over training our 2D CNN model from scratch. We believe that this can be related to different factors: the dataset used for pre-training the 2D CNN model is also small in size, the absence of Z coordinate feature in the PaHaW database, and using a pre-trained modal on unilingual dataset may be not suitable for multilingual dataset.

Secondly, jittering, scaling, time-warping, and synthetic data generation techniques are used for data augmentation. The challenging PD task is successfully tackled using the 1D CNN-BLSTM model described above and the combination of jittering and synthetic data augmentation methods. The accuracy performance is improved from 83.33 % to 97.62 %.

It is important to summarize a number of observations and conclusions obtained from this work. First of all, we found that the Z coordinates feature play an important role in PD classification. This can be explained by the fact that PD is characterized by tremor or irregular muscle contractions that introduce randomness to the movement during handwriting in the X-Y-Z space. The intensity of tremor usually decreases when hand is laid down on a surface. This means that tremor oscillations have larger amplitude when the pen is not touching the surface (in-air phase). For this reason, it will more efficient to analyze tremor along the three axes since there is no consensus on which axis tremor oscillations have larger

amplitudes. We have also found the effectiveness of data augmentation over transfer learning at reducing error and decreasing overfitting were shown. In addition, time-warping technique fails to improve the performance of PD classification due to the distortion of time intervals between samples that look similar to inter-samples time disturbances; which is one of the early marks of PD. Also, data augmentation methods applied for time series classification can increase deep learning model performance when time series are used directly with no need to convert them into images. Finally, augmenting training data using synthetic generated samples deteriorates the SVM classifier performances, where it improves deep learning performance so that it can surpass the models trained on global pre-engineered features even though the available data is small. We have proved in this study that despite the limited size of our dataset, short-term analysis with deep learning and data augmentation returns some interesting results in PD early detection task.

# 6 Effect of Voice's Sampling Rate and Unvoiced Sounds on Parkinson's Disease Early Detection Performance

Recently, the automatic detection of PD through speech has gained the interest of the scientific community. Existing works differ on many aspects: on the set of features considered, on the speech tasks used for analysis, and on the statistical approaches applied. Nevertheless, a language-independent model to detect PD using speech features has not been enough addressed. The main goal of this work is to build a language-independent system for assessment the motor disorders in PD patients based on speech signals, using SVM.

## 6.1 Speech related organs

The mechanism for generating the human voice can be subdivided into three parts; the pulmonary system, the vocal folds, and the vocal tracts (or articulators) as shown in Figure 6.1. Starting with the pulmonary system, which is composed of the lungs and the respiratory airways (tubes allowing the passage of air from the atmosphere to the lungs and vice versa). The role of the lungs is to assist metabolism through respiration, and to provide energy for the speech production system [Benesty et al., 2017]. Speech production system depends on the airflow along the respiratory tract, originated in the lungs and travelling along the trachea. Two different phases exist in the respiration: the inspiration and the expiration phases. In the inspiration, the lungs expend and air flows into the lungs, were in the expiration the lungs collapse and the air flows out. All these movements are controlled by the diaphragm muscles which expend and collapse.

The vocal folds (or vocal cords) are located above the trachea and across the larynx. During voiced speech sounds, the vocal folds are set into vibration by pressurizes the air passing through the membranous portion of the narrowed glottis (the gap between the free edges of the vocal folds) [Benesty et al., 2017]. The glottal airflow induces wave-like motion of the vocal fold membrane. When this oscillatory motion builds up, the vocal folds on either

side come into contact with each other. Many factors play major role in the vocal fold vibration such as the stiffness and mass of vocal folds and the width of the glottis.

The nose, mouth, tongue and lips are collectively referred to as the vocal tract. Whereas the vocal folds can be viewed as an oscillator, the vocal tract can be described as a resonator that amplifies certain acoustic frequencies and attenuates others [Titze, 2000].

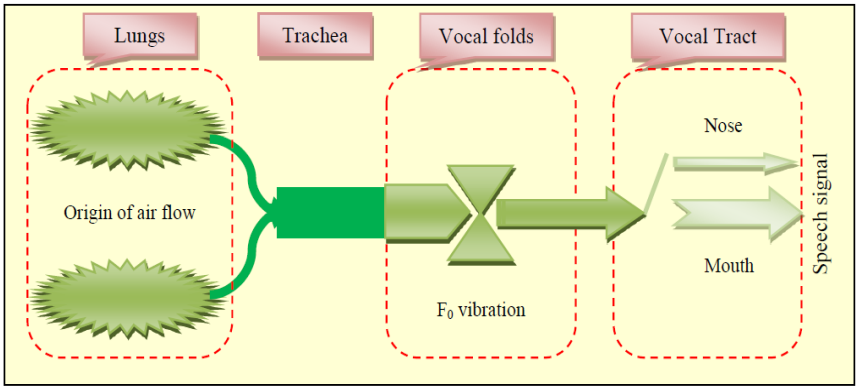



Figure 6.1. Schematic diagram of the major parts involved in the production of speech [Tsanas et al., 2012].

## 6.2 Speech and Parkinson's disease

As mentioned in chapter 2, the motor symptoms of PD include: muscle rigidity, slowness of movement, difficulty initiating movement, and tremor. One of the organs that are affected by this disease is the muscular control of the speech organs. Individuals with PD exhibit hypokinetic dysarthria that involves a disturbance in muscle control that results in weakness, slowness, and incoordination in speech production. Dysarthria covers all malfunctions related to respiration, phonation, articulation, resonance, and prosody.

**Respiration:** The rigidity associated with PD can often lead to a disruption of the respiratory process which serves to generate airflow and air pressures for speech. Due to rigidity, slowness of movement and difficulty initiating movements that occur in chest muscles, several factors can appear such as: reduced respiratory excursions, reduced vital capacity, paradoxical respiratory movements, rapid breathing cycles, and difficulty altering vegetative breathing (through nose) for speech [Duffy, 2013]. These factors can contribute significantly to reduce physiologic support for speech and some of the disorder's phonatory

and prosodic abnormalities, such as reduced loudness (decreasing in intensity), short phrases, short rushes of speech, and inappropriate pauses.

**Phonation:** Phonation is the vibration of the vocal folds to create sound. Rigidity of the larynx muscles results an increased stiffness of the vocal folds. The fundamental frequency can be represented as follow:

$$F_0 = \frac{1}{2L}\sqrt{\frac{k}{m}} \tag{6.1}$$

where L is the vocal length, K is fold stiffness, and m is tissue density [Goberman et al., 2002]. According to the fundamental equation (6.1), an increment in fold stiffness will result to a higher F0 or lower pitch. If the vocal folds get damaged it will affect the voice [Ackermann and Hughes, 2003]. If the folds cannot come together properly, then the air can escape between them causing breathy speech. For vowel prolongation, the vocal folds are unstable due to the difficulty in maintaining the laryngeal muscles in a fixed position [Goberman et al., 2002]. Jitter, shimmer and Harmonic to Noise Ration (HNR) are used as measures to assess the instability of vocal fold vibrations and the presence of incomplete closure of the vocal folds, producing a breathy voice (presence of noise). In addition, due to the incoordination of articulatory and laryngeal, transitions from vowels to following consonants within syllables may be voiceless. The duration of time from the articulatory release of a stop consonant to the onset of voicing for the following vowel is defined as VOT (voice onset time) [Goberman et al., 2002]. A summary of the respiratory and phonation deficits associated with PD, and the relationship between acoustic measures and the anatomy of PD is summarized in Figure 6.2.

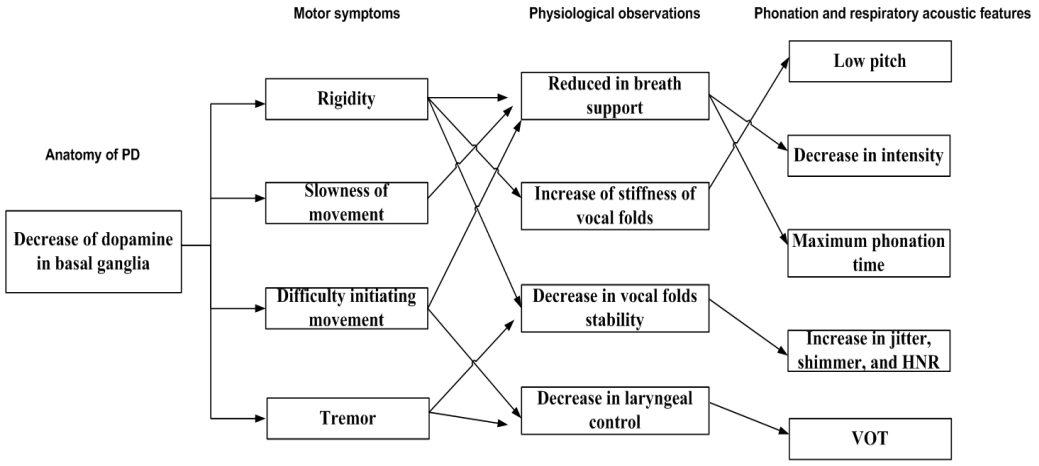



Figure 6.2. Relationship between PD motor symptoms and phonation and respiratory measures.

**Articulation:** Articulation is the modification of the position and shape of the speech organs (e.g., Tongue) in the creation of sound [Goberman et al., 2002]. Reduced articulator movement amplitude and reduced articulator strength lead to an inability to adequately close off the oral cavity. It is characterized by the replacement of a stop gap with low-intensity frication [Duffy, 2013]. This will lead to the "imprecise articulation". Another point to focus on is the range of movement. Due to rigidity and abnormal speed of articulatory movements (e.g., lips muscles rigidity, reduced velocity of lip and jaw movements) the range and speed of articulator movements are reduced. VSA is used to reflect the dynamics of the articulators [Duffy, 2013]. Vowels are in general produced by movements of the tongue, lip, and jaw creating several resonance cavities that have a certain frequency response at a certain frequency bands. The harmonics in such frequency response are called "formants". The first two formants, F1 and F2, are used for modeling vocal sounds. Due to the reduced articulatory speed, F1 and F2 transition rates should be reduced [Goberman et al., 2002]. In addition, the subtle dislocation in the movement of articulators results in varying energy with the frequency bands of speech signal. MFCCs compute the energy differences between the bands of speech frequency which can be used to discriminate between the varying energy levels of disturbed resonances [Khan, 2014]. The relationship between articulation acoustic measures and the PD motor symptoms is summarized in Figure 6.3.

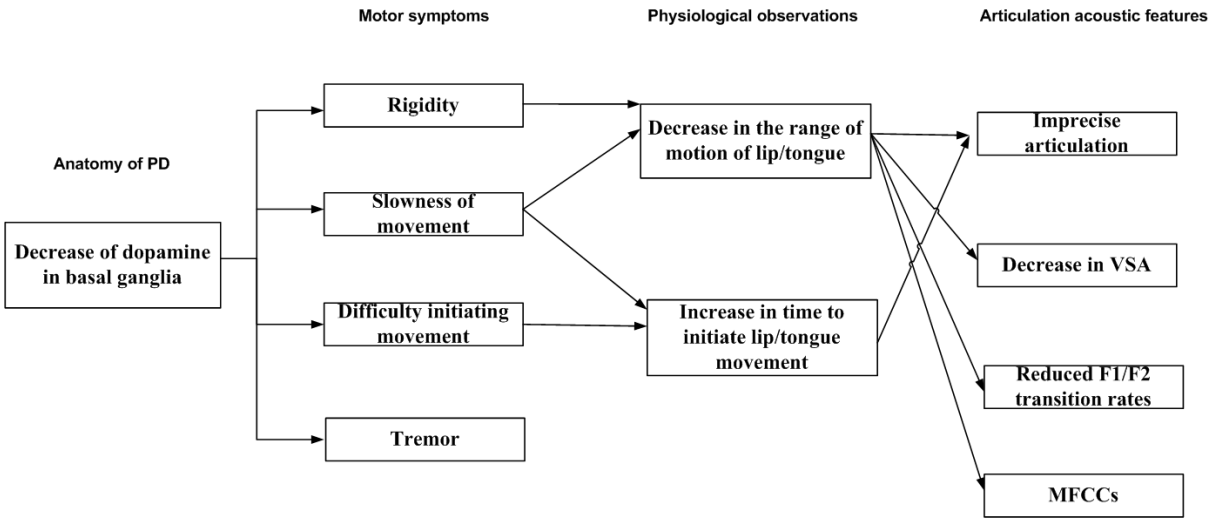



Figure 6.3. Relationship between PD motor symptoms and articulation measures.

**Prosody:** Prosody is the term applied to the natural variations in pitch, intensity, and timing accompanying natural speech [Goberman et al., 2002]. Comparing F0 of the final syllable of a sentence has been used to examine prosody. A decrease in F0 variation during reading tasks may reflect a prosodic deficit. Prosodic intensity changes should also be examined. A decrease in intensity variation during reading tasks may reflect a prosodic deficit. In addition, the rate of speech has been shown to be influenced by prosodic disturbances in PD [Goberman et al., 2002]. The rate disturbance associated with PD can cause speech rate to be accelerated or slowed. Due to the increment in breaths per utterance and the increment in time to initiate lip/tongue movement; the number of pause time (at the end of words and within polysyllabic words) should be increased [Duffy, 2013]. A summary of prosody deficits associated with PD, and the relationship between acoustic measures and the anatomy of PD are provided in Figure 6.4.

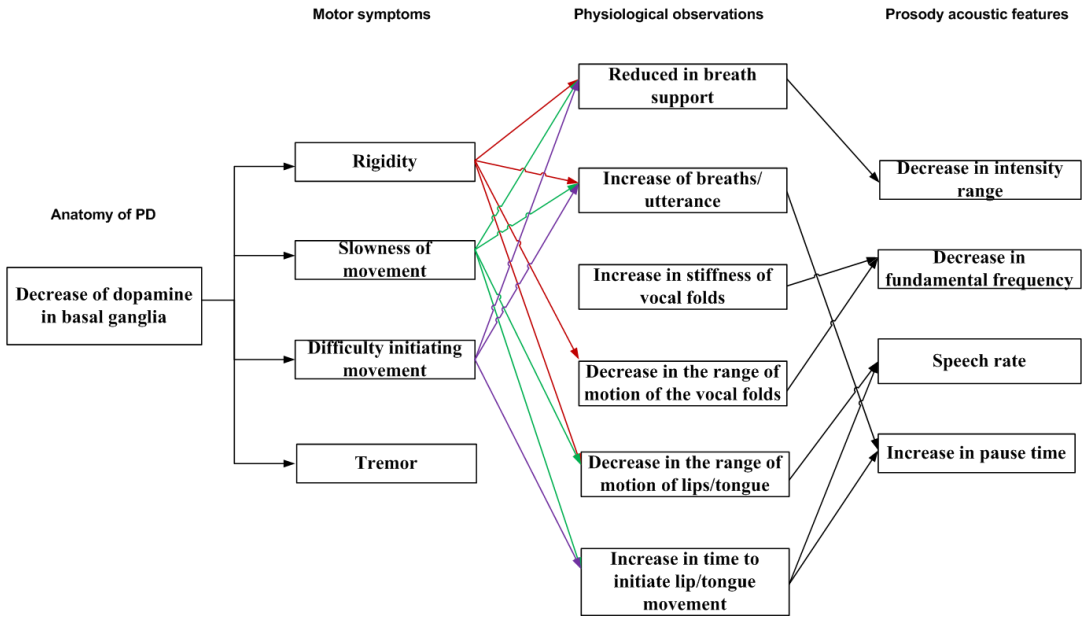



Figure 6.4. Relationship between PD motor symptoms and prosody measures.

## 6.3 Purpose of this work

The purpose of this work is building an acoustic feature set for assessment the motor disorders in PD patients, where these features combining different aspects (phonation, articulation, and prosody) are language-independent and suitable for each task under assessment. However, prosodic information, including intonation, stress, and rhythm, are considered language-specific and differ from language to another [Pinto et al., 2017]. It will be challenging to decide whether prosodic measures differences are related to PD disease or to the language spoken. For this reason, we have decided to study only phonation and articulation aspects to cover different types of motor signs at the same time, since not all the patients develop the same symptoms.

From the other side, the literature does not provide an accurate relationship between the voice sampling rate (Fs) and the accuracy and reliability of acoustic voice analysis. In this work, different sampling rate values are studied to explore the influence on voice analysis for PD early detection.

In addition, most of the works in literature focused on voiced sounds (produced when the vocal cords are closed and vibrating during the pronunciation) to detect PD, and neglect

the unvoiced sounds (produced when the vocal folds are open and stationary). All the vowel sounds are voiced, whereas some consonants are voiced (e.g. B, D, G, J, L, M, N, R) and some others are unvoiced (e.g. F, K, P, S, T). Some works showed that PD patients have impaired production of consonant sounds due to the inappropriate tongue elevation and reduced energy [Orozco-Arroyave, 2016]. Additionally, it is well known that PD patients suffer from movement disorders affecting the voice muscles in the larynx, and this disorder can develop certain involuntary movements called abductor spasms that cause the vocal folds to open [Stemple et al., 2020]. The vocal folds cannot vibrate when they are open too far, and the open position allows air to escape from the lungs during speech. As a result, the voice often sounds weak and breathy. The abductor spasms occur mostly on unvoiced consonants. In this work, to validate that the unvoiced sounds also play a role in PD detection, we have decided to study both voiced and unvoiced frames in a connected speech, where automatic segmentation is applied in order to get the voiced and unvoiced frames. The voiced frames are defined as a portion of speech where there is a detected pitch value, and unvoiced frames are those with no detected pitch (including silence).

In this work, the two speech tasks (sustained vowel 'a' and text reading) recorded for each of the 42 subjects (21 HC subjects and 21 PD patients) in their "on-state" and taken from SpeechPDMultiMC dataset are studied and analyzed.

## 6.4 Pre-processing

Before extracting the acoustic features from speech, several pre-processing steps are applied to the signal. Data is manually preprocessed in order to remove silence at the start and the end of the speech in addition to the speech that does not refer to the subject. Additionally, each spontaneous intervention introduced by the subject that was not directly related with the task is removed as well. Our signals are 2 channels sounds with 16-Bit depth and 44.1 KHz sampling rate. The next pre-processing step is converting the two-channel signals to a single channel to avoid extracting features for both channels individually. In addition, since one of the main goals of this work is to study the effect of sampling rate on voice analysis, voices of various sampling rates should be included. There are two methods to obtain data of this requirement. The first one is to record tasks at different sampling rate each time, or to record each task once at a given sampling rate and then re-sample the voice to a

given sampling rate. For simplicity and to ensure that sampling rate is the only factor that leads to the difference among voices, since it is difficult to obtain the same word pronunciation each time the subject repeat the task with different sampling rate. Down-sampling and up-sampling are two possible approaches to change the sampling frequency of a voice. However, up-sampling a voice to a higher sampling rate will increase the computational cost and time. For this reason, we have decided to generate voices of lower sampling rates using the ideal bandlimited interpolation method proposed by [Smith et al., 1984]. The sampling rates studied and compared in this work are as follows:

Sampling rate= {44.1, 32, 24, 16, 8} kHz.

No speech enhancement is applied prior to feature extraction since we are dealing with non-stationary background noise, and it is challenging to accurately estimate the local noise spectrum and keeping the important information. Some important information that can play a good role in PD detection might be removed. For this reason, to reduce the influence of noise in PD detection we have decided to normalize the features extracted in small segments where speech and noise can be considered stationary signals as described in the following sections.

## 6.5 Feature extraction

Audio signals contain attributes that change over time, that is why it is important to estimate these attributes in (quasi)-stationary segments instead of performing a single global analysis over the whole signal. In order to obtain a quasi-stationary signal within each frame, the frame's duration should not be too large. From the other side, the frame's duration must be greater of equal to the pitch period. In speech processing, the typical range of pitch frequency is between 80 and 500 Hz [Paliwal et al., 2011]. Based on this, we decided to extract features, from preprocessed audio files, over 20 ms frames shifted by 10 ms. To reduce discontinuity between adjacent frames, a window function (Hamming) is applied to the frame prior computing the low-level descriptors (LLD). Several other features frequently used for PD classification exist and are not considered in our work such as: articulation rate, percent pause time, number of pauses, VSA, VOT and many others. The extracted features are chosen in a way to assess articulation, and phonation dimensions of speech whatever are

the language and the task under assessment. To assess phonation aspect, fundamental frequency, probability of voicing, root mean square (RMS) energy, zero crossing rate (ZCR), jitter, shimmer, and HNR features are extracted. To evaluate articulation, and according to section 6.2, several features can be extracted for this goal. In this work, we decided to study only MFCC since the cepstral analysis helps in reducing the noise existing in the speech by considering only the first coefficients (filtering), and thus keeping the important information [Khan, 2014]. The list of acoustic features extracted per frame is described in the following subsections and summarized in Table 6.1, where openSMILE toolkit is used for extraction [Eyben et al., 2010].

### 6.5.1 Zero-crossing rate and energy

Zero-crossing rate is a measure of number of times in a given time interval/frame that the amplitude of the speech signal passes through a value of zero, and it is used to decide whether a frame is voiced or unvoiced (noise-like). In general, a voiced frame has low frequency, whereas an unvoiced frame has high frequency content. In conclusion, if the zero-crossing rate is relatively high, the speech signal is unvoiced, while if the zero-crossing rate is relatively low, the speech signal is voiced [Bachu et al., 2009].

Energy is one of the powerful audio descriptors. In our work we are focusing on $E_{rms}$ the root mean square energy defined in equation (6.2) for a given signal *x(n)* since it does not depend on the speech window size. Therefore, it is commonly used in speech.

$$E_{rms} = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x^2(n)} \tag{6.2}$$

### 6.5.2 Fundamental frequency F0 and probability of voicing

In this work, F0 is only determined for voiced frames, where it is considered zero for the unvoiced frames (since it is undefined in this case). To decide whether the frame is voiced or unvoiced, ZCR method is applied. Several methods exist to compute F0. These methods are distributed into 2 categories: time domain and frequency domain [Dey, 2019]. Even

though the time domain methods are simple and inexpensive, but they are not suitable for all kind of signals. On the other hand, the frequency domain methods are considered more versatile and accurate. In this work, Subharmonic Summation method (SHS) proposed by Hermes [Hermes, 1988] is used to estimate F0. Fourier transform is applied on windows short-term speech frames, where $X_M(m)$ denote the magnitude spectrum. All the local maxima (defined by $X_{M,peak}(m)$) are detected and their positions are stored.

After that, the spectrum $X_{M,peak}(m)$ is smoothed using a symmetric 3-tap filter in the spectral domain. The smoothed spectrum is defined in equation (6.3).

$$X_{M,peak,smoothed}(m) = \frac{1}{4}X_{M,peak}(m-1) + \frac{1}{2}X_{M,peak}(m) + \frac{1}{4}X_{M,peak}(m+1) \tag{6.3}$$

The values of the spectrum on a logarithmic frequency scale are then calculated from $X_{M,peak,smoothed}(m)$, for 48 equidistant points per octave by cubic-spline interpolation. The interpolated spectrum, $X_{M,inter}(m)$ is multiplied by a raised arc-tangent function W(m). The resulting spectrum $X_{M,weighted}(m) = X_{M,inter}(m)\mathrm{W(m)}$ is then shifted by a constant factor along the octave frequency axis, scaled by a factor $h_i$ and all the scaled versions of the spectrum are summed forming what is called subharmonic summation (SHS). The estimated F0 is the value returning the maximum SHS. In order to get the probability of voicing, the arithmetic mean (μs) of the subharmonic summation spectrum is computed. For each pitch candidate $i$ with a pitch candidate score $S_i$ (defined by peak amplitude) the voicing probability is computed as follows:

$$p_{v,i} = 1 - \frac{\mu s}{S_i} \tag{6.4}$$

The probability of voicing in this work is equal to the probability of voicing of the candidate with the highest refined magnitude $S_i$, and it is not set to zero for the unvoiced frames unlike the fundamental frequency.

### 6.5.3 Jitter

Jitter describes the variation of the length of the fundamental period from one single period $T_0$to the next [Schuller, 2013]. The absolute period to period jitter or absolute local jitter is defined in equation (6.5). This definition yields one value for each pitch period.

$$J_{pp}(n') = |T_0(n') - T_0(n'-1)| \text{ , n'>2} \tag{6.5}$$

The average local jitter per frame is obtained via equation (6.6), where N refers to the number of pitch periods within the frame.

$$\overline{J_{pp}} = \frac{1}{N-1}\sum_{n'=2}^{N} |T_0(n') - T_0(n'-1)| \tag{6.6}$$

To obtain the Jitter value that is independent of the underlying pitch period length (defined by the average relative Jitter and used in this work), the average local jitter defined in (6.6) is normalized by the average pitch period length as described in (6.7).

$$\overline{J_{pp,rel}} = \frac{\frac{1}{N-1}\sum_{n'=2}^{N} |T_0(n') - T_0(n'-1)|}{\frac{1}{N}\sum_{n'=1}^{N} T_0(n')} \tag{6.7}$$

The variance of Jitter across frames (known also as Jitter of Jitter) is defined in (6.8).

$$J_{ddp} = |J_{pp}(n') - J_{pp}(n'-1)| \text{ , n'>2} \tag{6.8}$$

The average of $J_{ddp}$ over a short time frame normalized by the average period length is known by delta Jitter and defined in (6.9).

$$\overline{J_{ddp,rel}} = \frac{\frac{1}{N-2}\sum_{n'=3}^{N} ||T_0(n') - T_0(n'-1)| - |T_0(n'-1) - T_0(n'-2)||}{\frac{1}{N}\sum_{n'=1}^{N} T_0(n')} \tag{6.9}$$

In this work, both the average relative Jitter and delta Jitter defined in (6.7) and (6.9) are extracted for each frame (and not set to 0 for unvoiced frames).

### 6.5.4 Shimmer

Shimmer describes amplitude variations of consecutive voice signal periods and defined in (6.10), where *A(n')* is defined as the peak to peak amplitude [Schuller, 2013].

$$S_{pp}(n') = |A(n') - A(n'-1)| \text{ , n'>2} \tag{6.10}$$

As for Jitter, the average relative shimmer defined in (6.11) is extracted in this work for each frame (not set to 0 for unvoiced frames).

$$\overline{S_{pp,rel}} = \frac{\frac{1}{N-1}\sum_{n'=2}^{N} S_{pp}(n')}{\frac{1}{N}\sum_{n'=1}^{N} A(n')} \tag{6.11}$$

### 6.5.5 Harmonic to noise ratio (HNR)

The HNR is defined as the ratio of the energy to harmonic signal components to the energy of noise like signal components. Many methods exist to compute the HNR, but the one used in this thesis is the one proposed by [Yumoto et al., 1982]. The idea here is that the acoustic signal is composed of 2 components: a periodic wave and an additive noise with zero mean. The acoustic wave *f(t)* is defined as a concatenation of waves $f_i(\tau)$ from each period. The average of a large number of $f_i(\tau)$ (the total number of periods defined by S) will neglect the noise. The resulting average wave is represented in equation (6.12). The harmonic energy is defined in equation (6.13), where $T_i$ refers to pitch period. The second element that should be calculated in order to get the HNR is the noise energy. The noise wave at each pitch period *i* can be obtained using equation (6.14), and the noise energy is calculated according to equation (6.15).

$$f_{Aver}(\tau) = \sum_{i=1}^{S} f_i(\tau)/S \tag{6.12}$$

$$H=\int_0^{T_i} {f_{Aver}}^2(\tau)d\tau \tag{6.13}$$

$$N_i(\tau) = f_i(\tau) - f_{Aver}(\tau) \tag{6.14}$$

$$N=\sum_{i=1}^{S} \int_0^{T_i} N_i(\tau)^2 \ d\tau \tag{6.15}$$

In this work, the logarithmic HNR defined in equation (6.16) is extracted for each voiced/unvoiced frame as follows:

$$HNR_{dB} = 10log_{10}(H/N) \tag{6.16}$$

## 6.5.6 Mel-frequency cepstral coefficients

As mentioned in section 6.2, MFCC coefficients are used to detect articulators' dislocation movements. However, speech signal can be modeled as the convolution of glottal source and vocal tract (see chapter 3). In this case, we desire the vocal tract component to extract the MFCCs and detect articulation problems since the glottal source signal contain only information related to pitch and speaker traits. Since these signals are convolved, it is difficult to separate them in time domain. The technique of calculating the MFCCs is shown in Figure 6.5. FFT is applied to transform each frame from time domain into frequency domain, where the convolution is converted into multiplication. After that, the spectrum amplitude is transformed into logarithmic scale. Some researchers have found that human ear resolution of frequencies does not follow a linear scale across the audio spectrum. Based on this, Mel scale is applied in the calculation of the MFCCs to estimate the frequency perception of the human ear. The Mel scale is approached by a bank of (15 to 30) triangular filters with a lower and upper cut-off frequencies [Benba et al., 2015]. For each frequency measured in Hertz (Hz), a subjective pitch is measured on the Mel scale as follows:

$$Mel(f) = 2595 * log_{10}(1 + \frac{f}{700}) \tag{6.17}$$

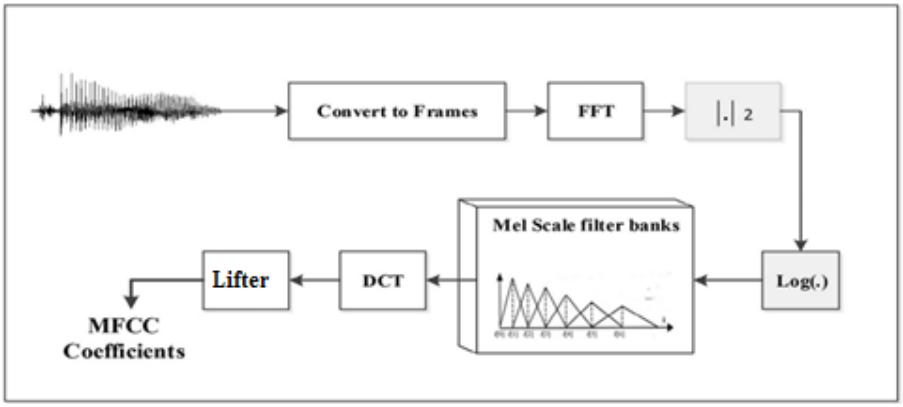



Figure 6.5. Block diagram of MFCCs extraction [Abou-Abbas et al., 2017].

After that, Discrete cosine transform (DCT) is applied on the Mel log amplitudes (defined by $m_j$) resulting in *K* Mel-cepstral coefficients and are calculated according to equation (6.18), where B refers to the number of filter banks and *i* varies from 1 to *K* [Benba et al., 2014].

$$c_i = \sqrt{\frac{2}{B}} \sum_{j=1}^{B} m_j \cos\left(\frac{\pi i}{B}(j + \frac{1}{2})\right) \tag{6.18}$$

After entering the cepstral domain, which is very similar to the time domain, two very separable components are produced: the glottal component at higher "times" (or higher order coefficients) and the vocal tract component at lower "times" (or lower order coefficients). Hence, to isolate the vocal tract component, a low-time lifter is applied. This was realized by Liftering the cepstral coefficients according to equation (6.19) where L is the Cepstral sine lifter parameter and $C_i$ represent the MFCC coefficients [Benba et al., 2014].

$$C_i = c_i\left(1 + \frac{L}{2}\sin\left(\frac{\pi i}{L}\right)\right) \tag{6.19}$$

In this work, the number of filters in the filter bank is set to 26 with lower cut-off frequency of 20 Hz (to avoid the influence of DC components) and upper cut-off frequency of Fs/2. The Cepstral sine lifter parameter L is set to 22. The first 14 MFCCs coefficients are extracted for both voiced and unvoiced frames in this work since most of the signal

information is represented by the first few MFCC coefficients. Typically the number of MFCCs coefficients is chosen between 12 and 16 for most speech tasks.

Table 6.1. Frame-level acoustic features extracted.

| Number | Feature | Category |
|---|---|---|
| 1. | RMS energy | Phonation aspect |
| 2. | ZCR | Phonation aspect |
| 3. | Fundamental frequency | Phonation aspect |
| 4. | Probability of voicing | Phonation aspect |
| 5. | Average Local Jitter | Phonation aspect |
| 6. | Delta Jitter | Phonation aspect |
| 7. | Average Local Shimmer | Phonation aspect |
| 8. | Logarithmic HNR | Phonation aspect |
| 9. | Mfcc [1-14] | Articulation aspect |

The described features above are defined by static features, since they depend only on the information existing in the frame, and can be represented as a vector of 22 LLD feature vectors $l_i(n)$ with $i$=1,…..,22 as follows:

$$L=\begin{bmatrix} l_1(n) \\ l_2(n) \\ \dots \\ l_{22}(n) \end{bmatrix}$$

Signal dynamics beyond the frame boundaries are not captured. In addition, short time analysis creates artifacts due to the artificial discontinuity happened in the waveform. To reduce the artifacts and to get dynamic features (containing dynamic information beyond the frame boundaries), we have decided to use the method proposed by Schuller [Schuller et al., 2006], where a moving average filter of window $W$ (must be old value) is applied to obtain the average features over a small number of neighboring frames. The smoothed low level feature vectors are expressed as follows:

$$l_{i,smo}\,(n) = \frac{1}{W} \sum_{j=-(W-1)/2}^{(W-1)/2} l_i(n+j) \tag{6.20}$$

We have decided to use small window size for smoothing to capture the short time-dynamics of the signal. Here in this work we assumed $W$=3. $k^{th}$ order Delta regression proposed by [Young, 1997] is another technique that can be applied to convert static features to dynamic features. In this work, we have decided to get the first order delta regression coefficients of the smoothed LLDs defined in (6.21) as follows:

$$l_{i,smo,delta}\ (n) = \frac{\sum_{j=1}^{W1} \mathrm{j}.(l_{i,smo}(\mathrm{n+j}) - l_{i,smo}(\mathrm{n-j}))}{2\sum_{j=1}^{W1} j^2} \tag{6.21}$$

where the window length $W1$ is equal to 2 in this work.

The total number of frame-level features extracted is equal to 44; where 22 of them refer to the smoothed LLDs, and the other 22 represent the 1st order delta regression coefficients defined in equations (6.20) and (6.21).

It is well known that the LLD feature vector length differs from subject to another. In order to get the same length for all the subjects, statistical functions such as mean, maximum, minimum, median and STD are calculated. This resulted in 220 global features that describe each subject's speech characteristics over the entire task. Many speaker level post-processing steps are applied in order to obtain the global features from the frame-level features. First of all, the acoustic signal contains a lot of variability, where some of them are related to PD disease, and some other variations are related to recording environment noise, speaking style or accents, etc. To minimize the effects of variations that are not related to the disease, and to obtain a noise and language–robust model, z-scored normalization is applied across the LLD features of each subject separately in order to obtain time series of zero mean and STD of one. In addition, as mentioned before, for each subjects the frames can be either voiced or unvoiced. In this work, only voiced sounds and the combination of both voiced and unvoiced frames are studied separately to check whether the unvoiced segments play a role in PD detection or not. For the voiced frames, all the 44 LLD features mentioned before are considered, where for the unvoiced frames the fundamental frequency, jitter and shimmer are not taken into consideration since they are equal to zero. Finally, the mean, maximum, minimum, median, and STD statistics are applied on each LLD feature separately. This resulted in 220 global features that describe each subject's speech characteristics over the entire task. In conclusion, the 220 global features are extracted for each task (sustained vowel 'a', and text reading) and the mean global features vector across the different tasks is obtained as shown in Figure 6.6.

## 6.6 Feature selection

The two-stage feature selection approach defined in chapter 5 is also applied here to remove the irrelevant features, where the first stage consists of a pure statistical analysis of the data and the second stage consists of applying a suboptimal approach that provides a kind of benchmark of the relevance of the features in the desired task. In the first stage, statistical tests are applied on each feature to validate if the underlying processes for PD and HC subjects are independent. To decide whether the sample features distribution is normal or not, Shapiro-Wilk test is applied for this purpose. Based on the results obtained with Shapiro-Wilk tests, multiple independent student t-test and Mann-Whitney tests are applied to normally and not normally distributed features respectively. Feature selection based on statistical tests with sequence of alpha values between 0 and 1 were tested on each of the 2 tasks and "All-tasks" separately as shown in Figure 6.6, and the one with the best validation accuracy was picked.

The set of selected features by statistical tests for "All-tasks is further reduced to a smaller subset of features in the second stage using the suboptimal approach also described in chapter 5. Also in this work, an SVM classifier is used for selection as well as for detection.

## 6.7 Classifier used

SVM with RBF kernel is applied also in this work for binary classification, where a grid search using cross-validation was applied on the RBF parameters C and $\gamma$ in order to identify the best values. Exponentially growing sequences of C and ɤ are used (C = $2^{-30}$, $2^{-29}$, ……., $2^{29}$,$2^{30}$ and $\gamma$= $2^{-30}$, $2^{-29}$, ……., $2^{29}$,$2^{30}$).

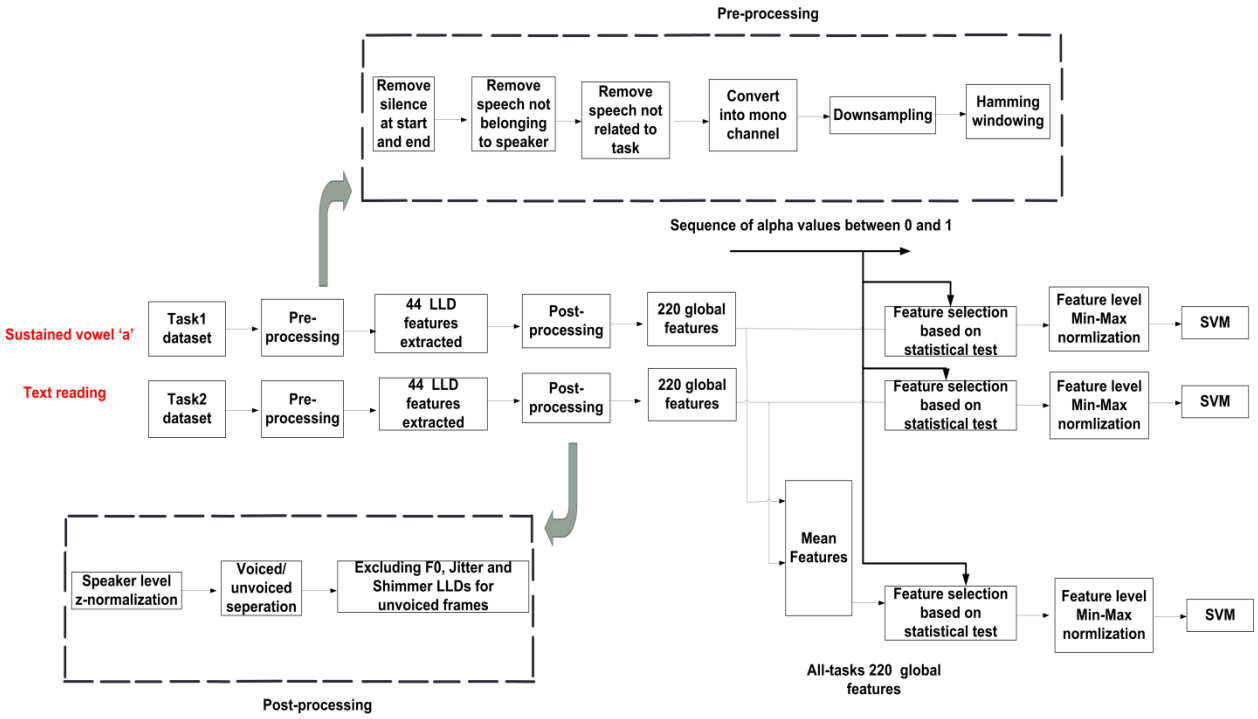



Figure 6.6. Feature selection based on statistical tests overview.

Due to the small data we worked on, the set with 42 subjects is divided into 3 folds, with the 66.66/33.33 % (training/validation) proportion using stratified sampling method. Sequentially, one fold is validated using the classifier trained on the remaining 2 folds. The total accuracy is obtained by calculating the mean of all the folds accuracies.

In this work, we decided not to use a separate test set due to a low database size. As a result the validation set can be considered as test set. Before classifying, feature-level Min-Max normalization is applied on each feature separately in order to avoid the dominance of features with greater numeric ranges (as shown in Figure 6.6). Equation (5.40) is applied for scaling, where upper and lower refer to the upper and the lower values of the scaling range (here the upper value is 1 and the lower value is -1), and minimum and maximum refer to the minimum and maximum feature values in the sample. The scaling factors (minimum and maximum) are obtained from training data and used to scale the test data.

## 6.8 Experiments and results

In this work, we investigate the effect of sampling rate and unvoiced sounds on classification performance of PD early detection through voice analysis. The same procedure

described above is performed for each sampling rate. First, the 44 low level features defined in section 6.5 are extracted over 20 ms frames on preprocessed audio files, where hamming window is applied to reduce the discontinuity between adjacent frames. Second, speaker level z-normalization is applied to reduce the effects of variations that are not related to the disease. The global features are calculated based on frame type; for voiced frames the 5 statistical functions are calculated for each of the 44 LLDs, where for unvoiced frames the 5 statistical functions are calculated for all the 44 LLDs except F0, jitter and shimmer. In this thesis, we decided to study the voiced frames alone and the combination of both voiced and unvoiced frames to validate that PD also produce abnormal unvoiced sounds, which can play an important role in classification. As a result, a 220 global features vector is obtained for each subject describing the speech characteristics over the entire task.

In the first stage, the effect of sampling rate and unvoiced sounds on the prediction performance for each of the 2 voice tasks is studied, where only statistical tests are applied for feature selection. The prediction performance is evaluated in term of the accuracy, sensitivity, and specificity defined in chapter 5. The numerical results achieved by the SVM classifier with 3 folds cross-validation using only statistical tests for feature selection are presented in Table 6.2, and the effect of sampling rate on classification accuracy is illustrated in Figure 6.7. To begin with, according to the results obtained it is noted that the sampling rate affect PD detection differently when different features and tasks are applied. For sustained vowel 'a' and voiced frames of text reading, it is observed that 44.1 KHz sampling rate returns the lowest performance, and as we decrease the sampling rate the performance starts to increase to reach the highest performance at 24 KHz sampling rate, before starting to decrease again. However, for the voiced and unvoiced frames of Text reading the highest performance is reached at 16 KHz. In addition, it is found that the studying and analyzing both voiced and unvoiced sounds in a connected speech is more effective than studying only the voiced sounds for PD detection. As it was expected, PD detection through voice analysis performance depends on sampling rate since two signals of different sampling rates provides different features even though both signals refer to the same speech, and unvoiced sounds play a an important role in PD detection. These findings can be explained by the fact that the signal with low sample rate (for example 8 KHz) cannot be able to save all of the voice characteristics that are needed for PD detection, so the characteristics that are extracted does

not fit the needed characteristics in a speech. In addition, some researchers [Stevens, 1998], [Ladefoged, 2003] found that the highest linguistically meaningful frequencies are below 11 KHz. Based on this, a sampling rate close to 22 KHz (twice the highest frequency) will be sufficient. The best sampling rates obtained in this work (24 KHz and 16 KHz) confirm the conclusion built in [Stevens, 1998], [Ladefoged, 2003]. From the other side, we found that it is not always a good idea to use a high sampling rate because it may tends to reduce the model performance and increase the processing cost.

Based on these findings, the next step is to form single feature vector combining information of both tasks: the sustained vowel 'a' and the combination of voiced and unvoiced sounds in the text reading. The global features vector (of size 220) is extracted for each task separately, where the sustained vowel 'a' is sampled at 24 KHz while the text reading is sampled at 16 KHz (based on the results obtained in Table 6.2). The mean feature vector is then obtained and the two stages feature selection approach is applied for "All-tasks". Statistical tests reduce the total number of features (for the two tasks and the "All-tasks") from 660 to 76 features. The number of selected features per task using statistical tests, the significance level with the best validation accuracy, and the numerical results achieved by the SVM classifier with 3 folds cross-validation using only statistical tests for feature selection are presented in Table 6.3.

For "All-tasks" the suboptimal incremental approach defined in chapter 5 is applied to select the most relevant features between the 19 selected in the first stage. The highest classification accuracy obtained at every iteration of the suboptimal approach is shown in Figure 6.8. The highest classification accuracy obtained is 97.62 % for N=16 features. The "All-tasks" performance obtained with one and two stage feature selection methods are shown in Table 6.4. It is clear that a smaller subset of features gave better classification accuracy. This indicates clearly that some of the features disturb the performance of the prediction system, mainly because of the curse of dimensionality, a critical factor in this particular case where a limited amount of training data is available. The 16 selected features providing the best "All-tasks" performance are listed in Table 6.5.

Table 6.2. The effect of sampling rate and unvoiced frames on the Performance of each task in PD classification using statistical tests for feature selection.

| Sampling rate (KHz) | Sustained vowel "a" | Text reading Voiced and unvoiced frames | Text reading Voiced frames |
|---|---|---|---|
| 44.1 | Acc: 80.95<br>Sens: 80.95<br>Spec: 80.95 | Acc: 92.86<br>Sens: 100<br>Spec: 85.71 | Acc: 90.48<br>Sens: 90.48<br>Spec: 90.48 |
| 32 | Acc: 88.1<br>Sens: 95.24<br>Spec: 80.95 | Acc: 90.48<br>Sens: 90.48<br>Spec: 90.48 | Acc: 92.86<br>Sens: 95.24<br>Spec: 90.48 |
| 24 | Acc: **92.86**<br>Sens: 90.48<br>Spec: 95.24 | Acc: 95.24<br>Sens: 95.24<br>Spec: 95.24 | Acc: **95.24**<br>Sens: 95.24<br>Spec: 95.24 |
| 16 | Acc: 85.71<br>Sens: 76.19<br>Spec: 95.24 | Acc: **97.62**<br>Sens: 100<br>Spec: 95.24 | Acc: 92.86<br>Sens: 95.24<br>Spec: 90.48 |
| 8 | Acc: 90.48<br>Sens: 95.24<br>Spec: 85.71 | Acc: 95.24<br>Sens: 100<br>Spec: 90.48 | Acc: 88.1<br>Sens: 95.24<br>Spec: 80.95 |

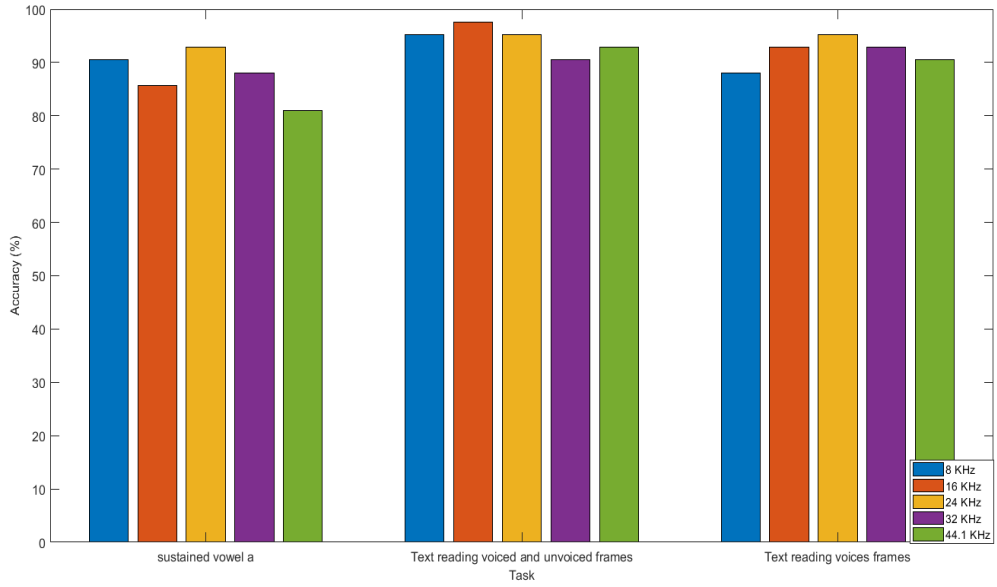



Figure 6.7. The influence of sampling rate on classification accuracy.

Table 6.3. Performance and number of selected features of each task in PD classification using statistical tests for feature selection.

| **Task** | **Accuracy (%)** | **Sensitivity (%)** | **Specificity (%)** | **Significance Level** | **# of selected features** |
|---|---|---|---|---|---|
| Sustained vowel 'a' | 92.86 | 90.48 | 95.24 | 0.0624 | 29 |
| Text reading<br>(voiced and unvoiced sounds) | 97.62 | 100 | 95.24 | 0.0279 | 28 |
| All | 95.24 | 90.48 | 100 | 0.0333 | 19 |

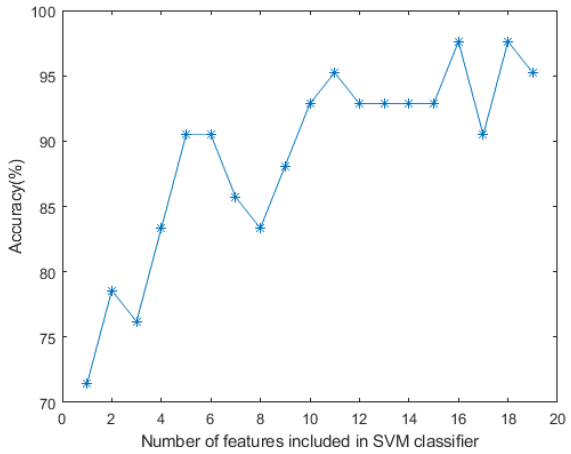



Figure 6.8. The highest classification accuracy obtained at every iteration of the suboptimal approach.

Table 6.4. Table of comparison between the "All-tasks" performance obtained with one and two stage feature selection methods.

| **Performance** | **1 stage Feature selection** | **2 stages Feature selection** |
|---|---|---|
| Accuracy (%) | 95.24 | 97.62 |
| Sensitivity (%) | 90.48 | 95.24 |
| Specificity (%) | 100 | 100 |

Table 6.5. The two stages selected features providing the best "All-tasks" performance.

| **Feature** | **Statistic** |
|---|---|
| Smoothed average local shimmer | Minimum |
| Smoothed RMS energy | Maximum |
| 1st smoothed MFCC coefficient | STD |
| 2nd smoothed MFCC coefficient | Minimum |
| 6th smoothed MFCC coefficient | STD |
| 7th smoothed MFCC coefficient | Median |
| 12th smoothed MFCC coefficient | Minimum |
| 12th smoothed MFCC coefficient | Median |
| 13th smoothed MFCC coefficient | Mean |
| 14th smoothed MFCC coefficient | Minimum |
| 1st order delta regression coefficient of the smoothed probability of voicing | Mean |
| 1st order delta regression coefficient of the 2nd smoothed MFCC coefficient | Median |
| 1st order delta regression coefficient of the 4th smoothed MFCC coefficient | STD |
| 1st order delta regression coefficient of the 5th smoothed MFCC coefficient | Mean |
| 1st order delta regression coefficient of the 9th smoothed MFCC coefficient | Minimum |
| 1st order delta regression coefficient of the 14th smoothed MFCC coefficient | Median |

# 6.9 Conclusions

The main goals of this part of the thesis are to build a language and task-independent acoustic feature set for assessing the motor disorders in PD patients, and to study the influence of sampling rate and unvoiced sounds on the performance. For this purpose, two different voice tasks taken from SpeechPDMuliMC subset are used, where this subset

includes speeches in three different languages: Arabic, French and English. Only phonation and articulation features that can be extracted for all the tasks under assessment are studied. The prosody features are excluded since they are related to intonation, stress, and rhythm, which are depending on the language spoken. Before extracting the low level features, many pre-processing steps are applied to the signal to ensure we get an efficient system. To minimize the effects of variations that are not related to the disease, and to obtain a noise and language–robust model, speaker level z-scored normalization is applied across the LLD features. In order to get the same global features vector length for all the subjects, statistical functions such as mean, maximum, minimum, median and STD are calculated.

To avoid the risk of falling in a curse of dimensionality, a two-stage feature selection approach is applied to remove the irrelevant features while keeping features that are necessary and sufficient to describe the target concept. Due to the small data we worked on, a 3-fold cross validation SVM classifier with RBF model was used for binary classification. We have succeeded to build a language-independent model for PD diagnosis through voice analysis with 97.62 % accuracy, 95.24 % sensitivity, and 100 % specificity.

A number of observations and conclusions are obtained in this work. Down-sampling a signal to a lower sampling rate will filter out high frequencies components. The high-frequencies components may include recording environment noise, tremor, or breathy voice related to the incomplete vocal fold closure (see chapter 6). First of all, it is noted that the sampling rate affects PD detection differently when different features and tasks are considered. We have found that signals with low sampling rate (lower than 16 KHz) can lose valuable information that can play a good role in PD detection, where a sampling rate of 24 KHz for sustained vowel 'a' task and text reading task (voiced sounds), and 16 KHz for text reading task (voiced and unvoiced sounds) are considered enough and provided more than enough information for the features analyzed here. From the other side, it was also found that high sampling rate may also reduce the model performance. The best sampling rates found in this work confirm with the conclusion that the highest linguistically meaningful frequencies are below 11 KHz. In addition, as we were expected, we found that not only the voiced sounds in a voice or speech contain important information about the disease, but also the unvoiced sounds may show up the existence of the disease. We believe that this can be explained by the idea that abductor spasms (defined by involuntary movements cause the

vocal fold to open) occurs mostly on unvoiced consonants. Finally, most of the selected features refer to the MFCCs coefficients. From a clinical point of view, PD affects the movement of the articulatory muscles (jaw, tongue, and lips), resulting varying energy in the frequency bands of speech signal. The MFCCs coefficients can effectively quantify the problems in speech articulation since they compute the energy differences between the bands of speech frequency which can be used to discriminate between the varying energy levels of disturbed resonances [Khan, 2014]. This can explain the frequent existence of MFCCs features in the selected set.

To the best of our knowledge and based on literature, there are several works considering different bio-signals to assess motor impairment of PD patients, most of these studies consider only one modality. Multimodal analyses (considering information from different sensors) have not been extensively studied. Although many improvements have been shown in several tasks, there is still an absence of a multimodal fusion system able to deliver an accurate prediction of PD disease. This has motivated us to build a multimodal system for PD early detection based on voice and handwriting signals that will be discussed in the next chapter.

# 7 Multimodal System for Early Parkinson's Disease Detection based on Handwriting and Speech

Based on the reviewed literature, a language-independent model to detect PD using multimodal signals has not been enough addressed. The main reason of focusing on multimodal analysis in this thesis is that there is no consensus on which aspect (handwriting, speech) is more appropriate to help on PD diagnosis in early stages; so combining and analyzing both handwriting and speech signals may deliver a more accurate PD prediction. In this chapter, two different learning approaches are applied: feature-based and deep learning approaches. These approaches are detailed in the following sections. Both HandPDMultiMC and SpeechPDMultiMC datasets taken from our PDMultiMC database are used, where the PD patients are also studied in their "on-state". The aim here is to extract important information from both modalities forming a multimodal vector that will be used for classification. Fusion of different modalities can be executed at different levels: data level, short term features level, global feature level, or decision level as shown in Figure 7.1 [Dumas et al., 2009]. In this work, only global features and decision level fusion are applied since data-level and short term feature-level fusions are used when the multiple raw data even come from a same type of modality source, or are synchronized (which is not our case) [Dumas et al., 2009].

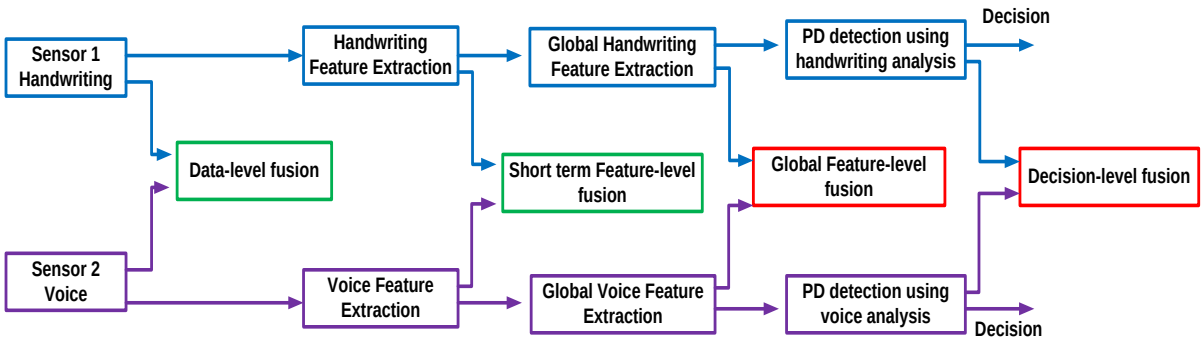



Figure 7.1. Levels of multimodal fusion of handwriting with speech.

## 7.1 Multimodal assessment of Parkinson's disease: feature-based approach

Previously, handwriting and voice were analyzed separately where language and task-independent pre-engineered global features were extracted from raw signals in a manner to assess the motor symptoms of PD. An excellent classification accuracy of 97 % was reached. The seven handwriting tasks and the two speech tests that are described in chapter 4 will be used to build the multimodal system. In this section, only global feature-level fusion is applied. Global feature-level fusion consists in combining two global features vectors, one for each modality. Each vector includes information of all tasks per subject. For each modality, the set of global features defined in previous chapters are extracted for each task then combined together to form a single feature vector. Two different methods are applied to combine the feature vectors in each modality: the first method will calculate the average feature vector across the different tasks, where the second method will concatenate the features vectors together as shown in Figure 7.2. At a later stage, the global features vectors obtained from the 2 modalities will be concatenated to form a multimodal global features vector that will be used to detect PD using an SVM model. An overview of the SVM model trained on multimodal pre-engineered features is shown in Figure 7.2.

The 189 'on-paper' global handwriting features and the 220 global voice features defined in chapters 5 and 6 are extracted for each handwriting and speech task forming even a 409 or 1763 multimodal global features vector as shown in Figure 7.2. Based on the results obtained in chapter 6, the sustained vowel 'a' is sampled at 24 KHz while the text reading is sampled at 16 KHz and both voiced and unvoiced frames in the text reading are studied. The two stages feature selection described in section 5.1.2 is also applied here to select from the multimodal features the most relevant ones; where a sequence of alpha values between 0 and 1 were tested and the one with the best validation accuracy was picked. The selected features are then used to classify PD and HC subjects using SVM model with RBF kernel.

For voices, all the pre-processing steps described in chapter 6 are also applied to the voice signal before extracting the LLD features, and speaker level z-normalization is applied to the LLD features to reduce the effects of variations that are not related to the disease (such

as recording environment noise, speaking styles or accent etc.). Later on, statistical functions (mean, maximum, minimum, median, and STD) are applied to obtain the 220 global features vector. SVM with RBF kernel is used for binary classification. Feature-level Min-Max normalization is applied on each feature separately before classification. Equation (5.40) is applied to scale features into range [-1, 1].

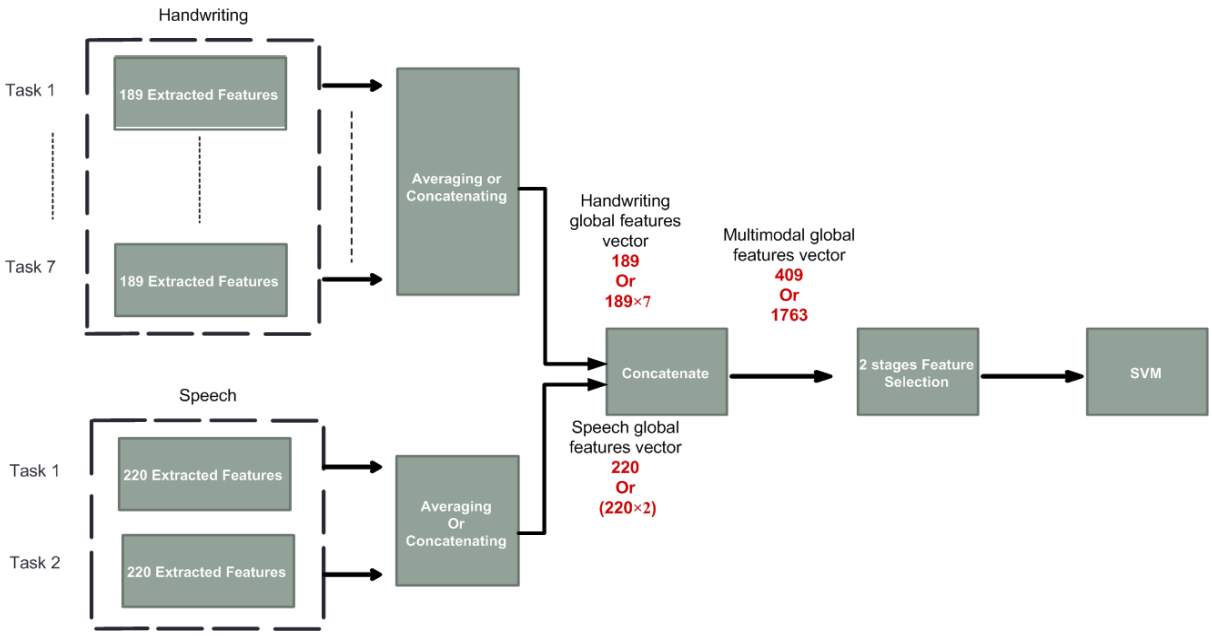



Figure 7.2. SVM model trained on multimodal pre-engineered features using global feature-level fusion where two methods are applied to combine modalities' features vectors: averaging or concatenation.

## 7.2 Multimodal assessment of Parkinson's disease: deep learning approach

In recent years deep learning have been successfully implemented to evaluate specific phenomena in speech, including the detection and monitoring of PD [Frid, et al., 2016], [Gunduz, 2019], [Caliskan et al., 2017]. To benefit from this, and due to the ability to extract features in an automatic way, deep learning approaches are also applied in this work and compared to the SVM model. As mentioned previously, our focus in this thesis is to build a multilingual model for PD early detection using multimodal signals. The core idea is to train our model with language-independent multimodal feature vector. For the SVM model that is

trained on pre-engineered multimodal features and described in section 7.1, the extracted features are chosen in a way to be language-independent as described in chapter 5 and 6. For deep learning approach, to obtain language-independent feature vector, the model is trained on all the languages so the features will not be biased toward a specific language. One of the deep learning models proposed in chapter 5, the 2D CNN model with spectrogram images (summarized in Figure 5.49), is applied in this work in both modalities. However, for the 1D CNN-BLSTM model with raw time series as input (presented in Figure 5.54), it is only applied here in handwriting modality, where we decided to crop the audio signal into segments of fix length and apply the 1D CNN-MLP model (shown in Figure 7.5) instead of the 1D CNN-BLSTM to overcome memory usage problem since we are dealing with very long signals and BLSTMs are more memory consuming than the MLPs.

Also in this work, we have studied the whole handwriting dynamic signals so we can extract both in-air and on-surface features. The handwriting and audio pre-processing steps mentioned in chapters 5 and 6 are also applied here (getting the same writing direction, normalizing the X and Y coordinates, removing silence at the start and the end of the speech, removing speech that does not refer to the subject and each spontaneous intervention introduced by the subject that was not directly related with the task, converting the 2 channels signal into mono signal, and down-sampling signal rate). To reduce computational time, cost and memory usage, voice sampling rate for both sustained vowel 'a' and text reading (voiced and unvoiced frames) are set to 8 KHz and not to the ones found in chapter 6. In addition, for all our models, 2D CNNs, 1D CNN-BLSTMs and 1D CNN-MLPs, all images and raw time series are normalized to the range (0, 1) using min-max normalization. In this work, Xavier normalized initialization (defined in section 5.3.1.3.1) is applied to initialize the weights of all the MLPs and CNNs, where biases are zeros initialized. Where for the BLSTMs, the 16 weights and the 8 biases are zero initialized, where the output is multiplied by a weight matrix, and added to a bias vector that are drawn from random normal distributions.

## 7.2.1 Convolutional neural network

Spectrogram 2D representation for both raw handwriting signals and voice signals are obtained by applying STFT. Blackman windowing function is applied, where both the window length and the number of FFT point are set to 256 and the overlapping rate is 50 %.

Lanczos technique is used to ensure that the number of input feature maps is identical for all subjects by resizing the spectrogram images to 64×64 pixels resolution. The 2D CNN model represented in Figure 5.49 in chapter 5 can be used for classification from a single image (voice case), or classification from k measurements, where each measurement is encoded into spectrogram image (handwriting case). The number of handwriting dynamic signals k is a hyper-parameter varying between one and seven. Fusion of both handwriting and voice modalities are executed here at two levels: global feature-level and decision-level.

#### 7.2.1.1 Multimodal assessment using 2D CNN and global feature-level fusion

The aim here is to form one feature vector with information of all tasks per subject and per bio-signal. To do this, for each modality individual 2D CNNs are trained per task as shown in Figure 7.3. For each modality, the feature maps obtained by the convolutional layers for each task are combined together whether by averaging or by concatenating as shown in Figure 7.3. For each modality, one feature vector is obtained. The embeddings obtained from the 2 modalities are concatenated to form a multimodal vector per subject. The created feature vectors are then used to classify PD patients and HC subjects using a hidden layer with 23 nodes (obtained using the empirical rule defined in equation (5.64)) and ReLU activation function and an output layer with 2 nodes (binary classification PD or HC) and softmax activation function.

#### 7.2.1.2 Multimodal assessment using 2D CNN and decision-level fusion

In this section, for voice modality, two cases are studied: the first case (defined by case 1) is when spectrogram is obtained for the whole audio signal, and the second case (defined by case 2) is when the audio signal is cropped into segments of fix size in order to increase the number of samples, and to keep the nonlinear variation over the time axis (since we do not need any more to normalize the time series into fixed dimension image). We decided to crop the audio signal into segments of size 4s because in [Ma et al., 2016] it was found that this length happens to be the best length to crop the signals. In the second case, spectrogram is obtained for each segment separately. All the training segment slices images are considered independent training instances. Segmentation is also applied when predicting

the label of a testing time series. No window slices referring to the same participant exist in training and test. To make the final prediction for each subject in the test set, a BLSTM is applied where the S probability vectors outputs of the 2D CNN models are considered as a multivariate sequence of length S, and are used as input to a dynamic BLSTM to decide the final prediction.

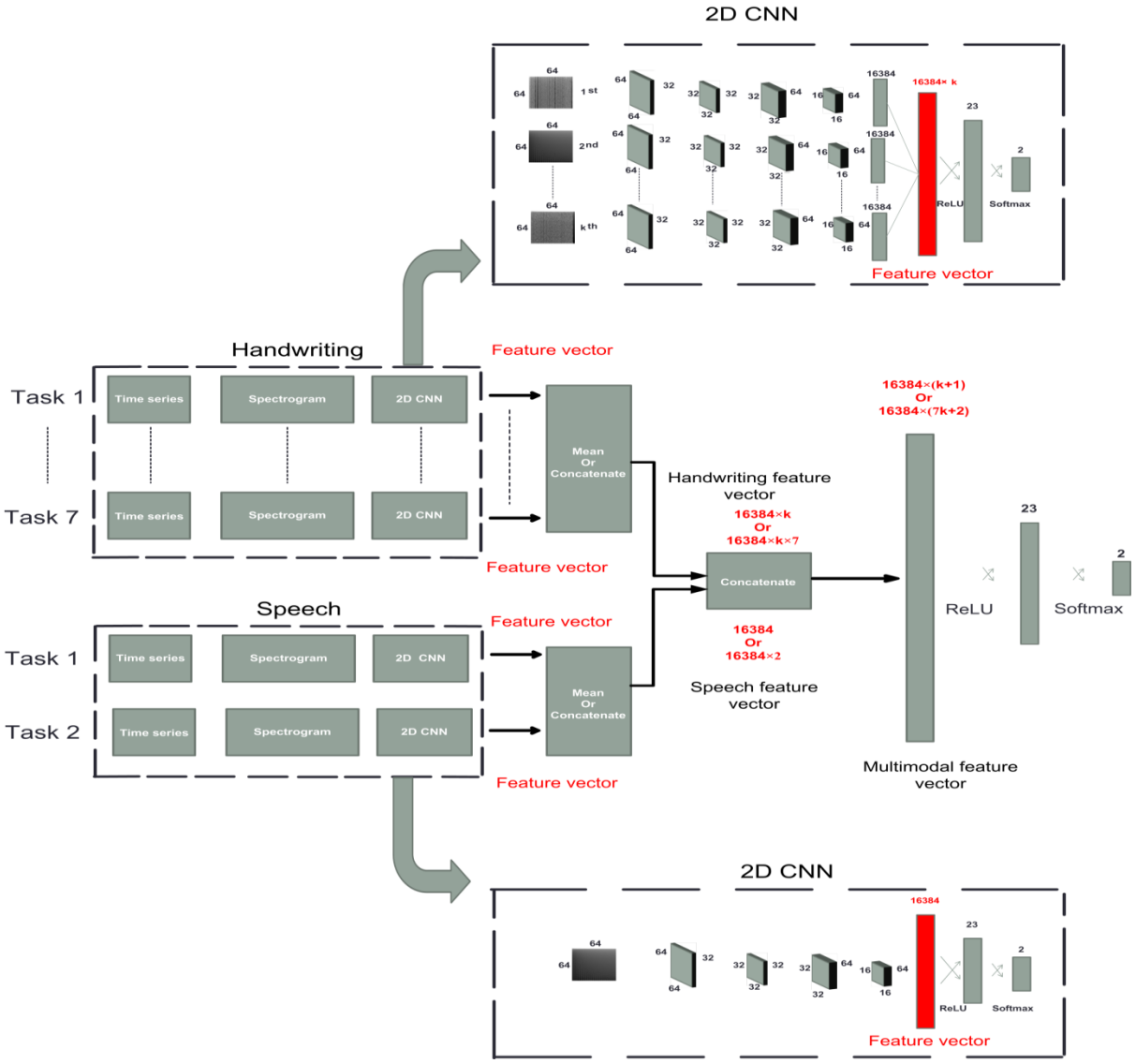



Figure 7.3. Multimodal assessment using 2D CNN models and feature-level fusion.

Individual 2D CNNs are trained for each task in the modality as shown in Figure 7.4. Two MLPs models (MLP1 and MLP2) are applied, where each one is used to combine the probability vectors (each of size 2) obtained by all tasks in each modality. At a later stage, another MLP model (MLP3) is used to combine the probability vectors provided by each of

MLP1 and MLP2 (each of size 2) in order to get the final prediction. For each MLP or 2D CNN model, the number of hidden nodes is chosen by using the empirical rule described in equation (5.64), where for BLSTM model, the number of hidden nodes is selected using equation (5.65).

## 7.2.2 Combination of 1D CNN-BLSTMs and 1D CNN-MLPs

As mentioned before, the 1D CNN-BLSTM (represented in Figure 5.54), and 1D CNN-MLP (represented in Figure 7.5 where in this case n is equal to 4×audio sampling rate and the number of hidden nodes is chosen by using the empirical rule described in equation (5.64)) are both applied here with handwriting and voice modalities respectively; where the raw signals are used directly. The number of handwriting dynamic signals k is a hyper-parameter varying between one and seven. Fusion of both handwriting and voice modalities are executed here at decision-level only since the feature maps in both modalities differ (the features map in handwriting modality is a sequence of length n/4 of vectors of size 32, where in voice modality the features map is a vector of length 32×n/4), in addition in each modality n varies from task to another and the number of audio segments of length 4s varies between tasks.

### 7.2.2.1 Multimodal assessment using the combination of 1D CNN-BLSTM and 1D CNN-MLP models and decision-level fusion

Individual 1D CNN-BLSTMs are trained for each task in handwriting modality, where individual 1D CNN-MLPs followed by BLSTMs to make the final prediction are trained for each task in voice modality as shown in Figure 7.6. We have ensured that there's no window slices referring to the same participant exist in training and test. Two MLPs models (MLP1 and MLP2) are applied, where each one is used to combine the probability vectors (each of size 2) obtained by all tasks in each modality. At a later stage, another MLP model (MLP3) is used to combine the probability vectors provided by each of MLP1 and MLP2 (each of size 2) in order to get the final prediction. Also for each MLP, the number of hidden nodes is chosen by using the empirical rule described in equation (5.64), where for each BLSTM the number of hidden nodes is selected using equation (5.65).

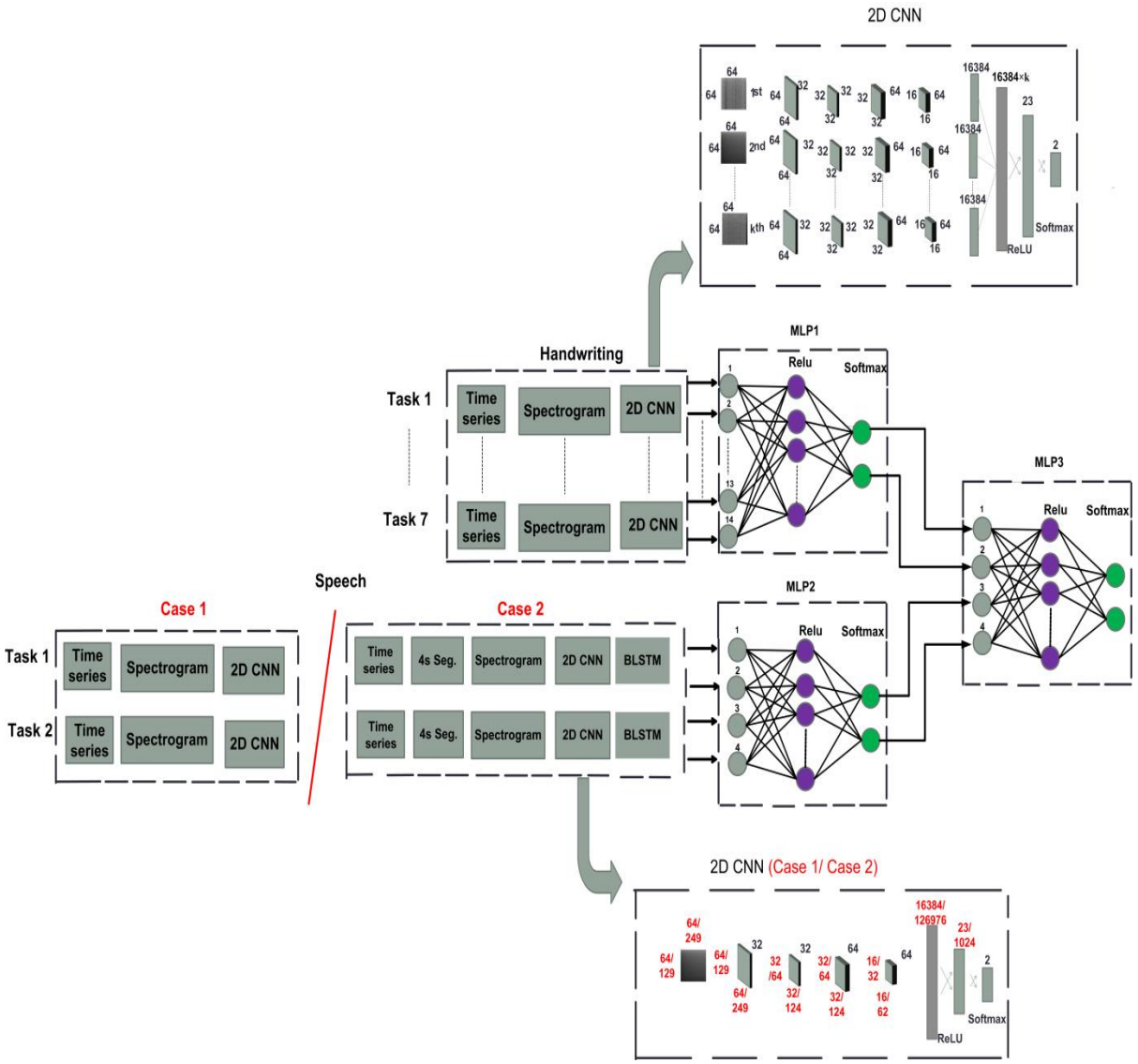



Figure 7.4. Multimodal assessment using 2D CNN models and decision-level fusion.

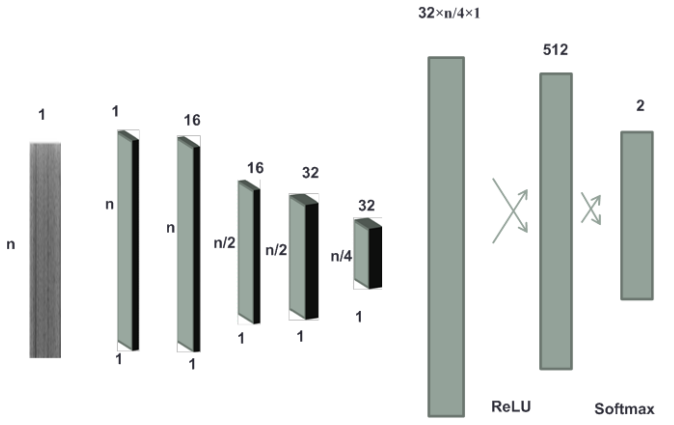



Figure 7.5. Single-task 1D CNN-MLP architecture on audio time series.

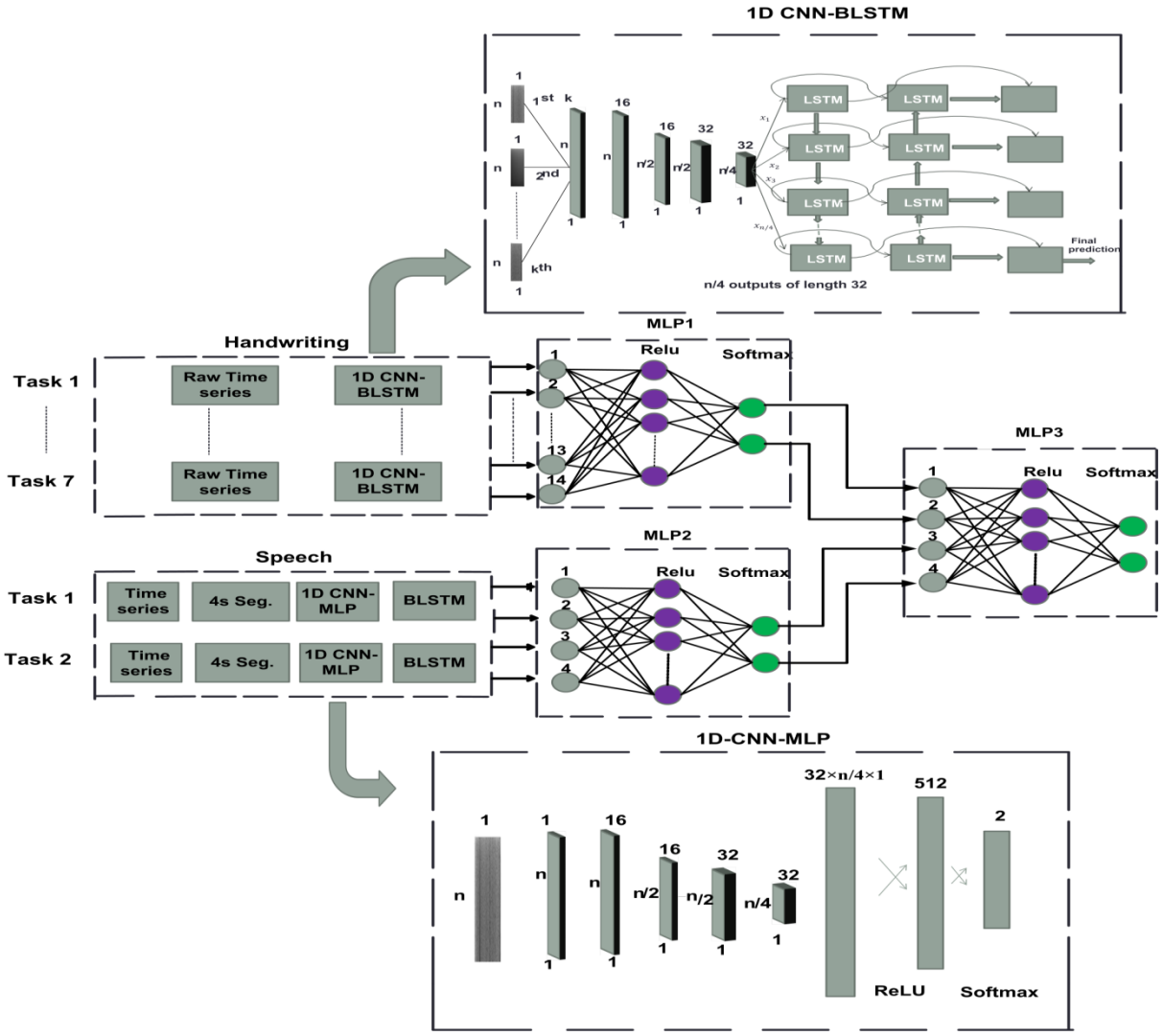



Figure 7.6. Multimodal assessment using 1D CNN-BLSTM and 1D CNN-MLP models and decision-level fusion.

## 7.2.3 Combination of 1D CNN-BLSTMs and 2D CNNs

The 1D CNN-BLSTMs and 2D CNNs models are combined in this section. For voice, 2D CNNs are applied with spectrogram as input, where for handwriting 1D CNN-BLSTMs are used with the raw signals as input. Also fusion of both handwriting and voice modalities are executed here at decision-level for the same reason mentioned in section 7.2.2, and the number of handwriting dynamic signals k is a hyper-parameter varying between one and seven.

### 7.2.3.1 Multimodal assessment using the combination of 1D CNN-BLSTM and 2D CNN models and decision-level fusion

In this section also, for voice modality, spectrogram is even obtained for the whole audio signal, or for each 4s segment as described in section 7.2.1.2. Individual 1D CNN-BLSTMs and 2D CNNs are trained for each task in handwriting and voice modalities respectively as shown in Figure 7.7, where in case of audio segmentation the 2D CNN is followed by BLSTMs to make the final prediction, and no window slices referring to the same participant exist in training and test.

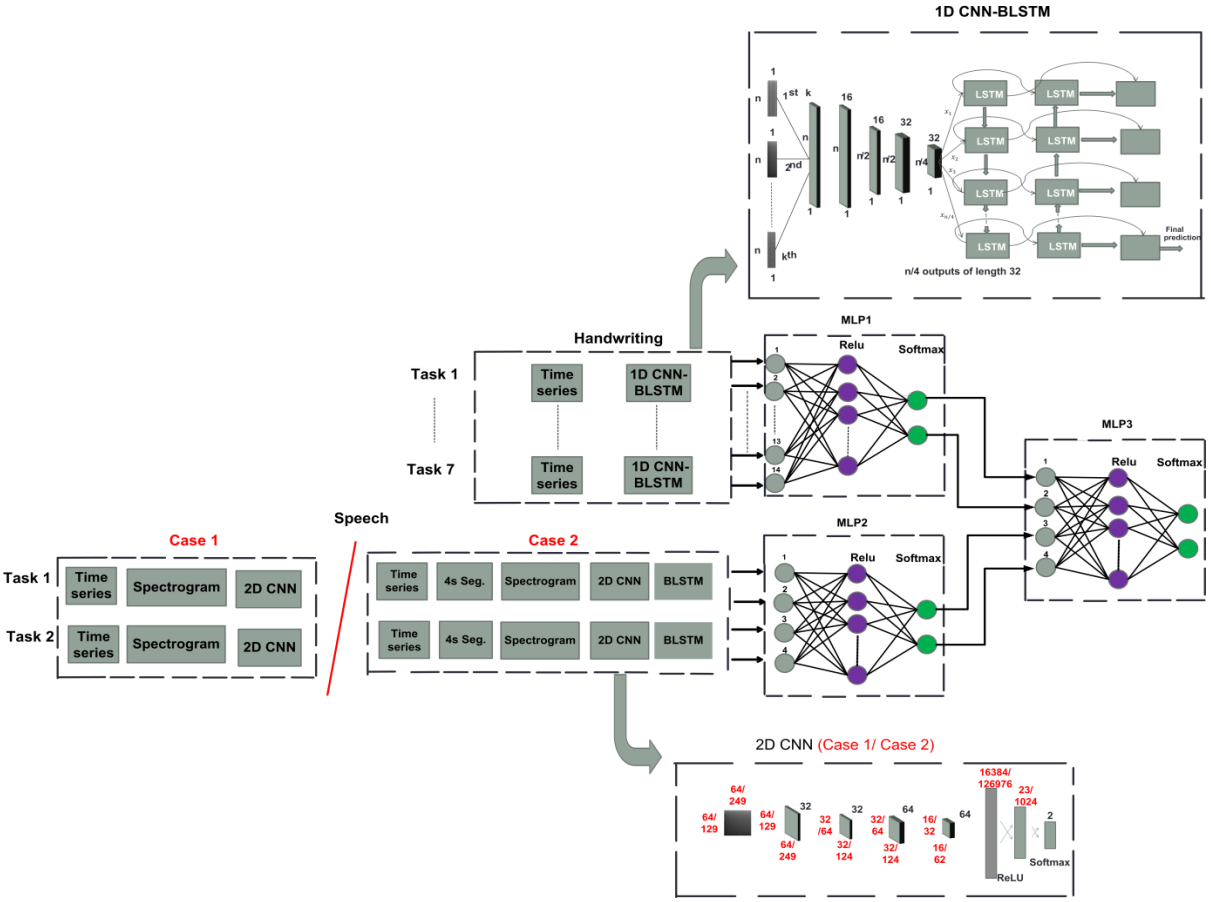



Figure 7.7. Multimodal assessment using the combination of both 1D CNN-BLSTM and 2D CNN models and decision-level fusion.

Two MLPs models (MLP1 and MLP2) are applied, where each one is used to combine the probability vectors (each of size 2) obtained by all tasks in each modality. At a later stage, another MLP model (MLP3) is used to combine the probability vectors provided by each of

MLP1 and MLP2 (each of size 2) in order to get the final prediction. For each 2D CNN or MLP model, the number of hidden nodes is chosen by using the empirical rule described in equation (5.64), where for the BLSTM the number of hidden nodes is chosen using equation (5.65).

## 7.3 Data augmentation

Based on the conclusions obtained in chapter 5, we have found how data augmentation improves the performance of deep models, and fails to improve the SVM model performance. In addition, we found the power of combining both jittering and synthetic data augmentation techniques with the 1D CNN-BLSTM model. Based on these findings, these techniques are applied to the best selected multimodal deep models. However, synthetic data generation is memory consuming method, and as long as we are working with long audio signals (more than 8000 samples/second) we will be facing memory problem. For this reason, only jittering data augmentation method is applied with voice modality; where in handwriting modality, jittering is applied with 2D CNNs and the combination of jittering and synthetic data is applied with the 1D CNN-BLSTMs. For jittering, several values of noise intensity are studied in order to explore its effect on classification. The training data is augmented twice in this work based on our previous finding (see chapter 5). In this case, the number of hidden nodes in each deep model defined in this chapter (and presented in Figure 7.3, Figure 7.4, Figure 7.6, and Figure 7.7) will be increased based on the empirical rules defined previously.

## 7.4 Experiments and results

In this work, multimodal assessment of PD is studied where both SVM model trained on handcrafted features and deep models are studied and compared, where the 42 subjects are divided into 3 folds, with the 66.66/33.33 % (training/validation) proportion using stratified sampling method. Sequentially, one fold is validated using the classifier trained on the remaining 2 folds. The total accuracy is obtained by calculating the mean of all the folds accuracies. Also in this work we decided not to use a separate test set due to a low database size. As a result the validation set can be considered as test set. Cross-entropy cost function

and Adam optimizer are also applied with deep models in this work, where early stopping procedure and dropout technique (where a dropout layer is added right after the hidden layer with 0.4 dropout rate) are applied to hidden layers to prevent the network from overfitting, and mini-batch training combined with shuffling at each epoch is applied for faster convergence. No batch normalization was applied.

Starting with the SVM model based on global feature-level fusion described in section 7.1, where 2 methods are applied to combine features from all tasks in each modality: average and concatenation. Once the multimodal feature vector is created, the two stage feature selection approach is applied to select the most relevant features. Statistical tests reduce the total number of features from 409 to 125 when average method is applied and, from 1,763 to 356 when concatenation method is used. The number of selected features using statistical tests, the significance level with the best validation accuracy, and the numerical results achieved by the SVM classifier with 3 folds cross-validation using only statistical tests for feature selection are presented in Table 7.1.

Table 7.1. Performance and number of selected multimodal features of each combination method in PD classification using statistical tests for feature selection.

| **Method** | **Accuracy (%)** | **Sensitivity (%)** | **Specificity (%)** | **Significance Level** | **# of selected features** |
|---|---|---|---|---|---|
| Average | 92.86 | 95.24 | 90.48 | 0.0933 | 125 |
| concatenation | 97.62 | 95.24 | 100 | 0.0350 | 356 |

The suboptimal incremental approach defined in chapter 5 is applied to select the most relevant features between the selected features in the first stage. The highest classification accuracy obtained with average method is up to 100 % for N=22 features (15 refers to handwriting features and 7 to acoustic features). In case of concatenation method, also the highest classification accuracy obtained is also up to 100 % for N=55 features (52 refers to handwriting features and 3 to acoustic features). The multimodal performances obtained with one and two stage feature selection methods are shown in Table 7.2. The selected features providing the best performance for both methods (average and concatenation) are represented in Table 7.3 and Table 7.4 respectively.

Table 7.2. Table of comparison between the performance obtained with one and two stage feature selection methods.

| Method | Performance | 1 stage Feature selection | 2 stages Feature selection |
|---|---|---|---|
| average | Accuracy (%)<br>Sensitivity (%)<br>Specificity (%) | 92.86<br>95.24<br>90.48 | 100<br>100<br>100 |
| concatenation | Accuracy (%)<br>Sensitivity (%)<br>Specificity (%) | 97.62<br>95.24<br>100 | 100<br>100<br>100 |

Table 7.3. The two stages selected features providing the best performance (average case).

| Modality | Feature | Statistic |
|---|---|---|
| Handwriting | Main part: correlation between pressure and horizontal velocity | 1st percentile |
| | Main part: correlation between pressure and vertical velocity | 99th percentile |
| | Main part: correlation between pressure and vertical acceleration | Mean |
| | Falling edge: correlation between pressure and vertical acceleration | 99th percentile |
| | Vertical jerk | Mean |
| | Acceleration | Mean |
| | Jerk | Mean |
| | Stroke width | Mean |
| | Stroke width | 99th percentile |
| | NCV/stroke | Median |
| | NCV/stroke | 1st percentile |
| | Conventional energy of Y coordinate | Scalar |
| | Second order Rényi entropy of the first empirical mode decomposition of Y coordinate | Scalar |
| | Third order Rényi entropy of the first empirical mode decomposition of Y coordinate | Scalar |
| | Third order Rényi entropy of the second empirical mode decomposition of Y coordinate | Scalar |
| Voice | 7th smoothed MFCC coefficient | Median |
| | 13th smoothed MFCC coefficient | Mean |
| | 1st order delta regression coefficient of the smoothed probability of voicing | Mean |
| | 1st order delta regression coefficient of the smoothed average local jitter | Median |
| | 1st order delta regression coefficient of the 8th smoothed MFCC coefficient | Mean |
| | 1st order delta regression coefficient of the 8th smoothed MFCC coefficient | Median |
| | 1st order delta regression coefficient of the 14th smoothed MFCC coefficient | Median |

Table 7.4. The two stages selected features providing the best performance (concatenation case).

| Modality | Task | Feature | Statistic |
|---|---|---|---|
| Handwriting | Task1 | Rising edge: correlation between pressure and vertical acceleration | STD |
| | | Falling edge: correlation between pressure and vertical velocity | Median |
| | | Falling edge: correlation between pressure and vertical velocity | 99th percentile |
| | | Falling edge: correlation between pressure and horizontal acceleration | 99th percentile |
| | | Jerk | 1st percentile |
| | | Conventional energy of Y coordinate | Scalar |
| | Task 2 | Main part: correlation between pressure and vertical acceleration | 99th percentile |
| | | Vertical acceleration | Mean |
| | | Stroke height | Median |
| | | Stroke height | 99th percentile |
| | Task 3 | Main part: NCP | STD |
| | | Main part: correlation between pressure and horizontal velocity | 99th percentile |
| | | Main part: correlation between pressure and vertical acceleration | Mean |
| | | Main part: correlation between pressure and vertical acceleration | 99th percentile |
| | | Falling edge: correlation between pressure and horizontal acceleration | Median |
| | Task 4 | Rising edge: NCP | Mean |
| | | Stroke time | Mean |
| | | Stroke time | 99th percentile |
| | | Stroke width | Mean |
| | | Stroke width | 99th percentile |

| | | | |
|---|---|---|---|
| | | NCV/stroke | Mean |
| | | NCV/stroke | 1st percentile |
| | | Second order Rényi entropy of the first empirical mode decomposition of Y coordinate | Scalar |
| | | Second order Rényi entropy of the second empirical mode decomposition of Y coordinate | Scalar |
| | | Third order Rényi entropy of the second empirical mode decomposition of Y coordinate | Scalar |
| | Task 5 | Rising edge: NCP | Median |
| | | Falling edge: correlation between pressure and horizontal velocity | 1st percentile |
| | | Stroke time | 1st percentile |
| | | Stroke time | 99th percentile |
| | | Stroke height | Mean |
| | | NCV/stroke | Mean |
| | | NCV/stroke | Median |
| | | NCV/stroke | STD |
| | | NCV/stroke | 1st percentile |
| | | NCV/stroke | 99th percentile |
| | | Shannon entropy of the second empirical mode decomposition of Y coordinate | Scalar |
| | Task 6 | Rising edge: NCP | 1st percentile |
| | | Horizontal acceleration | 99th percentile |
| | | Vertical acceleration | 1st percentile |
| | | Stroke time | 1st percentile |
| | | Shannon entropy of Y coordinate | Scalar |
| | Task 7 | Rising edge: NCP | Mean |
| | | Rising edge: NCP | Median |
| | | Rising edge: NCP | STD |
| | | Rising edge: NCP | 99th percentile |
| | | Vertical jerk | Mean |
| | | Movement Time | Scalar |
| | | Stroke time | 1st percentile |
| | | Stroke time | 99th percentile |
| | | Shannon entropy of the second empirical mode decomposition of X coordinate | Scalar |
| | | Shannon entropy of the first empirical mode decomposition of Y coordinate | Scalar |
| | | Shannon entropy of the second empirical mode decomposition of Y coordinate | Scalar |
| Voice | Text Reading | Smoothed average local jitter | Minimum |
| | | 1st smoothed MFCC coefficient | Median |
| | | 2nd smoothed MFCC coefficient | Minimum |

According to Table 7.3 and Table 7.4, most of the selected features providing the best performance in both cases (average and concatenation) include kinematic, pressure, and correlation between kinematic and pressure features for handwriting modality, and MFCC coefficients for voice modality; agreed with the conclusion found in chapters 5 and 6.

Moving to the multimodal assessment using deep learning, the three different combinations described in sections 7.2.1, 7.2.2, and 7.2.3 are studied and compared. The first experiment is where for both modalities 2D-CNNs are applied with spectrogram as input (go back to section 7.2.1), and where both global feature-level and decision-level fusion methods are applied. The second experiment refers to the case where 1D CNN-BLSTMs and 1D

CNN-MLPs are applied with raw signals for both handwriting and voice modalities respectively, and where only decision-level fusion method is applied (see section 7.2.2). The last experiment is where 1D CNN-BLSTMs and 2D CNNs are applied for both handwriting and voice modalities respectively, and where decision-level fusion method is applied (section 7.2.3). The results obtained are summarized in Table 7.5.

Table 7.5. Multimodal classification of PD and HC performance.

| Hand. Model/ Voice Model | Hand. Data input/ Voice Data input | Fusion method | Modality combination method | Best handwriting Features Combination | Multimodal Performance (%) |
|---|---|---|---|---|---|
| 2D CNN /2D CNN | Whole signal Spectrogram image/ Whole signal Spectrogram image | Feature-level | Averaging | Z | Acc:78.57 Sens:85.71 Spec:71.43 |
| | Whole signal Spectrogram image/ Whole signal Spectrogram image | Feature-level | Concatenating | X+Y+Z+ Pressure+ Altitude+Azimuth | Acc: 61.9 Sens: 61.9 Spec: 61.9 |
| | Whole signal Spectrogram image/ Whole signal Spectrogram image | Decision-level | | X+Y+Z+ Pressure+ Altitude | **Acc: 83.33** Sens: 87.71 Spec: 80.95 |
| | Whole signal Spectrogram image/ 4s segments Spectrogram images | Decision-level | | X+Y+Z+ Pressure+ Altitude | **Acc: 83.33** Sens: 87.71 Spec: 80.95 |
| 1D CNN BLSTM /1D CNN-MLP | Whole raw signal/ 4s segments raw signals | Decision-level | | X+Y+Z+ Pressure+ Altitude+Azimuth | **Acc: 88.1** Sens: 80.95 Spec: 95.24 |
| 1D CNN-BLSTM /2D CNN | Whole raw signal/ Whole signal Spectrogram image | Decision-level | | X+Y+Z+ Pressure+ Altitude+Azimuth | **Acc: 88.1** Sens: 80.95 Spec: 95.24 |
| | Whole raw signal/ 4s segments Spectrogram images | Decision-level | | X+Y+Z+ Pressure+ Altitude+Azimuth | **Acc: 88.1** Sens: 80.95 Spec: 95.24 |

Table 7.6.Handwriting, audio and multimodal classification performance obtained with decision-level fusion method.

| Hand. Model/ Voice Model | | Hand. Data input/ Voice Data input | Best Handwriting Features Combination | Handwriting Performance (%) | Voice Performance (%) | Multimodal Performance (%) |
|---|---|---|---|---|---|---|
| M1 | 2D CNN /2D CNN | Whole signal Spectrogram image/ Whole signal Spectrogram image | X+Y+Z+ Pressure+ Altitude | Acc: 83.33 Sens: 87.71 Spec: 80.95 | Acc: 54.76 Sens: 76.19 Spec: 33.33 | Acc: 83.33 Sens: 87.71 Spec: 80.95 |
| M2 | 2D CNN /2D CNN | Whole signal Spectrogram image/ 4s segments Spectrogram images | X+Y+Z+ Pressure+ Altitude | Acc: 83.33 Sens: 87.71 Spec: 80.95 | Acc: 52.38 Sens: 61.9 Spec: 42.86 | Acc: 83.33 Sens: 87.71 Spec: 80.95 |
| M3 | 1D CNN-BLSTM /1D CNN-MLP | Whole raw signal/ 4s segments raw signals | X+Y+Z+ Pressure+ Altitude+Azimuth | Acc: 88.1 Sens: 80.95 Spec: 95.24 | Acc: 54.76 Sens: 23.81 Spec: 85.71 | **Acc: 88.1** Sens: 80.95 Spec: 95.24 |
| M4 | 1D CNN-BLSTM /2D CNN | Whole raw signal/ Whole signal Spectrogram image | X+Y+Z+ Pressure+ Altitude+Azimuth | Acc: 88.1 Sens: 80.95 Spec: 95.24 | Acc: 54.76 Sens: 76.19 Spec: 33.33 | **Acc: 88.1** Sens: 80.95 Spec: 95.24 |
| M5 | 1D CNN-BLSTM /2D CNN | Whole raw signal/ 4s segments Spectrogram images | X+Y+Z+ Pressure+ Altitude+Azimuth | Acc: 88.1 Sens: 80.95 Spec: 95.24 | Acc: 52.38 Sens: 61.9 Spec: 42.86 | **Acc: 88.1** Sens: 80.95 Spec: 95.24 |

Table 7.7. Task-wise system and "All-tasks" system accuracies (in %) for the various models presented in Table 7.6.

| Task | M1 | M2 | M3 | M4 | M5 |
|---|---|---|---|---|---|
| Repetitive cursive letter 'l' | **54.76** | **54.76** | **54.76** | **54.76** | **54.76** |
| Triangular wave | 50.00 | 50.00 | **83.33** | **83.33** | **83.33** |
| Rectangular wave | **64.29** | **64.29** | 59.52 | 59.52 | 59.52 |
| Repetitive "Monday" | 61.90 | 61.90 | **73.81** | **73.81** | **73.81** |
| Repetitive "Tuesday" | **64.29** | **64.29** | 54.76 | 54.76 | 54.76 |
| Repetitive "Name" | **64.29** | **64.29** | 38.10 | 38.10 | 38.10 |
| Repetitive "Family Name" | **71.43** | **71.43** | **71.43** | **71.43** | **71.43** |
| Sustained vowel 'a' | **54.76** | **54.76** | 52.38 | **54.76** | **54.76** |
| Text reading | 64.29 | **73.81** | 61.9 | 64.29 | **73.81** |
| All tasks | 83.33 | 83.33 | **88.1** | **88.1** | **88.1** |

For 2D CNN model, decision-level fusion method performs better than feature-level fusion. This can be related to the fact that feature-level fusion is effective when time synchronized modalities are to be fused (fusion of speech and eye movements for example) [Dumas et al., 2009]. The best models are selected and presented in Table 7.6, where the "All-tasks" performances of both modalities beside the multimodal system are shown. For each model presented in Table 7.6, Task-wise system accuracies for each modality are presented in Table 7.7.

Moving to the results obtained with data augmentation techniques described in section 7.3, where for jittering a random additive scalar is sampled from a Gaussian distribution with zero mean and STD of 0.3 for handwriting modality and zero mean and STD of 0.1 for voice modality. The performances of both modalities ("All-tasks") beside the multimodal system are presented in Table 7.8; where Task-wise accuracies are presented in Table 7.9 respectively.

Based on these results, we can see that the best handwriting features combination found is the same as the one found with handwriting modality (see chapter 5), and it is clear how deep learned audio features has no effect on PD detection, and how the results obtained when applying deep learning to detect PD from raw speech signals are not satisfactory and not nearly compelling as the ones obtained with handwriting analysis. It is challenging to learn acoustic deep models from raw signals and especially when very few convolutional layers are used for acoustic feature extraction, which might be insufficient for building high-level discriminative features [Dai et al., 2017]. In our case, since we are working with small dataset, it will not be a good idea to enlarge our model. It will be more efficient to build the

model using some low-level audio descriptors instead of applying the raw audio waveforms directly. In addition, the log-spectrogram offers a rich representation of the temporal and spectral structure of the input signal. Considering the log-spectrograms as input features may improve our acoustic deep model performance. According to Table 7.8, the results show how the accuracy performance is improved from 52.38 % to 71.43 % after considering the log-spectrogram as input instead of the raw signal. However, the achieved results are still non-satisfactory compared to the results obtained with handwriting. A possible explanation of this behavior is that in spectrogram representations, it is difficult to separate simultaneous sounds since they all sum together into a distinct whole. This means that a particular observed frequency in a spectrogram cannot be assumed to belong to a single sound. In addition, moving a sound vertically in a spectrogram might influence the meaning. Therefore, the spatial invariance that 2D CNNs provide might not perform as well for this form of data. Finally, periodic sounds comprised of a fundamental frequency and a number of harmonics which are most often non-locally distributed on the spectrogram. Finding local features in spectrograms using 2D convolutions will be complicated in this case [Lonce, 2017].

Table 7.8. Performance measures obtained after applying data augmentation and decision-level fusion, where the best handwriting features combination are the ones found in Table 7.6.

| Hand. Model/ Voice Model | | Hand. Data input/ Voice Data input | Augmentation Technique | Handwriting Performance (%) | Voice Performance (%) | Multimodal Performance (%) |
|---|---|---|---|---|---|---|
| M1 | 2D CNN /2D CNN | Whole signal Spectrogram image/ Whole signal Spectrogram image | Hand: Jitter<br>Voice: Jitter | Acc: 83.33<br>Sens: 87.71<br>Spec: 80.95 | Acc: 57.14<br>Sens: 95.24<br>Spec: 19.05 | Acc: 83.33<br>Sens: 87.71<br>Spec: 80.95 |
| M2 | 2D CNN /2D CNN | Whole signal Spectrogram image/ 4s segments Spectrogram images | Hand: Jitter<br>Voice: Jitter | Acc: 83.33<br>Sens: 87.71<br>Spec: 80.95 | Acc: 71.43<br>Sens: 76.13<br>Spec: 66.67 | Acc: 85.71<br>Sens: 71.43<br>Spec: 100 |
| M3 | 1D CNN-BLSTM /1D CNN-MLP | Whole raw signal/ 4s segments raw signals | Hand: Jitter+Syn<br>Voice: Jitter | **Acc: 97.62**<br>Sens: 95.24<br>Spec: 100 | Acc: 52.38<br>Sens: 33.33<br>Spec: 71.43 | **Acc: 97.62**<br>Sens: 95.24<br>Spec: 100 |
| M4 | 1D CNN-BLSTM /2D CNN | Whole raw signal/ Whole signal Spectrogram image | Hand: Jitter+Syn<br>Voice: Jitter | **Acc: 97.62**<br>Sens: 95.24<br>Spec: 100 | Acc: 57.14<br>Sens: 95.24<br>Spec: 19.05 | **Acc: 97.62**<br>Sens: 95.24<br>Spec: 100 |
| M5 | 1D CNN-BLSTM /2D CNN | Whole raw signal/ 4s segments Spectrogram images | Hand: Jitter+Syn<br>Voice: Jitter | **Acc: 97.62**<br>Sens: 95.24<br>Spec: 100 | Acc: 71.43<br>Sens: 76.13<br>Spec: 66.67 | Acc: 95.24<br>Sens: 95.24<br>Spec:95.24 |

Table 7.9. Task-wise system and "All-tasks" system accuracies (in %) for various models and training schemes presented in Table 7.8.

| Task | M1 | M2 | M3 | M4 | M5 |
|---|---|---|---|---|---|
| **Augmentation technique** | **Jitter** | **Jitter** | **Jitter/Syn. Data** | **Jitter/Syn. Data** | **Jitter/Syn. Data** |
| Repetitive cursive letter 'l' | **69.05** | **69.05** | 59.52/47.62 | 59.52/47.62 | 59.52/47.62 |
| Triangular wave | 71.43 | 71.43 | **80.95**/78.57 | **80.95**/78.57 | **80.95**/78.57 |
| Rectangular wave | 61.9 | 61.9 | 71.43/**76.19** | 71.43/**76.19** | 71.43/**76.19** |
| Repetitive "Monday" | 59.52 | 59.52 | **78.57**/76.19 | **78.57**/76.19 | **78.57**/76.19 |
| Repetitive "Tuesday" | **71.43** | **71.43** | 57.14/59.52 | 57.14/59.52 | 57.14/59.52 |
| Repetitive "Name" | 52.38 | 52.38 | **57.14**/50 | **57.14**/50 | **57.14**/50 |
| Repetitive "Family Name" | **71.43** | **71.43** | 69.05/64.29 | 69.05/64.29 | 69.05/64.29 |
| **Augmentation technique** | **Jitter** | **Jitter** | **Jitter** | **Jitter** | **Jitter** |
| Sustained vowel 'a' | 52.38 | **66.67** | 52.38 | 52.38 | **66.67** |
| Text reading | 54.76 | **73.81** | 54.76 | 54.76 | **73.81** |
| All tasks | 83.33 | 85.71 | **97.62** | **97.62** | 95.24 |

From the other side, cropping the audio signal into short segments of the same length (or short-term analysis) and getting the log-spectrograms of each segment seems to be more effective than getting the log-spectrograms of the whole signal (global analysis) with 2D CNN (accuracy performance is improved from 57.14 % to 71.43 %). We believe that this is due to the larger number of training samples, and to the idea of maintaining the nonlinear variation over the time axis.

From a quick analysis of the Task-wise accuracies presented in Table 7.7 and Table 7.9, we notice that the text reading task performs better than the sustained phonation vowel 'a' in PD detection. We believe that the text reading task is richer in terms of acoustic and prosodic information, which makes them more convenient for automatic PD detection in contrast to maximum phonation time of vowel 'a' which contains less information [Pompili et al., 2017]. If we go back to our SVM model trained on handcrafted acoustic features, we can also notice the same conclusion (see Table 6.3 in chapter 6).

Finally, we can see how data augmentation improves the 2D CNN model performance when audio signal segmentation is applied, and fails to improve the performance of the 1D CNN-MLP with raw signal and the 2D CNN without segmentation. In our previous work (chapter 5), we have found that data augmentation does not improve the 2D CNN model with online handwriting spectrograms since the augmented time series are converted into spectrograms then normalized to a fix dimension. While here, since no normalization is

applied on spectrograms, this means that the model may benefit the most from the new generated signals.

In our opinion, it may be more efficient to build a deep model using some low level acoustic descriptors instead of using speech signal, which might build deep speech features without the need to enlarge our model. From the other side, the acoustic handcrafted features extracted in our previous work (chapter 6) have achieved very good results in pure speech and in multimodal corpuses. In general, deep learning models are basically selected to avoid handcrafted features extraction that needs an expert knowledge of the field, or to achieve a better result by extracting deep features. Since our multimodal system built with handcrafted features reaches 100 % accuracy, and we believe that deep models in voice modality require some handcrafted low level audio features in order to be more effective, we agree that it does not make sense to apply deep learning in this case.

## 7.5 Conclusions

The aim of this part is to build a language-independent multimodal system for PD early detection by combining handwriting and voice. SVM and deep learning models are both studied and compared. The aim here is to extract important information from both modalities forming a multimodal vector that will be used for classification; where both global Feature-level and Decision-level fusion methods are applied. In case of SVM model, for each bio-signal, the handcrafted features described in previous chapters are extracted per task for each subject, and then even the average or the concatenation of the different tasks in each modality is calculated. For each bio-signal one feature vector is obtained; the embeddings obtained from the 2 bio-signals will be concatenated to form a multimodal vector per subject. In case of deep learning, five different combinations are also studied and compared. Starting by the case where for both modalities, 2D CNNs are applied with spectrogram of the whole signal as input, and where both global feature-level and decision-level fusion methods are applied. The second multimodal system combines both 2D CNNs with spectrogram of the whole handwriting signal as input, and 2D CNNs with spectrogram of each 4s voice segment; where only decision-level fusion method is applied. The third system refers to the case where 1D CNN-BLSTMs are applied with the whole raw handwriting signal, and 1D CNN-MLPs are applied with raw voice signals in a 4s segments; where only decision-level fusion method is

applied. The last 2 models combine 1D CNN-BLSTMs with the whole handwriting signal as input and 2D CNNs with spectrogram of the whole voice signal or with spectrogram of each 4s segment; where decision-level fusion method is applied. Once the best models are selected, data augmentation techniques are applied on both modalities. The results obtained with the SVM model are higher than the ones obtained with deep learning. We have found how SVM with the combination of information from both handwriting and speech modalities deliver a more accurate PD prediction (accuracy up to 100 % that need to be confirmed on larger scaled data) than pure handwriting and speech analyses.

A number of observations and conclusions are obtained from this work and summarized in this section. First of all, we found that decision-level fusion method is more efficient than feature-level fusion in case of combining handwriting and voice, since we are dealing with non-synchronized bio-signals. In addition, we have noticed how it is challenging to learn acoustic deep models from raw signals and especially when very few convolutional layers are used for acoustic feature extraction, which might be insufficient for building high-level discriminative features. Feeding the CNN with 2D spectrograms of the raw audio signals improves the results, since the log-spectrogram offers a rich representation of the temporal and spectral structure of the original signal. However, the results are still non-satisfactory; we believe that this can be related to the difficulty in separating simultaneous sounds in a spectrogram, the influence of moving a sound vertically on the meaning, and the difficulty in finding local features in spectrograms due to the non-locally distribution of the fundamental and harmonic frequencies.

We also confirm higher number of training data samples and preservation of the signal nonlinear variation over the time axis improve the performance. It was also found that text reading task performs better than the sustained phonation vowel 'a' due to the high probability of existing acoustic and prosodic information that may be considered sufficient for PD detection. In addition, data augmentation methods applied on voice signals can increase deep learning model performance when the raw signals are converted into 2D spectrograms and the non-linearity over time axis is preserved. Finally, in case of working with small database, it may be more efficient to build a deep model using some low level acoustic descriptors instead of using speech signal, which might build deep speech features without the need to enlarge the model.

## 7.6 Correlation between hand-crafted and deep learned features

To enhance interpretability of the extracted handwriting/acoustic deep features, we have conducted correlation analysis between the features maps obtained by the convolutional layers of the 2D CNN model (shown in Figure 5.49 with k spectrogram input images; where k here is equal to 7 for handwriting and 1 for speech), and the hand-crafted features defined in sections 5.1 and 6.5. Deep learned and handcrafted features were obtained within each modality for each task separately forming two matrices of size (42×7, 114688) and (42×7, 189) for handwriting and two other matrices of size (42×2, 16384) and (42×2, 220) for speech. In this section, for simplicity, we decided to sample all the speech tasks at 8 KHz. Two techniques were applied to see how strong the relationship between deep and handcrafted features is.

In the first technique, the correlation between each handcrafted and deep learned feature is obtained (resulting a correlation matrix of size (189, 114688) for handwriting and another one of size (220, 16384) for speech). The correlation maps for both modalities are presented in Figure 7.8-a and Figure 7.9-a respectively, where we only show the highest 189 correlated handwriting deep features and the highest 220 correlated acoustic deep features. In these heat maps, blue and red colors show positive and negative correlation, respectively. The darker the color, the stronger is the relation. We can consider that hand-crafted and deep features are highly correlated for handwriting, where for speech the correlation is not as strong as it is with handwriting.

The second technique applied is the multiple linear regression (MLR) defined in [Olive, 2017], where we made the assumption that: Y=AX (where the dependent variable denoted by Y refers to hand-crafted feature and the independent variables denoted by X refer to deep learned features since deep learned features may include more variety of features than the handcrafted features). Each hand-crafted feature is regressed separately on the deep learned features. Both hand-crafted and deep features were standardized in order to obtain features of zero mean and STD of one. Then, PCA was applied on deep learned features to reduce the size by finding the best linear combinations of the original variables so that the

variance along the new variable is maximum [Pal, 2020]. The MLR errors obtained on each hand-crafted feature (handwriting or speech) for different PCA-size reduction are shown in Figure 7.8-b and Figure 7.9-b respectively, where we made sure that the number of samples is at least half more than the number of parameters (PCA-size reduction). The best results were obtained when we reduced the size of handwriting deep features to 189, and acoustic deep features to 50. Some of the handcrafted features are not varying along time, leading to zero STD. This can explain why they have reported zero MLR errors.

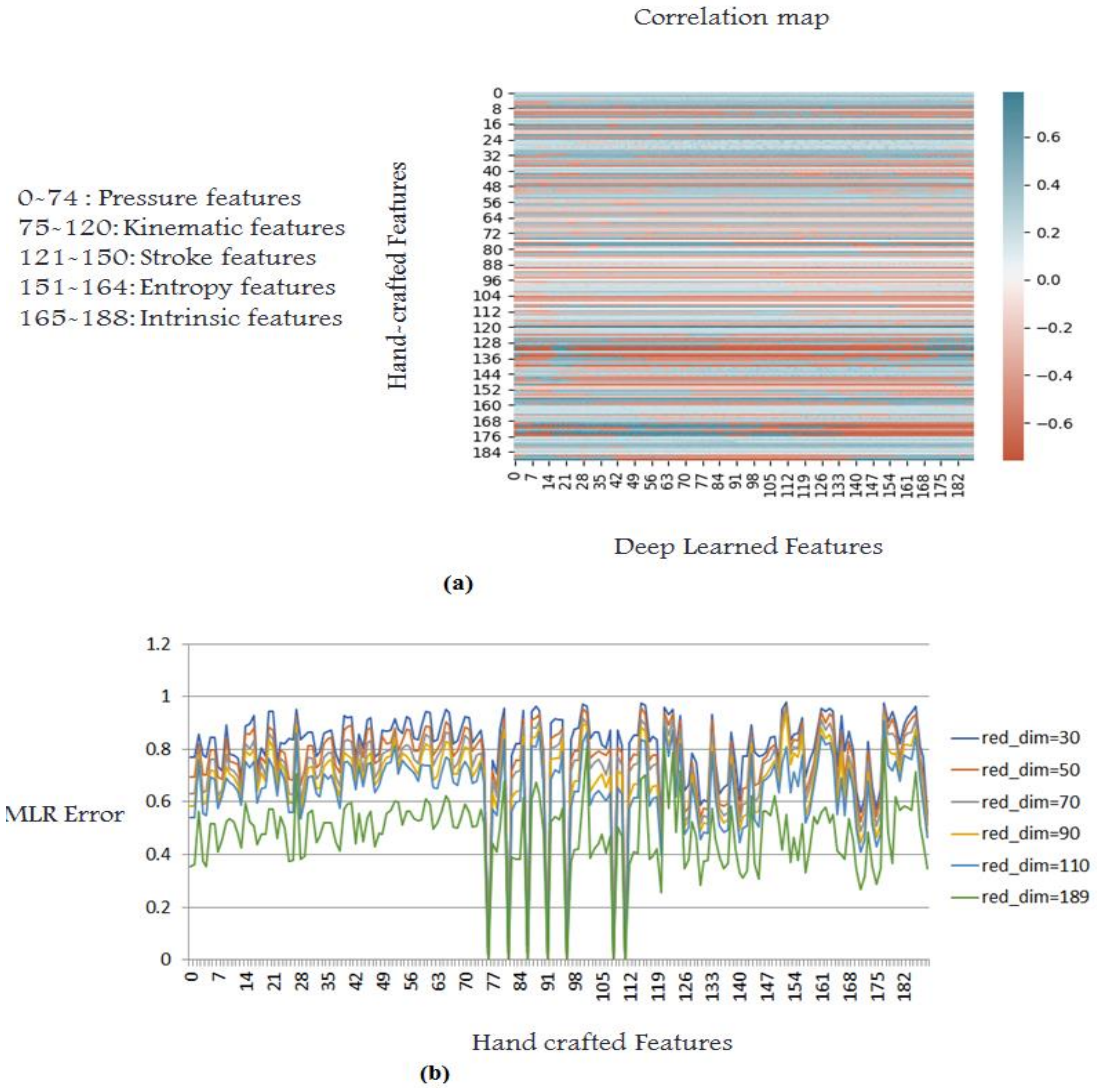



Figure 7.8. Relationship between handwriting hand-crafted and Deep Learned features.

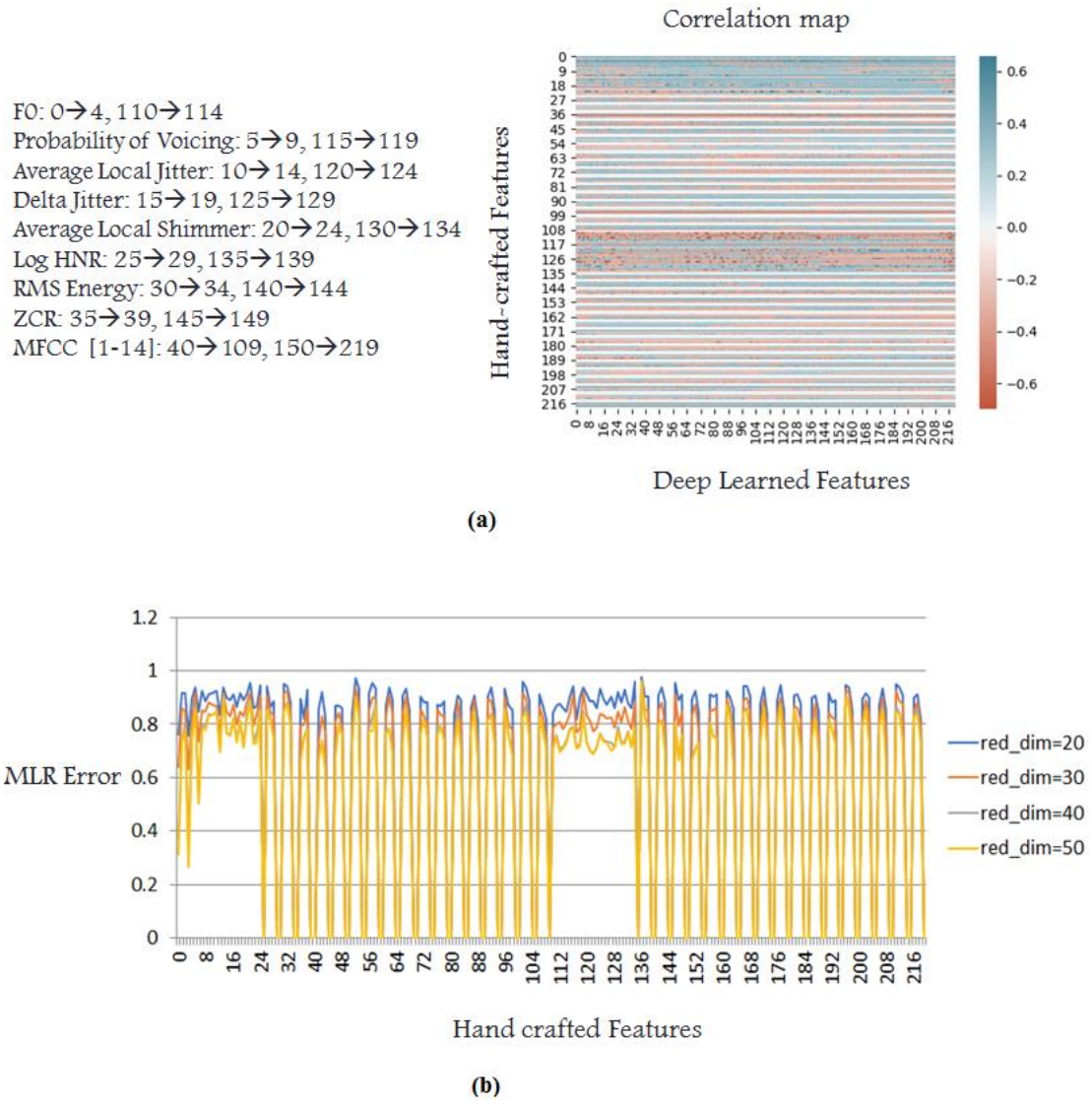



Figure 7.9. Relationship between acoustic hand-crafted and Deep Learned features.

For handwriting, we can see as overall that the results obtained are acceptable, where some hand-crafted features are highly correlated with deep features that others (the ones with errors below 0.4). Whereas, for speech we can see that very few hand-crafted features are highly correlated with deep features, where most of them are not strongly correlated. Based on the findings in both techniques, we can say that the handwriting hand-crafted and deep learned features are highly correlated, and among the highly correlated hand-crafted features, some refer to Pressure and Kinematic features, which confirm the importance of such features in PD detection. For speech, we found that the acoustic hand-crafted and deep learned features are not strongly correlated as the handwriting features (confirmed with the conclusion found in section 7.5 concerning audio spectrogram representation with CNN).

# 8 Conclusions and Future Work

The aim of this thesis is to build a language-independent multimodal system for assessment the motor disorders in PD patients at an early stage based on combined handwriting and speech signals where both: handcrafted features and classical classifiers, and deep learning approaches are studied. Even though it is difficult to diagnose PD at an early stage, small differences in handwriting and voice is machine-detectable. In this thesis, the handcrafted features and classical classifiers were implemented in Matlab [Etter, 2018], where deep models were created in Tensorflow [Abadi et al., 2016] (Python library for deep learning [Guttag, 2012]) so we can run the model on a GPU that allowed us to run fast prototypes.

Toward this aim and due to the lack of such PD databases, a multimodal and multilingual database that includes handwriting, voice, and eye movements recordings collected from PD patients and HC subjects was constructed. Once the multimodal database was collected, an automatic system that assists in the decision-making process for the diagnosis of PD and the prediction of the modified H&Y stage using global handcrafted features extracted from handwriting was first built. Advanced language-independent handwriting markers based on kinematic, stroke, pressure, entropy, and intrinsic features were extracted from the “on-paper” periods in each handwriting task. A two-stage feature selection approach was applied to avoid the risk of falling in a curse of dimensionality (chapter 5- section 5.1). A binary SVM classifier trained on a set of kinematic, pressure, and correlation between kinematic and pressure features succeeded in detecting PD with an accuracy of 96.87 %. Pressure features offer further detailed information that can not be obtained from the kinematic features, here comes the importance of combining kinematic and pressure features.

Based on these selected features, a multi-class SVM classifier was built for stage detection, where windowed segmentation was applied on each task to deal with small database and multiclass distribution (chapter 5- section 5.2). To acquire the final scores for classification decision, an MLP model was applied to combine the scores of all segments from one subject. Many obstacles were faced in this part such as: large variability over

patients in term of symptoms and stage of disease, imbalanced data and class distribution, and limited number of samples. Re-balancing training data, 1-off accuracy, stage group classification, ensuring the same class distribution in all the subsets are all studied to see how the factors mentioned before affect the classification performance. The results obtained confirm that similar class distribution in all the subsets plays a major role in the classification performance, and that either the 1-off accuracy or stage group classification can overcome the large variability over patients' problem. In addition, we have demonstrated that learning a classifier with real balanced training data (and not artificially balanced) return better results than a trained classifier with real unbalanced data. Finally, combining training and validation datasets after hyper-parameters selection improves the classification performance and emphasizes the importance of the amount of data available for PD stage prediction.

The next work done was to build a system for PD early detection based on short-term features and deep learning to avoid handcrafted features extraction, to benefit from its ability to extract deep features, and to avoid losing some important information while applying global features extraction. Two based learning models for end to-end time series classification were proposed (the 2D CNN and the 1D CNN-BLSTM), where two different frameworks were proposed to encode time series into images for the 2D CNN model (spectrogram and modified GAF). The entire handwriting dynamic signals were studied so both in-air and on-surface features will be extracted. We have demonstrated the importance of both: a deep architecture based on the combination of 1D CNN and BLSTM recurrent layers, and a 2D CNN model with spectrograms as input in PD detection. These models have the ability to tackle the variation of information in time series either by explicitly considering the local short term information on the time axis of the non-stationary online handwriting signals or by dealing with raw time series directly. To cope with the limited data, and to improve our deep models, transfer learning and data augmentation approaches were applied. In previous work [Mormont et al., 2018] it was proved that transfer learning is efficient with CNN models. For this reason, multiple transfer learning strategies were investigated and compared across only the 2D CNN model with spectrogram images as input. For data augmentation, different techniques were applied on raw time series to generate new synthetic samples. These new synthetic samples are even encoded into 2D spectrogram images or applied directly to train and evaluate the 2D CNN or the 1D CNN-BLSTM model. The challenging PD task is

successfully tackled using the 1D CNN-BLSTM model described above and the combination of jittering and synthetic data augmentation methods yielding an accuracy of 97.62 %. A number of observations and conclusions were obtained in this part (chapter 5- section 5.4.5). It was found that the more convolutional layers included in the fine-tuning in transfer learning, the better performance we get. In addition, we believe that to benefit from transfer learning a large dataset should be used to pre-train the model. Secondly, we have shown that Z coordinates are important in PD classification, and it was shown that data augmentation (jittering and synthetic data techniques) is more effective than transfer learning at improving deep models performance, especially when time series are applied directly to the model, where it deteriorates the SVM classifier performances. Finally, in case of small database, we have proven that deep models can perform better than the classical machine learning models when proper data augmentation methods are applied. To validate these conclusions, two more experiments were conducted. The first one is when the best model obtained was trained and tested on PaHaW database, and the second one is when the best model obtained was trained and tested on HandPDMultiMC dataset where the Z coordinates attribute was excluded. The results obtained with both experiments are close to each other but worse than the one obtained with the existence of Z coordinates, which confirm the importance of this feature and the relevance of the results obtained. PD is characterized by tremor or irregular muscle contractions that introduce randomness to the movement during handwriting in the X-Y-Z space. The intensity of tremor usually decreases when hand is laid down on a surface. This means that tremor oscillations have larger amplitude when the pen is not touching the surface (in-air phase). It will more efficient to analyze tremor along the three axes since there is no consensus on which axis tremor oscillations have larger amplitudes. This can explain the importance of the Z feature in PD detection.

After succeeding in building a language-independent model for PD diagnosis using handwriting analysis, the next work was to build a language and task-independent acoustic feature set for assessing the motor disorders in PD patients, and to study the influence of sampling rate and unvoiced sounds on the performance. Only phonation and articulation handcrafted features that can be extracted for all the tasks under assessment are studied, where the prosody features are excluded since they are related to intonation, stress, and rhythm, which are depending on the language spoken. LLD features were extracted per frame

from the processed voice signal (chapter 6). Global features were obtained from the z-scored LLD features by applying some statistical functions. Unvoiced sounds and sampling rate effects on classification performance of PD detection through voice analysis were studied. We have succeeded to build a language-independent SVM model for PD diagnosis through voice analysis with 97.62 % accuracy. It was found that the effect of sampling rate on PD classification may depend on task and features used. We have found that signals with low sampling rate (less than 16 KHz) can lose valuable information that can play a good role in PD detection, where a sampling rate of 24 KHz for sustained vowel ‘a’ and text reading (voiced sounds) tasks, and 16 KHz for text reading task (voiced and unvoiced sounds) are appropriate for the features analyzed here. Actually, High-frequencies signal components will be lost when the original signal is down-sampled to a lower sampling rate. The high-frequencies components may include recording environment noise, or noise due to tremor, or the breathy voice related to the incomplete vocal fold closure (see chapter 6). The best sampling rate values found here confirm with the conclusion that the highest linguistically meaningful frequencies are below 11 KHz built in [Stevens, 1998], [Ladefoged, 2003]. We have also demonstrated that unvoiced frames also play a role in the detection related to the abductor spasms occurring mostly on unvoiced consonants. Finally, the importance of MFCC coefficients to quantify the problems in speech articulation and to detect the disease was shown.

In the last part of this thesis we focused on building a language-independent multimodal system for PD early detection by combining handwriting and voice signals, where classical SVM model and deep learning models (such as 2D CNN with spectrogram image as input, 1D CNN-BLSTM, and 1D CNN-MLP) were both analyzed. The aim here was to extract complementary information from both modalities forming a multimodal vector that will be used for detection; where both global feature-level and decision-level fusion methods were applied (chapter 7). Data augmentation was applied on the best deep model selected, and compared to the classical SVM model. Classification accuracy up to 100 %, that need to be confirmed on larger scaled data, was obtained when handcrafted features from both modalities are combined and applied to the SVM. We found that decision-level fusion method performs better than feature-level fusion in case of combining non-synchronized bio-signals. In addition, we have noticed how it is challenging to learn acoustic deep models from

raw signals and especially when very few convolutional layers are used for acoustic feature extraction, which might be insufficient for building high-level discriminative features. Feeding the 2D CNN with 2D spectrograms of the raw audio signals improves the results, since the log-spectrogram offers a rich representation of the temporal and spectral structure of the original signal. However, despite the observed improvement, the results were still non-satisfactory. We believe that this can be related to the difficulty in separating simultaneous sounds in a spectrogram (since they all sum together into a distinct whole), the influence of moving a sound vertically in a spectrogram on the meaning (the spatial invariance that 2D CNNs provide might not perform as well for this form of data), and the difficulty in finding local features in spectrograms (using 2D convolutions) due to the non-locally distribution of the fundamental and harmonic frequencies. We have also noticed that higher number of training data samples and preservation of the signal nonlinear variation over the time axis improve the performance. Text reading task performs better than the sustained phonation vowel 'a' most probably due to the existence of acoustic and prosodic information that are considered sufficient for PD detection. In addition, data augmentation methods applied on voice signals may increase deep learning model performance when the raw signals are converted into 2D spectrograms and the non-linearity over time axis is preserved. Finally, in case of working with small database, it may be more efficient to build a deep model using some low level acoustic descriptors instead of using raw speech signal, which might build deep speech features without the need to enlarge the model. At the end of this work, we have studied the correlation between hand-crafted and deep learned features to enhance interpretability of the extracted handwriting/acoustic deep features.

The classical and the deep models built in this thesis for PD early detection are considered language-independent since the studied signal (handwriting or speech) can be considered as a summation of three basic components including linguistic information $l_n$, channel information $c_n$, and disease information $d_n$. In case of classical models, the global features were obtained by applying some statistical functions (such as mean, median, STD, etc.) to the short-term features. The average of channel information ($C_c$) is the same for all the subjects; where the average of the linguistic information ($C_l$) can be considered very similar between subjects since averaging may remove all the language specificity existing in a speech. As a conclusion, the average of the observed signal can be considered as the

summation of a certain noise ($C_c + C_l$) with disease characteristics; where the noise will not interfere in the classification. The same conclusion can also be applied for deep models, since it is well known that such models can get the average in a better way than the linear one.

This thesis has touched upon many topics and has raised a number of important ideas that can be used as starting points of further studies:

- Despite the encouraging results obtained, there is a long way to go before putting our PD early detection multimodal model into clinical use due to the fact that we have few subjects. The results obtained can not be generalized directly to thousands of subjects, but we assumed that the best architecture found on our samples, will be also a good architecture when dealing with a broader and more diverse database. For this reason, the observations and conclusions obtained, and the relevance of our system need to be validated on a larger scaled database.

- Many factors that influence handwriting, voice, or eye movements exist and can affect classification decision. In this thesis, patients in their "on-state" were studied. Medication can affect the movements of patients which can then impact the classification process. Further studies are needed to approve the conclusions drawn in this work in both the "on-state" and the "off-state" cases.

- According to the reviewed literature, most of the studies have focused on motor skills while ignoring the non-motor ones. Parkinson's patients report an awareness of change in their cognitive abilities before detecting deterioration in motor skills [Siegel, 2013]. Cognitive impairment in PD is not only a late-stage feature of the disease, but may be evident in about 20-30 % of early newly diagnosed PD patients; that is why the identification of prodromal non-motor symptoms may contribute to the precious diagnosis of PD. Future studies should explore the associations between cognitive and motor aspects of PD and handwriting, voice and eye movements.

- Predicting PD stage and progression from handwriting or voice is a very challenging task depending on many factors such as the large variability over

patients, the irrelevant motion interference, the difficulty in obtaining a large and balanced database, and the influence of the medication on the stage of the disease. Levodopa medication in most cases reduces the symptoms of PD, where it also may contribute to the development of dyskinesia or uncontrolled movements (mentioned in Chapter 2). In both cases, it is difficult to separate between the stages because patient can even get nearer to the non-PD or early stages, or nearer higher stages. As a future work, it will be important to perform our modal described in section 5.2 on a large and balanced dataset in both the “on-state” and the “off-state” cases.

- Building deep models with some low level acoustic descriptors as inputs instead of raw speech signals is needed to approve its effectiveness in early PD detection.

- It will be also important as a future work to build a PD early diagnosis automatic system based on eye movements, and another one based on the combination of handwriting, voice, and eye movements signals.

During the following years I hope that I will be able to contribute new ideas and concepts both in biomedical applications, and also in more wide impact, wide applicability statistical machine learning algorithms.

# 9 List of Publications

## Journal Papers

Taleb, C., Likforman-Sulem, L., Mokbel, C., & Khachab, M. (2020). Detection of Parkinson's disease from Handwriting using Deep Learning: a Comparative Study. *Evolutionary Intelligence*. doi:10.1007/s12065-020-00470-0

## Conference papers

Taleb, C., Likforman, L., Khachab, M., and Mokbel, C. (2017). *Feature Selection for an Improved Parkinson's Disease Identification Based on Handwriting*. Arabic Script Analysis and Recognition (ASAR), 2017 1st International Workshop, Nancy, France.

Taleb, C., Likforman, L., Khachab, M., & Mokbel, C. (2018). *A Reliable Method to Predict Parkinson's Disease Stage and Progression based on Handwriting and Re-sampling Approaches*. Arabic Script Analysis and Recognition (ASAR), 2018 2nd International Workshop, the Alan Turing Institute, London-UK.

Taleb, C., Likforman, L., Khachab, M., and Mokbel, C. (2019). *Visual Representation of Online Handwriting Time Series for Deep Learning Parkinson's Disease Detection*. Arabic Script Analysis and Recognition (ASAR), 2019 3rd International Workshop, Sydney, Australia.

Taleb C., Likforman-Sulem L., Mokbel C. (2020). Improving Deep Learning Parkinson's Disease Detection through Data Augmentation Training. In: Djeddi C., Jamil A., Siddiqi I. (eds), *Pattern Recognition and Artificial Intelligence* (pp. 79–93). Cham: Springer International Publishing. (Got the best student paper award at the 3rd Mediterranean Conference on Pattern Recognition and Artificial Intelligence, 2019).

# 10 Détection de la Maladie de Parkinson par Analyse Multimodale Combinant Signaux d’Écriture et de Parole

## 10.1 Introduction, hypothèse et objectifs

La maladie de Parkinson (MP) est un trouble neurologique causé par une diminution du niveau de dopamine dans le cerveau. Cette maladie est caractérisée par des symptômes moteurs et non moteurs qui s'aggravent avec le temps. Aux stades avancés de la maladie de Parkinson, le diagnostic clinique est clair. Cependant, dans les premiers stades, lorsque les symptômes sont souvent incomplets ou subtils, le diagnostic devient difficile et, parfois, le sujet peut rester non diagnostiqué. De plus, le suivi de la progression de la maladie dans le temps nécessite des visites cliniques répétées du patient [Nilashi et al., 2016]. La difficulté à détecter et à suivre la progression de la MP est une forte motivation pour les outils d'évaluation informatisés, les outils d'aide à la décision et les instruments de test qui peuvent aider au diagnostic précoce et à la prévision de la progression de la maladie. Une détection rapide, de préférence à un stade plus précoce que ce qui est actuellement possible, et une intervention ultérieure pourraient être extrêmement bénéfiques, de sorte que le patient pourrait avoir accès à une thérapie modificatrice de la maladie pour ralentir la progression de la maladie.

La détérioration de l'écriture et la déficience vocale et oculaire peuvent être l'un des premiers indicateurs de l'apparition de la maladie. Selon la littérature, un modèle indépendant du langage pour détecter la MP à l'aide de signaux multimodaux n'a pas été suffisamment étudié. L'objectif principal de cette thèse est de construire un système multimodal indépendant du langage pour évaluer les troubles moteurs chez les patients atteints de la MP à un stade précoce, basé sur des signaux combinés d'écriture et de parole, en utilisant des techniques d'apprentissage automatique. Dans ce but, et en raison de l'absence d'un ensemble de données multimodales et multilingues, une telle base de données, également répartie entre les témoins et les patients atteints de la MP, a d'abord été construite, où les patients seront examinés avant et après la prise de médicaments à base de L-dopa. La base de données doit

inclure l'enregistrement de l'écriture, de la parole et des mouvements des yeux. Toutefois, comme notre objectif est la détection précoce de la maladie et qu'il est très difficile de collecter une vaste base de données aux premiers stades (avec de nombreux échantillons à chaque stade), nous partons du principe que les patients atteints de la maladie de Parkinson sous traitement ont des imperfections dans leur écriture, leur parole et leurs mouvements oculaires plus proches des premiers stades de la maladie. C'est pourquoi, dans cette thèse, les patients seront étudiés après avoir pris le médicament L-dopa.

Le deuxième objectif de la thèse est de construire un système automatique de diagnostic de la maladie de Parkinson à partir de l'écriture. Deux approches doivent être considérées, étudiées et comparées: l'approche globale des caractéristiques artisanales et des classificateurs classiques et les caractéristiques à court terme avec l'augmentation des données et l'approche de l'apprentissage profond.

Dans cette thèse, nous nous concentrons principalement sur la détection précoce de la MP à partir de l'écriture mais nous présentons également les premières expériences de détection à partir de la parole où les approches décrites au point précédent seront également étudiées ici. Nous combinons ensuite les modalités de l'écriture et de la parole afin de compenser le manque de données et d'améliorer la fiabilité de la détection de la maladie.

## 10.2 La maladie de Parkinson

Parkinson est une maladie neurodégénératif causé par des neurones dopaminergiques endommagés ou morts dans la Substantia Nigra, une zone du cerveau située dans les noyaux basaux. Il y a une Substantia Nigra du côté droit du cerveau et une du côté gauche, et souvent un côté est affecté avant l'autre. De ce fait, les personnes atteintes de la maladie de Parkinson ressentent souvent les symptômes principalement d'un côté de leur corps, en particulier dans les premiers stades [Weiner et al., 2013]. Les symptômes de la maladie de Parkinson ne deviennent perceptibles qu'après la mort d'environ 80 % des cellules de la substance noire, car le système nerveux humain comporte de multiples facteurs de sécurité et redondances. Pendant longtemps, ces facteurs de sécurité ont pu prendre le relais des activités des cellules mourantes.

Les symptômes caractéristiques de la MP se répartissent en symptômes moteurs et non moteurs. Parmi les symptômes moteurs, les plus courants sont: Tremblements, lenteur des mouvements, rigidité musculaire, micrographie, trouble vocal, trouble de la saccade. En ce qui concerne les symptômes non moteurs, les plus courants sont: dépression et anxiété, constipation, problèmes de communication, démence et autres problèmes cognitifs. Ces symptômes peuvent se manifester à des degrés et selon des combinaisons variables selon les individus.

Aucun test de laboratoire ou étude radiologique définitif n'est disponible pour diagnostiquer la maladie de Parkinson, mais certains tests ou scanners d'imagerie (tels que l'imagerie par résonance magnétique (IRM) et la tomographie axiale informatisée (TAO)) peuvent être utilisés pour exclure la maladie de Parkinson. Les changements cérébraux à l'origine de la maladie de Parkinson sont microscopiques, au niveau chimique, et ne sont pas révélés par ces scanners [Weiner et al., 2013]. Certains types de tomographie par émission de positrons (TEP) et de tomographie par émission monophotonique (TEMP) sont utilisés pour évaluer le système de dopamine dans le cerveau. Toutefois, ces deux types de scanners ne sont pas largement disponibles et sont très coûteux. Ainsi, le diagnostic doit être basé sur le jugement clinique du médecin, en rassemblant les indices historiques et les résultats d'un examen physique profond.

Les patients atteints de la maladie de Parkinson souffrent de difficultés dans la coordination et le contrôle des différents systèmes musculaires. De l'autre côté, les doigts, le poignet et le bras génèrent des composantes spécifiques des mouvements d'écriture [Teulings et al., 1997]. Les doigts produisent le mouvement vertical comme dans les mouvements de haut en bas, en se rapprochant et en s'éloignant du corps dans le plan horizontal. Le poignet produit le mouvement horizontal local comme dans les mouvements gauche-droite. L'avant-bras produit la progression horizontale de gauche à droite comme dans les lignes d'écriture horizontales étendues [Teulings et al., 1986]. Par conséquent, les troubles de la coordination chez les patients atteints de la maladie de Parkinson peuvent être détectés dans les mouvements des doigts et du poignet, et même dans les flexions et les déviations cubitales du poignet, ce qui peut contribuer aux troubles de l'écriture observés chez les patients atteints de la maladie de Parkinson [Teulings et al., 1986]. En conclusion, l'écriture est considérée comme idéale pour étudier le contrôle moteur et pour détecter la maladie de Parkinson.

La parole requiert l'intégrité et l'intégration de nombreuses activités neurocognitives, neuromotrices, neuromusculaires et musculo-squelettiques [Duffy, 2013]. Ces activités sont résumées dans Figure 10.1.

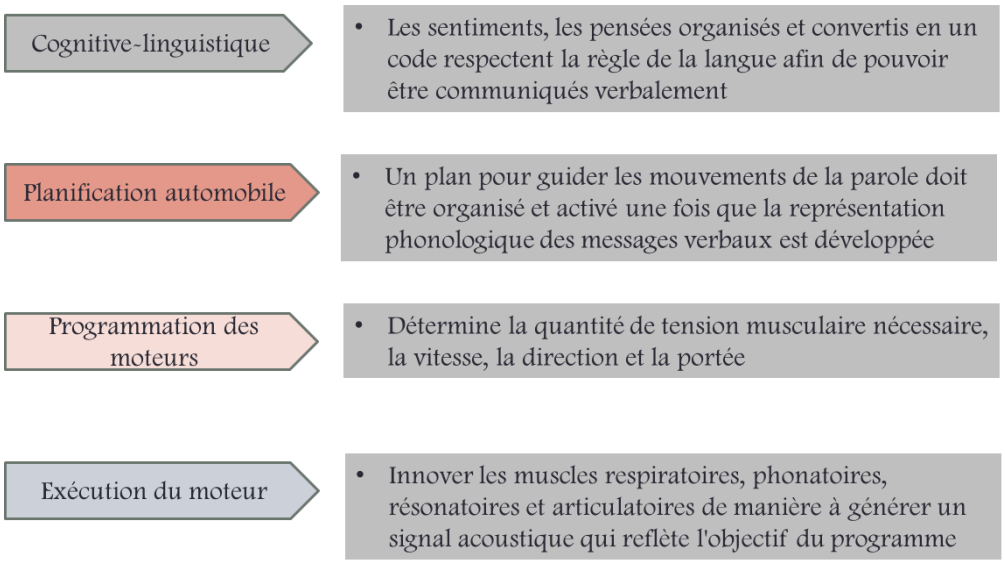



Figure 10.1. Les activités neurocognitives, neuromotrices et neuromusculaires de la parole.

Les processus combinés de planification, de programmation, de contrôle et d'exécution de la parole sont appelés processus moteurs de la parole. Lorsque le système nerveux est perturbé, la production de la parole peut l'être également. Les troubles de la parole associés à la MP sont appelés dysarthries hypokinétiques.

# 10.3 État de l'art

Cette section passe en revue les différentes méthodologies liées à l'analyse et à la caractérisation de l'écriture, de la parole et de la combinaison de multiples modalités dans la détection précoce de la MP.

## 10.3.1 Analyse de l'écriture manuscrite

De nombreuses études ont été proposées qui utilisent l'écriture pour détecter et surveiller la MP, car une écriture anormale est une manifestation bien reconnue de la MP. Les anomalies de l'écriture peuvent apparaître des années auparavant aux premiers stades de la

maladie et peuvent donc être l'un des premiers signes du MP. Plusieurs bases de données d'écriture manuscrite en ligne existent dans la littérature. Les plus cohérentes qui sont accessibles au public sont résumées dans le Tableau 3.1.

Ces études sont divisées en deux catégories: caractéristiques artisanales et classificateur classique et des approches d'apprentissage profond. Dans la première catégorie, certains auteurs ont appliqué des modèles SVM formés sur des caractéristiques artisanales globales pour la détection précoce de la MP, comme Drotar et al. [Drotar et al., 2015-a], qui ont constaté qu'une combinaison de caractéristiques cinématiques, temporelles, de pression et intrinsèques donne une précision de classification de 89.09 %, tandis que Mucha et al. [Mucha et al., 2018] ont proposé une autre approche prometteuse qui donne une précision de 97.14 % lorsqu'on utilise une combinaison de caractéristiques cinématiques et temporelles extraites à la fois pour le signal quand le stylo touche la surface de la tablette (on-paper) et le signal quand le stylo est en l’air (in-air).

Dans la deuxième catégorie, Certains auteurs ont appliqué des modèles CNN pour la détection précoce de la MP, où chacun proposait une méthode distincte pour coder les signaux dynamiques de l'écriture manuscrite en une seule image. Pereira et al. [Pereira et al., 2018] codent tous les signaux dynamiques d'écriture manuscrite en une seule image définie par l'extraction dynamique d'une image basée sur les signaux et ont rapporté une précision de classification de 93.42 %, alors que Moetesum et al. [Moetesum et al., 2019] ont retourné une précision de classification de 83 % lorsqu'ils ont exploité les 3 représentations de l'image statique des attributs de l'écriture manuscrite combinées où des CNNs sont utilisés pour l'extraction des caractéristiques, et un SVM est utilisé pour la classification. Khatamino et al. dans [Khatamino et al., 2018] ont étudié à la fois l'image dynamique extraite basée sur le signal proposée par Pereira et al. [Pereira et al., 2018] et l'image d'attributs visuels proposée par Moetesum et al. [Moetesum et al., 2019] il a été constaté que les deux représentations d'images prédisent la MP avec une précision de 88 %. Quelques autres auteurs tels que Gallicchio et al. [Gallicchio et al., 2018] ont proposé un réseau d'état d'écho profond traitant l'ensemble des signaux dynamiques d'écriture brute comme entrée pour la détection des MP, avec une précision de 89 %.

### 10.3.2 Analyse de la parole

Egalement les études dans la littérature sont divisées en deux catégories: caractéristiques artisanales et classificateur classique et des approches d'apprentissage profond. Dans la première catégorie, un certain nombre de chercheurs ont travaillé sur la classification de MP tels que Vasquez-Correa et al. [Vasquez-Correa et al., 2015] qui ont proposé différents ensembles de caractéristiques audio globales extraites de trames vocales et non vocales séparément et appliquées à un SVM pour une classification aboutissant à des précisions de 86 % et 99 % respectivement. Hariharan et al. [Hariharan et al., 2014] encore a obtenu une précision élevée atteignant 100 % en combinant l'extraction de caractéristiques globales, la pondération des caractéristiques, la sélection et la classification des caractéristiques. D'autre part, certains travaux tels que [Schuller et al., 2015] ont estimé le stade de la MP (UPDRS) en utilisant un ensemble de caractéristiques acoustiques artisanales et un apprentissage en profondeur de la régression vectorielle linéaire de soutien; où le coefficient de corrélation de Spearman (ρ) a été utilisé comme mesure d'évaluation.

Dans la deuxième catégorie, les études disponibles peuvent être divisées en deux groupes; dans le premier groupe, les modèles profonds traitent directement des signaux vocaux bruts tels que Frid et al. [Frid et al., 2016] qui ont obtenu une précision de 60 % en appliquant un 1D CNN qui traite du signal vocal brut, où le signal est copié dans des trames de taille 20 ms et le vote majoritaire a été appliqué pour la combinaison. Dans le second groupe, les modèles profonds traitent de certaines caractéristiques acoustiques artisanales au lieu de signaux bruts tels que [Caliskan et al., 2017] et [Gunduz, 2019] qui ont même appliqué un DNN ou un 1D CNN pour la classification lorsqu'un ensemble de caractéristiques artisanales a été appliqué en entrée en renvoyant des précisions proches de 86 %. Alors que, Jeancolas et al [Jeancolas et al, 2020] ont proposé une nouvelle étude pour évaluer si leur technique proposée (x-vecteurs) est meilleure que le MFCC-GMM. Il s'est avéré que les x-vecteurs, combinés à des analyses discriminantes, sont plus pertinents que la classification MFCC-GMM pour les tâches textuelle, et particulièrement adaptés à la détection de la MP chez les femmes (avec classification EER de 30 %).

Sur la base des études révisées dans cette section, il a été constaté que, dans le cas de l'analyse de la parole, les modèles profonds fonctionnent mieux avec des caractéristiques acoustiques artisanales qu'avec des signaux bruts.

### 10.3.3 Analyse multimodale

Les analyses multimodales pour une prédiction précise de la maladie de Parkinson n'ont pas fait l'objet d'études profond. Peu d'études ont appliqué des caractéristiques artisanales avec des classificateurs classiques pour la détection de la MP; Barth et al. [Barth et al., 2012] ayant combiné des modalités d'écriture et de démarche au niveau des caractéristiques globales où plusieurs classificateurs ont été étudiés, et Pham et al. [Pham et al., 2019] ayant combiné des modalités d'écriture et de parole au niveau des décisions globales où plusieurs classificateurs ont été combinés au sein des modalités.

Vásquez-Correa et al. [Vásquez-Correa et al., 2019] utilise une approche d'apprentissage profond pour l'évaluation multimodale de la MP, où ils combinent l'écriture, la parole et la démarche au niveau des caractéristiques globales où les CNN sont appliqués pour l'extraction des caractéristiques, et les SVM pour la classification.

### 10.3.4 Résumé et conclusions

Basé sur la littérature, il manque une analyse à laquelle il faudrait remédier, comme le développement d'un système multilingue robuste assisté par ordinateur capable d'évaluer la MP avec de hautes performances à partir d'un vecteur multimodal de caractéristiques, où les caractéristiques globales et le classificateur classique avec une sélection efficace de caractéristiques et des caractéristiques à court terme et des approches d'apprentissage profond sont étudiés et comparés, et où des techniques appropriées d'augmentation des données sont appliquées pour surmonter la limitation des données.

## 10.4 Construction de notre base de données multimodale sur la maladie de Parkinson

Comme nous l'avons vu dans la section 10.2, les troubles de l'écriture, de la parole et des mouvements oculaires surviennent au stade précoce de la maladie et peuvent être utilisés pour une détection précoce. Sur cette base et pour atteindre l'objectif de détection précoce et de suivi du développement de la maladie de Parkinson par le biais de signaux multimodaux, une base de données multimodale (que nous appelons la Collection multimodale de la maladie de Parkinson (PDMultiMC)) a été construite. Cette base de données, qui sera bientôt disponible sur l'IAPR TC11, comprend des données démographiques et des caractéristiques cliniques, des enregistrements en ligne de l'écriture (HandPDMultiMC), de la parole (SpeechPDMultiMC) et des mouvements oculaires (EyePDMultiMC) recueillis auprès de 21 patients atteints de la MP et 21 sujets de contrôle sains (CS) qui sont appariés en fonction de l'âge, des années d'éducation et de la dominance de la main. Les patients atteints de la maladie de Parkinson ont été sélectionnés parmi ceux qui ont consulté un neurologue expérimenté à l'hôpital Saint George- Liban, où les sujets témoins ont été choisis dans mon entourage. Les patients sélectionnés ont été examinés dans leur état " hors médication " (avant de prendre le médicament dopaminergique), et " sous médication " (1 heure après avoir pris leur dose régulière de médicament dopaminergique).

Cette base de données comprend des échantillons en trois langues: l'arabe, le français et l'anglais (où la représentation n'est pas équilibrée). Par souci de confidentialité, chaque sujet est représenté par un numéro d'identification et non par son nom. Les caractéristiques démographiques (âge, sexe, dominance de la main, années d'études) et cliniques (mini-examen de l'état mental (MMSE), échelle unifiée d'évaluation de la maladie de Parkinson (UPDRS), le stade de la maladie, les années de la maladie, l'état du patient (médicament en cours ou non) et le dosage de la dopamine) concernant chaque sujet sont enregistrées.

Pour collecter cette base de données, une application a été développée. Cette application est composée de quatre parties: la première partie consiste à remplir un questionnaire avec les caractéristiques démographiques et cliniques de chaque participant. La deuxième partie de l'application consiste à saisir les données d'écriture manuscrite en ligne

tout en réalisant un modèle d'écriture manuscrite préparé à l'aide d'une tablette de numérisation. Chaque point capturé contient des informations multidimensionnelles. Il contient des informations sur la trace de la pointe du stylet (coordonnées X-Y-Z), la pression de la pointe du stylet sur la surface, les angles du stylet par rapport à la tablette (altitude et azimut) et l'horodatage [Naik, 2012]. Le modèle d'écriture se composait de deux parties: la première partie est l'écriture libre, ou les sujets devaient écrire leur prénom et leur nom avec leur langue familière 5 fois avec leur propre vitesse et taille chaque fois sur une ligne différente. Dans la partie copie, 3 différents motifs en boucle (lettre 'l' répétitive, ondes triangulaires et rectangulaires) avec les mots ' lundi ' et ' mardi ' dans 3 langues différentes sont imprimes sur le côté gauche du papier placée sur la tablette. Les sujets devaient commencer à copier les motifs et procéder de gauche droite jusqu'à ce qu'ils aient termine 10 cycles, puis ils devaient copier 'lundi' et 'mardi' 5 fois de suite avec leur langue familière. La segmentation des mots et des motifs a été effectuée manuellement afin d'extraire différentes caractéristiques liées aux mots ou aux motifs existant dans le segment (voir Figure 4.1-b). Sept tâches d'écriture ont été étudiées dans cette thèse; où la lettre cursive répétitive (lettre 'l'), l'onde triangulaire, l'onde rectangulaire, le "lundi" répétitif, le "mardi" répétitif, le nom du sujet répétitif et le nom de famille du sujet répétitif représentent respectivement les sept tâches.

La troisième partie est un enregistrement vocal, en utilisant le microphone interne du PC, où chaque participant doit accomplir deux tâches différentes: produire une seule voyelle "a" et maintenir la hauteur de celle-ci aussi constante que possible, aussi longtemps que possible, et lire un texte apparaissant sur l'écran du PC écrit dans un langage familier.

Enfin, la dernière partie de l'application est l'enregistrement des modifications des caractéristiques du visage. Dans cette partie, le participant sera invité à lire le même texte que celui utilisé dans la partie précédente qui apparaîtra sur l'écran du PC pendant que la webcam commencera à enregistrer le changement de caractéristiques du visage.

Dans cette thèse, en raison du temps limité, nous nous sommes concentrés sur l'analyse de l'écriture et de la parole pour la détection précoce de la MP. Les enregistrements des mouvements oculaires constituent une partie supplémentaire de la base de données qui n'a pas été étudiée dans le cadre de cette thèse, mais qui peut être utilisée pour des travaux futurs.

## 10.5 Détection précoce automatique non invasive de la maladie de Parkinson basée sur l'écriture manuscrite

La détection de la maladie de Parkinson par l'analyse de l'écriture représente la plus grande partie de notre thèse, où elle est divisée en deux parties: les caractéristiques artisanales globales et le classificateur SVM, et les caractéristiques à court terme et l'apprentissage profond. Notre objectif ici est de construire un modèle indépendant du langage pour la détection précoce de la maladie de Parkinson aux premiers stades où les symptômes moteurs ne sont pas graves, en se basant sur les caractéristiques de l'écriture, où les échantillons d'écriture sont prélevés dans le sous-ensemble HandPDMultiMC décrit à la section 10.4. Il est important d'étudier les patients atteints de la maladie de Parkinson dans leur état actuel, car le traitement à la dopamine peut réduire les symptômes moteurs, et de construire un ensemble de caractéristiques générales, indépendantes du langage et adaptées à chaque tâche évaluée.

### 10.5.1 Caractéristiques artisanales globaux et SVM classificateur

#### 10.5.1.1 Classification de la MP par rapport aux CS

La principale contribution de cette étude est de trouver une approche de sélection des caractéristiques pour une détection précoce améliorée de la MP basée sur les caractéristiques de l'écriture manuscrite suggérées par Drotar et al. [Drotar et al. 2015a], [Drotar et al. 2015b], où il a été proposé d'appliquer une superposition en boucle avant l'extraction des caractéristiques afin d'obtenir des caractéristiques plus précises (voir Figure 5.5). Comme notre base de données est de petite taille, nous avons choisi pour commencer un SVM pour la classification de la MD qui peut être appris sur un petit nombre d'échantillons.

Les données brutes acquises par le numériseur ne sont pas améliorées au moyen des algorithmes standard de traitement du signal: filtrage, réduction du bruit et lissage, car cela pourrait entraîner la perte d'informations importantes qui peuvent jouer un rôle important dans la détection de la maladie.

Les symptômes moteurs peuvent se manifester à des degrés et selon des combinaisons variables selon les individus. L'analyse d'un seul de ces symptômes est insuffisante pour détecter la maladie de Parkinson, car tous les patients ne développent pas les mêmes symptômes. Pour y remédier, nous avons essayé de déterminer quelles caractéristiques de l'écriture permettent d'évaluer au mieux la plupart des symptômes moteurs. La relation entre les symptômes moteurs caractéristiques et les mesures de l'écriture manuscrite est présentée dans la Figure 5.1. Ces mesures sont réparties en 5 groupes: caractéristiques de l'accident vasculaire cérébral, caractéristiques cinématiques et temporelles, caractéristiques de pression, caractéristiques d'entropie et d'énergie, et caractéristiques intrinsèques. Nous avons extrait ces caractéristiques lorsque le stylo touche le papier. Les valeurs extraites peuvent être soit une valeur unique, soit une séquence de valeurs extraites à travers le temps [Drotar et al., 2013]. Dans le cas où il y a une séquence résultante, 5 caractéristiques fonctionnelles de base sont calculées pour la représenter: la médiane moyenne, l'écart-type, le 1er percentile et le 99ème percentile [Drotar et al., 2015a]. Il en résulte un vecteur de caractéristiques globales de taille 189 extraits pour chaque tâche.

Les tâches 1, 2 et 3 consistent en un seul segment de modèle, tandis que les tâches 4 à 7 consistent en différents segments de mots. Pour chaque segment (qu'il s'agisse d'un mot ou d'un modèle), les 189 caractéristiques décrites ci-dessus sont extraites. L'objectif est de former un vecteur de caractéristiques par tâche avec les informations de tous les segments. Pour ce faire, pour chaque tâche, nous extrayons les caractéristiques par segment, puis nous calculons la moyenne des différents segments. Une fois que les vecteurs d'entités sont extraits pour chaque tâche, nous obtenons le vecteur d'entité moyen pour les différentes tâches. Une vue d'ensemble du système d'extraction des caractéristiques est présentée dans Figure 5.4.

Une approche de sélection des caractéristiques en deux étapes a été proposée pour supprimer les caractéristiques non pertinentes. Une approche en deux étapes est proposée pour la sélection des caractéristiques. La première étape est une analyse statistique pure des données où des tests t et des tests de Mann-Whitney sont appliqués pour les caractéristiques distribuées normalement et non normalement respectivement. Une séquence de valeurs alpha comprises entre 0 et 1 a été testée sur chacune des 7 tâches et " toutes tâches " séparément, comme le montre la Figure 5.7, et celle ayant la meilleure précision de validation a été choisie. La deuxième étape est celle où une approche sous-optimale qui fournit une sorte de

point de repère n'est appliquée qu'aux " toutes tâches ". Les caractéristiques résultant de l'étape 1 sont utilisées seules pour classer un ensemble de CV, puis les caractéristiques seront ajoutées progressivement en sélectionnant à chaque itération celle qui donne la plus grande précision de validation.

Il convient de noter que, tant pour la sélection que pour la détection, le SVM avec le classificateur de noyau RBF est utilisé, où une recherche par grille utilisant la validation croisée a été appliquée pour sélectionner les paramètres du noyau. Une validation croisée quadruple a été utilisée avec la technique d'échantillonnage stratifié pour assurer la même distribution de classe dans tous les plis et pour garantir l'efficacité des résultats. Dans ce travail, nous avons décidé de ne pas utiliser un ensemble de tests séparé en raison de la petite taille de la base de données. En conséquence, le jeu de validation peut être considéré comme un jeu de test. Avant d'appliquer le SVM, les caractéristiques sont mises à l'échelle dans la plage [-1, 1] afin d'éviter la prédominance de caractéristiques avec des plages numériques plus importantes [Hsu et al., 2003], où les facteurs d'échelle sont obtenus à partir des données de formation et utilisés pour mettre à l'échelle les données de test. Les performances et le nombre de caractéristiques sélectionnées pour chaque tâche à l'aide de tests statistiques pour la sélection des caractéristiques sont présentés dans le Tableau 5.8. Selon le Tableau 5.8, les tâches 2 et 3 sont les plus précises dans la détection de la maladie de Parkinson. Ces tâches sont longues et quelque peu complexes et elles nécessitent la plus grande force cognitive et expliquent l'effet de la maladie sur l'écriture.

L'approche progressive sous-optimale décrite ci-dessus est appliquée aux system "toutes tâches". Un sous-ensemble plus petit de caractéristiques a donné une précision de classification de 96.87 %. Les caractéristiques sélectionnées qui offrent les meilleures performances sont une combinaison de pression, de cinématique et de corrélation entre la pression et les caractéristiques cinématiques.

Nous avons réussi à construire un modèle indépendant du langage pour le diagnostic de la maladie de Parkinson en utilisant l'analyse de l'écriture manuscrite avec une précision de 96.87 %, une sensibilité de 93.75 % et une spécificité de 100 %. D'un point de vue clinique, l'accélération et la taille de la course sont réglées par le mécanisme de contrôle du mouvement du poignet et des doigts, un mécanisme qui est inexact ou absent dans la MP. De

plus, les caractéristiques de pression peuvent fournir des informations détaillées qui ne peuvent être obtenues à partir des caractéristiques cinématiques, d'où l'importance de montrer la relation entre les caractéristiques cinématiques et de pression [Drotar et al., 2016].

#### 10.5.1.2 Prédiction de l'étape H&Y

Le stade H&Y est une mesure clinique standard de la progression de la MP. Il indique la présence et la gravité des symptômes de la maladie de Parkinson et peut être utilisé pour évaluer le parkinsonisme et quantifier le degré d'altération causé par les symptômes parkinsoniens. Les stades H&Y de chaque sujet sont indiqués dans le Tableau 5.12, où la répartition des échantillons entre les classes pour chacun des stades H&Y est présentée à la Figure 5.16.

Comme il s'agit d'une classification multi-classes, où chaque classe contient certains échantillons, l'utilisation d'un vecteur de caractéristiques globales peut être pénalisante à cet effet puisque certains détails sont lissés dans les caractéristiques extraites. Nous avons proposé une analyse à court terme où chaque tâche a été découpée en 41 segments. Deux types de fusion ont été appliqués pour obtenir la prédiction finale et sont résumés dans la Figure 5.29. Un classificateur SVM multi-classes à validation croisée quadruple avec noyau RBF est appliqué. Sur les 4 sous-échantillons, un sous-échantillon est retenu comme données de validation, un autre sous-échantillon est retenu comme données de test et 2 sous-échantillons sont utilisés comme données de formation. La validation croisée est faite par participant; ce qui signifie qu'il n'y a pas de fenêtre dans la formation, la validation et les tests se référant au même participant. Les paramètres du noyau sont sélectionnés par la méthode de recherche de grille, la technique d'échantillonnage stratifié est utilisée et les caractéristiques sont mises à l'échelle dans la plage [-1, 1] avant d'appliquer le SVM. En commençant par une fusion au niveau du score où les valeurs de score fournies par le classificateur SVM pour chacun des N segments (N est un hyper-paramètre qui varie entre 1 et 41) ont été utilisées comme caractéristiques d'entrée MLP du réseau neuronal (N×C - caractéristique dimensionnelle) où C représente le nombre de classes correspondant aux étapes H&Y. L'hyperparamètre N est sélectionné à l'aide du jeu de validation.

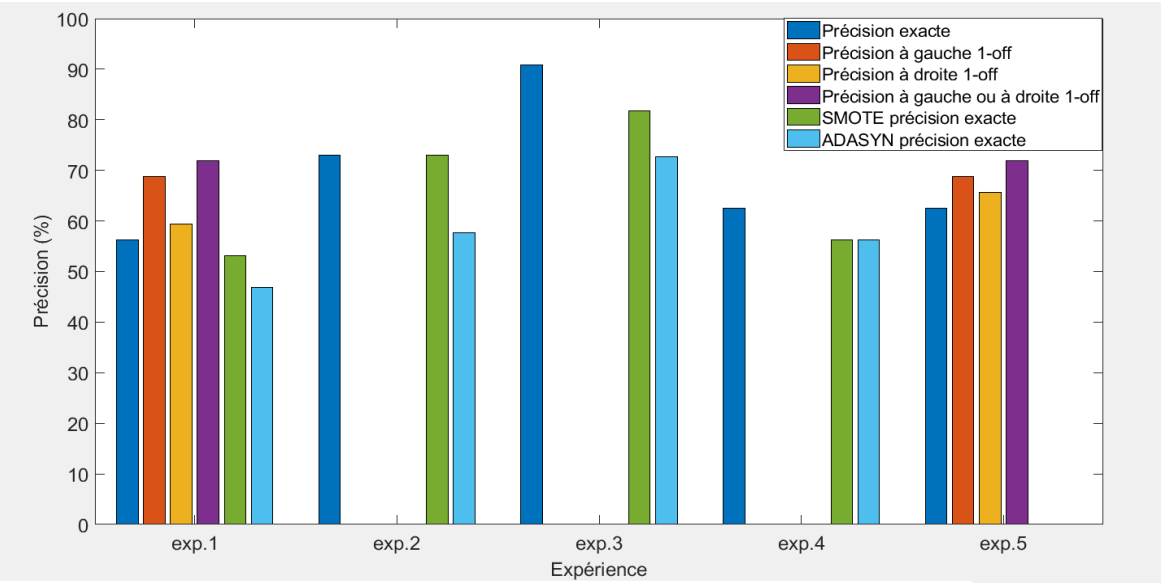
Pour chaque sujet, M observations ont été incluses; où M est le nombre de combinaisons possibles de N segments adjacents. Le modèle MLP formé prédit l'étiquette de chacune des observations M, où la fonction de coût de l'entropie croisée et l'optimiseur de descente de gradient stochastique (SGD) sont appliqués. La fusion des vecteurs de probabilité de sortie a été appliquée pour combiner les probabilités de toutes les M observations d'un sujet afin d'obtenir le vecteur de probabilité final pour la décision de classification. Le vecteur de probabilité de sortie final est calculé en obtenant la moyenne de toutes les M sorties, et le label de classe est identifié en déterminant la classe ayant la valeur la plus élevée dans le vecteur de sortie final.

De nombreux obstacles ont été rencontrés dans ce travail, tels que grande variabilité sur les patients et interférence de mouvement non pertinente, disponibilité limitée des données, données déséquilibrées, et distribution inégale des classes dans les plis. Pour réduire la grande variabilité sur les patients, nous avons décidé d'appliquer la prédiction de stade de groupe, et d'obtenir la précision 1-off (lorsque le résultat est décalé d'une étiquette de stade adjacente à gauche ou à droite). Pour surmonter l'obstacle des données déséquilibrées, nous avons décidé de n'étudier que les classes qui sont approximativement équilibrées, ou d'appliquer certaines techniques d'échantillonnage telles que ADASYN et SMOTE (voir section 5.2.4). Pour le problème de la répartition inégale des classes, nous avons décidé de n'étudier que les classes qui sont à peu près équilibrées. Enfin, pour surmonter la limitation de la taille de nos données, nous avons proposé de combiner formation et validation après sélection des paramètres. Les précisions obtenues pour chaque expérience sont résumées dans la Figure 10.2.

Nous avons découvert comment la précision unique ou la classification par groupe de stade peut permettre de surmonter la grande variabilité du problème des patients, l'importance d'avoir une distribution de classe similaire et des données équilibrées, et l'importance de la quantité de données disponibles. Nous avons également constaté que l'utilisation de méthodes de ré-échantillonnage pour rééquilibrer les données d'apprentissage n'est pas simple, car elle dépend de la facilité ou non de séparer les classes.

En conclusion, la performance obtenue dans ce travail n'est pas satisfaisante et n'est pas aussi convaincante que celle obtenue avec la classification binaire (section 10.5.1.1).

Prédire le stade de la maladie de Parkinson en utilisant des données limitées est une tâche difficile, même avec la technique du fenêtrage, surtout lorsque les patients sont étudiés en état de marche, car le médicament lévodopa réduit dans la plupart des cas les symptômes de la maladie de Parkinson, où il peut également contribuer au développement de mouvements incontrôlés. Dans les deux cas, il est difficile de séparer les stades car le patient peut même se rapprocher des stades non PD ou des stades précoces, ou des stades plus élevés. Dans le cadre de nos futurs travaux, il sera important de réaliser notre modal sur un ensemble de données large et équilibré dans les cas "sous médication " et " hors médication ".




exp1: Tous les labels de classe inclus
exp2: N'inclure que des étiquettes de classe qui assureront la même distribution de classe
exp3: Seuls les labels de classe équilibrée sont inclus
exp4: Prévision de l'étape de groupe (toutes les étiquettes de classe incluses)
exp5: Combiner la formation et la validation après la sélection des paramètres (tous les labels de classe inclus)

Figure 10.2. Effet de la grande variabilité, des données déséquilibrées, de la distribution des classes et de la taille des données sur la performance de la classification.

### 10.5.2 Caractéristiques à court terme et apprentissage profond

Cependant, comme le modèle de caractéristiques artisanales nécessite une connaissance experte du domaine, et comme nous travaillons avec une petite base de données, cela nous motive à apprendre des caractéristiques basées sur le stylo au moyen d'un apprentissage profond où l'analyse à court terme est appliquée pour éviter de perdre certaines

informations importantes en appliquant l'extraction de caractéristiques globales. Contrairement aux approches classiques d'apprentissage machine, les approches d'apprentissage profond extraient automatiquement des caractéristiques des signaux/enregistrements bruts. C'est un avantage évident puisque le modèle de réseau neuronal profond (DNN) est capable d'extraire les meilleures caractéristiques associées à chaque tâche, de manière automatique grâce à un apprentissage de bout en bout.

#### 10.5.2.1 Apprentissage profond pour la classification des séries chronologiques

Puisque chaque capteur émet le signal entier acquis pendant la tâche d'écriture, nous pouvons représenter ces données sous la forme d'une série chronologique, comme le montre la Figure 5.35 qui représente la sortie de la tâche1 d'un sujet sain et d'un patient atteint de la maladie de Parkinson. Les différences entre les dessins ne sont pas intuitivement reconnaissables, lorsque les signaux extraits du patient atteint de la maladie de Parkinson sont plus bruyants que ceux du sujet témoin. En conclusion, les signaux dynamiques d'écriture en ligne peuvent fournir des informations plus détaillées et plus complexes pour la détection de la MP. D'où l'importance d'étudier les signaux extraits plutôt que les dessins. Dans ce travail, nous avons décidé d'étudier les signaux dynamiques de l'écriture manuscrite entiers afin de pouvoir extraire des caractéristiques à la fois dans l'air et à la surface.

Ces dernières années, les réseaux neuronaux convolutifs ont montré d'excellentes performances dans les tâches de classification d'images [Gamboa et Borges, 2017]. Afin d'en tirer parti, Pereira et al. [Pereira et al., 2018] ont proposé de transformer une série temporelle en une image et de l'utiliser comme entrée pour CNN, qui sera en mesure d'apprendre les caractéristiques utilisées pour distinguer les individus en bonne santé des patients atteints de la maladie de Parkinson. D'autre part, les réseaux de mémoire à long terme et à court terme (LSTM) sont une famille de réseaux neuronaux qui excellent dans l'apprentissage à partir de données séquentielles et peuvent traiter des séries chronologiques de longueur variable [Atienza, 2017]. Les LSTM sont très populaires dans le traitement des données textuelles et ont connu un grand succès dans la traduction et la génération de textes [Burakhimmetoglu, 2017]. Comme les LSTM peuvent stocker des informations pendant de longs intervalles de

temps, ils sont donc également adaptés au traitement de séries chronologiques représentant des signaux d'écriture manuscrite [Burakhimmetoglu, 2017].

Dans ce travail, une comparaison entre les modèles 1D CNN-BLSTM et 2D CNN est effectuée. Pour le modèle 2D CNN, deux nouvelles approches ont été proposées pour coder les séries chronologiques brutes en images, et comparées à celle proposée par Pereira et al. [Pereira et al., 2018]. La fonction de coût de l'entropie croisée et l'optimiseur Adam sont appliqués aux modèles profonds proposés. Pour éviter les sur-ajustements, deux techniques ont été appliquées: l'arrêt précoce (avec 20 époques maximum sans progression) et la technique d'abandon (où une couche d'abandon est ajoutée juste après la couche cachée avec un taux d'abandon de 0.4) appliqué aux couches cachées. Pour obtenir une convergence d'entraînement plus rapide, un entraînement par mini-batch combiné à un brassage à chaque époque est appliqué [Bengio, 2012]. Puisque la technique d'abandon a été appliquée, nous avons décidé de ne pas ajouter batch-normalisation car elle remplit certains des mêmes objectifs que la technique d'abandon [Loffe et Szegedy, 2015].

Chaque tâche d'écriture est composée de n lignes (longueur de la série chronologique) et de 7 colonnes (X, Y, Z, pression, altitude, azimut et horodatage). Afin d'obtenir la meilleure combinaison de signaux, le nombre de caractéristiques des séries temporelles utilisées est un hyperparamètre variant entre 1 et 7 (nombre total de signaux extraits) et défini par k. Trois cadres ont été utilisés pour coder les séries temporelles sous forme d'images; un cadre transforme l'ensemble des données (matrice n×k) en une image (images basées sur des séries temporelles) comme proposé par Pereira et al. [Pereira et al., 2018], et les deux autres transforment chaque signal de caractéristique en une image séparée (champ angulaire gramian modifié et images de spectrogramme) (voir sections 5.3.1.2.1, 5.3.1.2.2, et 5.3.1.2.3). Certaines étapes de prétraitement ont été appliquées avant d'encoder les séries temporelles en images: obtenir la même direction d'écriture, et normaliser les coordonnées X et Y. En outre, pour tous nos modèles, 2D CNN et 1D CNN-BLSTM, toutes les images et séries temporelles brutes sont normalisées à l'intervalle (0, 1) en utilisant la normalisation min-max, où les facteurs d'échelle sont obtenus à partir des données de formation et utilisés pour mettre à l'échelle les données de test. Pour l'angle de gramian modifié, chaque série temporelle est divisée en M segments de longueur 64 et chaque segment est converti en image en utilisant la méthode GAF. Pour les images basées sur des séries temporelles et spectrogramme, la taille

de l'image convertie dépend de n, où n est la longueur de la série temporelle. La taille des séries temporelles n diffère d'une personne à l'autre et d'une collection à l'autre pour la même personne. Pour que le nombre des caractéristiques d'entrée reste identique pour la 2D CNN, la taille des images doit être identique pour tous les sujets et expériences. Dans cette étude, toutes les images sont en niveaux de gris et ont été redimensionnées à une résolution de 64×64 pixels en utilisant la technique de Lanczos [Ye et al., 2005].

Le modèle 2D CNN mis en œuvre dans le cadre de ce travail est résumé à la Figure 5.49. Ce modèle 2D CNN peut être utilisé pour la classification à partir d'une seule image comprenant k mesures (basée sur des séries chronologiques), ou pour la classification à partir de k mesures, où chaque mesure est codée dans une image (GAF et spectrogramme modifiés). Dans le cas du GAF modifié, lors d'une formation utilisant le modèle 2D CNN à entrée k représenté sur la Figure 5.49, toutes les images des tranches de la fenêtre de formation sont considérées comme des instances de formation indépendantes. Le découpage des fenêtres est également appliqué lors de la prédiction de l'étiquette d'une série chronologique de test. Le modèle 2D CNN à entrée k formé prédit l'étiquette de chacune des tranches de fenêtre. Il n'existe aucune tranche de fenêtre se référant au même participant dans la formation et le test. Pour faire la prédiction finale pour chaque sujet de la série de tests, 3 opérations différentes sont utilisées: utiliser un vote majoritaire parmi toutes ces tranches, ou obtenir la plus forte probabilité du produit des vecteurs de probabilité M, ou utiliser le modèle BLSTM représenté dans la Figure 5.50. Le second modèle proposé est le 1D CNN-BLSTM résumé dans la Figure 5.54. Ce modèle est appliqué directement sur les séries temporelles et se compose de couches 1D CNN pour l'extraction des caractéristiques et pour éviter la disparition du gradient, et de BLSTMs pour la prédiction des séquences.

Les expériences sont divisées en deux cycles: une évaluation unique et une évaluation combinée. Dans l'évaluation unique, nous analysons chaque tâche séparément, tandis que dans l'évaluation combinée, nous combinons les résultats de chaque modèle 2D CNN/1D CNN-BLSTM en utilisant le vote majoritaire afin d'obtenir le résultat final. Une approche incrémentale sous-optimale a été utilisée pour établir l'hyperparamètre k.

Les modèles décrits ci-dessus sont testés, où la performance et le nombre de mesures des paramètres appris dans le Tableau 5.17 représentent la moyenne de 3 tours (3 plis de

validation croisée où nous avons décidé de ne pas utiliser un ensemble de tests séparé en raison de la petite taille de la base de données). En considérant l'approche de combinaison de vote majoritaire. Le classificateur 2D CNN alimenté par des spectrogrammes 2D, et le 1D CNN-BLSTM alimenté par des séries temporelles (avec le plus petit nombre de paramètres) fonctionnent mieux qu'un seul 2D CNN alimenté par des images basées sur des séries temporelles, où cette représentation 2D est inspirée de Pereira et al. [Pereira et al., 2018]. Ces résultats peuvent s'expliquer par le fait que les spectrogrammes calculent les informations locales à court terme qui existent dans les signaux d'écriture manuscrite non stationnaires en ligne, et que les BLSTM apprennent la dynamique d'activation des caractéristiques temporelles. Les meilleurs ensembles de caractéristiques sélectionnés pour les deux modèles ne comprend pas l'horodatage. Cette constatation est évidente puisque les séries temporelles multivariées sont générées avec un échantillonnage fixe synchronisé sur toutes les dimensions.

Lorsqu'il s'agit de séries chronologiques d'écriture manuscrite en ligne, la variation sur l'axe du temps définit un défi pour les modèles nécessitant une entrée de dimension fixe. Les meilleurs modèles trouvés ont la capacité d'aborder la variation des informations dans les séries temporelles soit en considérant explicitement les informations locales à court terme sur l'axe du temps des signaux d'écriture manuscrite en ligne non stationnaires, soit en traitant directement les séries temporelles brutes. La classification de la MP par apprentissage profond est une tâche difficile en raison de la disponibilité limitée des données. C'est pourquoi nous avons étudié des approches d'apprentissage par transfert et d'augmentation des données basées sur ces modèles pour effectuer la détection précoce de la MP sur des données à grande échelle qui seront décrites dans la section suivante.

#### 10.5.2.2 Améliorer la détection précoce de la MP en profondeur par l'augmentation des données

La formation d'un réseau neuronal de taille adéquate avec une petite quantité de données peut amener le réseau mémoriser tous les exemples d'apprentissage ce qui entraîne une mauvaise performance sur un autre ensemble de données [Brownlee, 2019]. Ce phénomène, également connu sous le nom de "overfitting", peut être résolu en utilisant différentes techniques telles que la collecte de nouvelles données plus étiquetées (ce qui est

dans notre cas difficile à obtenir), en utilisant la méthode d'apprentissage par transfert, ou en utilisant l'augmentation des données. Des techniques d'apprentissage par transfert et d'augmentation des données pour les séries temporelles sont proposées ici pour surmonter le problème du sur-ajustement et pour augmenter la précision de reconnaissance et la robustesse des deux meilleurs modèles trouvés dans la section 10.5.2.1 et résumés dans les Figure 5.55 et Figure 5.56: le modèle 2D CNN avec la combinaison d'images spectrogrammes se référant aux caractéristiques X, Y, Z, pression et altitude comme entrée et le modèle 1D CNN-BLSTM avec la combinaison de séries temporelles brutes X, Y, Z, pression, altitude et azimut comme entrée.

L'idée derrière l'apprentissage par transfert est d'utiliser les poids et l'architecture obtenus à partir de modèles préformés qui ont été précédemment formés sur de grands ensembles de données, et d'appliquer l'apprentissage à notre problématique [Marcelino, 2019]. Nous avons appliqué le processus d'apprentissage par transfert uniquement sur le modèle 2D CNN représenté dans la Figure 5.55, car des travaux antérieurs ont montré son efficacité avec les CNN [Mormont et al., 2018]. En comparant notre propre base de données avec celles qui existent dans la littérature, nous avons sélectionné la base de données PaHaW car elle est la plus proche de notre base de données en termes de tâches et de signaux. Les différences entre les deux ensembles sont que la caractéristique des coordonnées Z est absente dans PaHaW, et que le nombre de tâches est de 8 au lieu de 7. Pour faire correspondre les deux ensembles de données, seules les tâches avec des boucles et des mots répétitifs dans PaHaW sont étudiées (les 7 premières tâches), et la caractéristique des coordonnées Z dans HandPDMultiMC est éliminée. L'ensemble des données PaHaW est utilisé pour la formation préalable à ce travail.

Différentes stratégies d'apprentissage par transfert sont étudiées et comparées pour valider les gains de l'apprentissage par transfert par rapport à la formation de notre modèle 2D CNN à partir de zéro. Ces stratégies sont résumées dans la Figure 5.58. La première stratégie d'apprentissage par transfert fige toutes les couches du modèle formées par le PaHaW et un nouveau classificateur softmax est formé en utilisant les images de formation de l'ensemble de données HandPDMultiMC. Nous avons également étudié deux stratégies de gel partiel: la première consiste à geler toute la base convolutionnelle du modèle formé par PaHaW et à recycler la partie la plus proche du classificateur en utilisant les images

d'apprentissage du jeu de données HandPDMultiMC, et la seconde consiste à geler uniquement les premières couches de la base convolutionnelle et à recycler le reste du modèle. Enfin, nous considérons la congélation complète de toutes les couches du modèle 2D CNN formé par le PaHaW.

En ce qui concerne l'augmentation des données, différentes techniques peuvent être utilisées pour générer des données artificielles: transformation géométrique (décalage, échelle, rotation/réflexion, la distorsion du temps, etc.), ajout de bruit [Um et al., 2017], et modèles générateurs profonds. En raison de l'absence de grandes bases de données dans la littérature pour former des modèles générateurs profonds, nous avons décidé de n'appliquer que des transformations géométriques et du bruit additif pour augmenter nos données. L'augmentation des données sera appliquée sur les meilleures combinaisons de séries temporelles trouvées dans la section 10.5.2.1. La gigue, la mise à l'échelle, la distorsion temporelle et les techniques de génération de données synthétiques sont utilisées pour générer de nouveaux échantillons de séries temporelles; lorsque la gigue et la mise à l'échelle sont considérées comme générant des bruits de capteurs additifs et multiplicatifs, la distorsion temporelle est un moyen de perturber la position temporelle en déformant de manière régulière les intervalles de temps entre les échantillons [Um et al., 2017], et la génération de données synthétiques fera la moyenne d'un ensemble de séries temporelles et la série temporelle moyenne peut être utilisée comme un nouvel exemple [Fawaz et al., 2018]. L'effet de l'intensité du bruit (STD) et du multiple augmenté (m) sur le travail d'augmentation des données de séries temporelles a été étudié.

Les expériences décrites dans ce travail sont également divisées en deux cycles: évaluation unique et évaluation combinée. Dans l'évaluation combinée, nous combinons les résultats des 7 modèles (un par tâche) afin de trouver le label final. Deux schémas de combinaison sont considérés, le vote majoritaire et les combinaisons basées sur le MLP; où le vote majoritaire a été utilisé dans le cas de l'apprentissage par transfert, et le modèle MLP défini dans la section 5.4.3 a été utilisé dans le cas de l'augmentation des données. La fonction de coût de l'entropie croisée et l'optimiseur Adam sont également appliqués ici, où la procédure d'arrêt précoce et la technique d'abandon (où une couche d'abandon est ajoutée juste après la couche cachée avec un taux d'abandon de 0.4), et une formation par mini-batch

combinée à un remaniement à chaque époque sont appliquées. Aucune batch-normalisation n'a été appliquée.

Sur la base des résultats obtenus et présentés dans le Tableau 5.19, il a été constaté dans la stratégie d'apprentissage par transfert que plus les couches convolutionnelles incluses dans le réglage fin sont nombreuses, plus les performances sont élevées. Cependant, il n'y a pas de gains de transfert d'apprentissage par rapport à la formation de notre modèle 2D CNN à partir de zéro, ceci peut être lié au fait que l'ensemble de données de pré-formation est unilingue, limité en taille, et n'inclut pas les coordonnées Z. En ce qui concerne la stratégie d'augmentation des données, les principaux résultats des techniques proposées sont présentés dans le Tableau 5.20. La meilleure précision est obtenue lorsque les données d'apprentissage sont augmentées deux fois. Nous avons constaté que la mise à l'échelle ne parvient pas à améliorer les performances du 1D CNN-BLSTM parce que la modification de l'intensité du signal peut modifier les étiquettes [Um et al., 2017]. D'autre part, la gigue, la distorsion du temps, et la création de séries temporelles synthétiques en calculant la moyenne d'un ensemble de séries temporelles utilisées avec le modèle 1D CNN-BLSTM améliorent la précision de la classification de la MP de 7.15 %. Cependant, l'augmentation des données ne parvient pas à améliorer les performances de 2D CNN puisque les séries temporelles ne sont pas appliquées directement au modèle, qui ne bénéficiera pas au maximum des techniques d'augmentation des données pour les séries temporelles.

Nous réalisons également des expériences en combinant les résultats obtenus à partir du modèle 1D CNN-BLSTM avec les meilleures méthodes d'augmentation des données trouvées en utilisant un modèle MLP, où les résultats sont présentés dans le Tableau 5.22. Nous avons constaté que la combinaison des résultats de deux différentes méthodes d'augmentation des données montre de meilleures performances que celle d'une seule augmentation des données. La précision la plus élevée est de 97.62 % obtenue en combinant les méthodes d'augmentation de données la gigue et la création de séries temporelles synthétiques. Cependant, lorsque nous combinons les résultats de trois méthodes d'augmentation des données, nous constatons que l'existence de la méthode d'augmentation La distorsion du temps dans la combinaison détériore la performance (une diminution de 97.62 % à 92.86 %). D'un point de vue clinique, la distorsion entre les échantillons ressemble aux perturbations temporelles inter-échantillons; l'un des premiers signes de la MP. Le

meilleur modèle final qui classifie les personnes atteintes de la MP et les témoins sains avec une précision de 97.62 % est résumé dans la Figure 5.64.

Afin de pouvoir comparer avec d'autres études, nous avons mené deux autres expériences, la première consistant à former et à tester notre meilleur modèle sur la base de données PaHaW et la seconde sur le jeu de données HandPDMultiMC où les coordonnées Z ont été exclues. Les performances obtenues dans les deux expériences sont indiquées dans le Tableau 5.24. Ces résultats confirment l'importance de la caractéristique des coordonnées Z (Le tremblement est plus évident lorsque la main est levée) et la pertinence des résultats obtenus.

Nous avons prouvé dans cette section que malgré la taille limitée de notre ensemble de données, une analyse à court terme avec un apprentissage profond et une augmentation des données donne des résultats intéressants qui dépassent le modèle SVM formé sur les caractéristiques globales définies dans la section 10.5.1.1.

## 10.6 Effet du taux d'échantillonnage de la voix et des sons non vocaux sur les performances de détection précoce de la maladie de Parkinson

La partie suivante de cette thèse consiste en une détection précoce de la MP basée sur l'analyse de la voix, où le but de ce travail est de construire un ensemble de caractéristiques acoustiques pour l'évaluation des troubles moteurs chez les patients atteints de maladie de Parkinson, où ces caractéristiques combinant différents aspects (phonation, articulation et prosodie) sont indépendantes du langage et adaptées à chaque tâche évaluée. Cependant, les informations prosodiques, y compris l'intonation, le stress et le rythme, sont considérées comme spécifiques au langage et diffèrent d'un langage à l'autre [Pinto et al., 2017]. C'est pourquoi nous avons décidé de n'étudier que les aspects liés à la phonation et à l'articulation. D'autre part, la littérature ne fournit pas de relation précise entre le taux d'échantillonnage de la voix (Fs) et la précision et la fiabilité de l'analyse acoustique de la voix. Dans ce travail, différentes valeurs de taux d'échantillonnage sont étudiées pour explorer l'influence sur l'analyse de la voix pour la détection de la MP.

En outre, la plupart des travaux de la littérature se sont concentrés sur les sons vocaux pour détecter la MP, et négligent les sons non vocaux. Pour valider le fait que les sons non vocaux jouent également un rôle dans la détection de la MP, nous avons décidé d'étudier les trames vocales et non vocales dans un discours connecté.

Les deux tâches vocales (décrites dans la section 10.4) tirées du jeu de données SpeechPDMultiMC sont étudiées et analysées dans le "on-state". Avant d'extraire les caractéristiques acoustiques de la parole, plusieurs étapes de prétraitement sont appliquées au signal: suppression du silence au début et à la fin de la parole, suppression de la parole qui ne se réfère pas au sujet et de chaque intervention spontanée introduite par le sujet qui n'était pas directement liée à la tâche, conversion du signal 2 canaux en signal mono, et réduction du taux d'échantillonnage du signal. Puis les caractéristiques sont extraites sur des trames de 20 ms, décalées de 10 ms, sur des fichiers audio prétraités, où une fonction de fenêtre de hamming est appliquée à la trame avant de calculer les descripteurs de bas niveau (LLD). L'ensemble des caractéristiques extraites est résumé dans le Tableau 6.1 , où la boîte à outils openSMILE a été utilisée pour l'extraction. Ces caractéristiques ont été choisies de manière à évaluer les dimensions d'articulation et de phonation de la parole, quelles que soient la langue et la tâche à évaluer.

Les caractéristiques LLD sont définies comme des caractéristiques statiques puisqu'elles dépendent uniquement des informations contenues dans le cadre. Afin d'obtenir les caractéristiques dynamiques et de réduire les artefacts, nous avons décidé d'appliquer à la fois: un filtre à moyenne mobile de la fenêtre W proposé par Schuller et al. [Schuller et al., 2006], et la régression delta du premier ordre proposée par Young [Young, 1997].

Le signal acoustique contient beaucoup de variabilité, dont certaines sont liées à la maladie de Parkinson, et d'autres variations sont liées à l'enregistrement du bruit de l'environnement, du style de parole ou des accents, etc. Avant l'extraction des caractéristiques globales, afin de minimiser les effets des variations qui ne sont pas liées à la maladie, et d'obtenir un modèle de bruit et de robustesse du langage, la normalisation en Z est appliquée aux caractéristiques de la maladie de Parkinson de chaque sujet séparément. Un vecteur de 220 caractéristiques globales a été obtenu après application de certaines fonctions statistiques (moyenne, maximum, minimum, médiane et écart-type) sur les LLDs.

L'approche de sélection des caractéristiques en deux étapes définie dans la section 10.5.1.1 est également appliquée ici, où un classificateur SVM à triple validation croisée avec le modèle RBF a été utilisé pour la classification binaire, où nous avons décidé de ne pas utiliser un ensemble de tests séparé. Avant d'appliquer le SVM, les caractéristiques sont mises à l'échelle dans la plage [-1, 1].

La première partie des expériences consiste à étudier l'effet du taux d'échantillonnage et des trames non vocales sur la précision de chaque tâche dans la classification de la MP, où seuls des tests statistiques ont été appliqués pour la sélection des caractéristiques. Les résultats numériques obtenus sont présentés dans le Tableau 6.2. Sur la base des résultats, nous avons constaté que le taux d'échantillonnage affecte différemment la détection de la MP lorsque différentes caractéristiques et tâches sont prises en compte, et nous avons constaté que pour la tâche de voyelle "a" soutenue et la tâche de lecture de texte (sons prononcés uniquement), la performance la plus élevée a été atteinte à 24 KHz, alors que pour la lecture de texte (sons prononcés et non prononcés), 16 KHz indiquent la performance la plus élevée. Sur la base de ces résultats, l'étape suivante consiste à former un vecteur unique combinant les informations des deux tâches: la voyelle soutenue "a" (échantillonnée à 24 KHz) et la combinaison de sons vocaux et non vocaux dans la lecture du texte (échantillonnée à 16 KHz). L’approche de sélection des caractéristiques en deux étapes est appliquée. La plus grande précision de classification obtenue est de 97.62 % pour N=16 caractéristiques. Le Tableau 6.4 présente les performances "toutes tâches" obtenues avec les méthodes de sélection des caractéristiques en un et deux temps.

En conclusion, les meilleurs taux d'échantillonnage trouvés dans ce travail confirment avec la conclusion que les plus hautes fréquences linguistiquement significatives sont en dessous de 11 KHz. En outre, comme on s'y attendait, les sons non vocaux jouent également un rôle important dans la détection de la MP puisque les spasmes du ravisseur (définis par des mouvements involontaires provoquant l'ouverture du pli vocal) se produisent principalement sur les consonnes non vocales. Enfin, l'importance des coefficients MFCC pour quantifier les problèmes d'articulation de la parole et pour détecter la maladie a été démontrée.

## 10.7 Système multimodal de détection précoce de la maladie de Parkinson basé sur l'écriture et la parole

Basé sur la littérature, un modèle indépendant du langage pour détecter la MP à l'aide de signaux multimodaux n'a pas été suffisamment étudié. La raison principale pour laquelle cette thèse se concentre sur l'analyse multimodale est qu'il n'y a pas de consensus sur l'aspect (écriture, parole) le plus approprié pour aider au diagnostic de la maladie de Parkinson à un stade précoce; ainsi, la combinaison et l'analyse des signaux de l'écriture et de la parole peuvent fournir une prédiction plus précise de la MP. Dans cette section, deux approches d'apprentissage différentes sont appliquées: les approches basées sur les caractéristiques et l'apprentissage profond. L'objectif est ici d'extraire des informations importantes des deux modalités formant un vecteur multimodal qui sera utilisé pour la classification. La fusion de différentes modalités peut être effectuée à différents niveaux: niveau des données, niveau des caractéristiques à court terme, niveau des caractéristiques globales ou niveau de décision, comme le montre la Figure 7.1 [Dumas et al., 2009]. Dans ce travail, seules les fusions de caractéristiques globales et de niveau de décision sont appliquées puisque les fusions de niveau de données et de caractéristiques à court terme sont utilisées lorsque les multiples données brutes proviennent d'un même type de source de modalités ou sont synchronisées (ce qui n'est pas notre cas) [Dumas et al., 2009].

### 10.7.1 Une approche basée sur les caractéristiques

Les sept tâches d'écriture et les deux tests vocaux décrits dans la section 10.4 seront utilisés pour construire le système multimodal. Dans cette section, seule la fusion globale au niveau des caractéristiques est appliquée. Les 189 caractéristiques globales d'écriture sur papier et les 220 caractéristiques globales de voix définies dans les sections 10.5.1.1 et 10.6 sont extraites pour chaque tâche d'écriture et de parole, puis deux méthodes de combinaison différentes ont été appliquées pour combiner les vecteurs de caractéristiques des modalités: le calcul de la moyenne ou la concaténation. Sur la base des résultats obtenus dans la section 10.6, la voyelle soutenue "a" est échantillonnée à 24 KHz tandis que le texte lu est échantillonné à 16 KHz et les trames vocales et non vocales du texte lu sont étudiées. Toutes les étapes de prétraitement décrites dans la section 10.6 sont également appliquées au signal

vocal avant d'extraire les caractéristiques du LLD. La sélection des caractéristiques en deux étapes décrite à la section 10.5.1.1 est également appliquée ici pour sélectionner parmi les caractéristiques multimodales celles qui sont les plus pertinentes. Les caractéristiques sélectionnées sont ensuite utilisées pour classer les patients de la MP et les sujets de CS en utilisant le modèle SVM avec le noyau RBF, Avant d'appliquer le SVM, les caractéristiques sont mises à l'échelle dans la plage [-1, 1].

## 10.7.2 Une approche d'apprentissage profond

Puisque notre objectif dans cette thèse est de construire un modèle multilingue pour la détection précoce de la MP en utilisant des signaux multimodaux, l'idée principale est de former notre modèle avec un vecteur de caractéristiques multimodales indépendant de la langue. Pour le modèle SVM qui est formé sur des caractéristiques multimodales pré-élaborées et décrites dans la section 10.7.1, les caractéristiques extraites sont choisies de manière à être indépendantes du langage comme décrit dans les sections 10.5.1.1 et 10.6. Pour l'approche d'apprentissage profond, afin d'obtenir un vecteur de caractéristiques indépendant de la langue, le modèle est entraîné sur toutes les langues afin que les caractéristiques ne soient pas biaisées vers une langue spécifique. Encore dans ce travail, nous avons étudié les signaux dynamiques de l'écriture manuscrite entiers afin de pouvoir extraire des caractéristiques à la fois dans l'air et à la surface. Les étapes de prétraitement manuscrit et audio mentionnées dans les sections 10.5.2.1 et 10.6 ont également été appliquées dans ce travail. Pour réduire le temps de calcul, le coût et l'utilisation de la mémoire, le taux d'échantillonnage de la voix pour la voyelle "a" soutenue et la lecture de texte (trames vocales et non vocales) est fixé à 8 KHz. En outre, pour tous nos modèles, 2D CNN, 1D CNN-BLSTM et 1D CNN-MLP, toutes les images et séries temporelles brutes sont normalisées à l'intervalle (0, 1) en utilisant la normalisation min-max.

### 10.7.2.1 2D CNN/2D CNN

L'un des modèles d'apprentissage profond étudié est le modèle 2D CNN avec images spectrogrammes (résumé dans la Figure 5.49), qui est appliqué dans ce travail selon les deux modalités. On applique la transformée de Fourier à court terme, où la technique de Lanczos est appliquée pour le redimensionnement de l'image (64×64). Le nombre de signaux

dynamiques d'écriture k est un hyper-paramètre (entre 1 et 7) et la fusion des deux modalités s'effectue à deux niveaux: le niveau global des caractéristiques et le niveau de décision. Pour la fusion globale au niveau des caractéristiques, dans chaque modalité, des 2D CNN individuels sont formés par tâche. Ensuite, les caractéristiques obtenues par les couches convolutionnelles pour chaque tâche sont combinées ensemble, soit par moyenne, soit par concaténation, comme le montre la Figure 7.3. Les encastrements obtenus à partir des 2 modalités sont concaténés pour former un vecteur multimodal par sujet. Les vecteurs de caractéristiques extraites sont ensuite utilisés pour classer les patients atteints de la maladie de Parkinson et des sujets témoins sains en utilisant un réseau entièrement connecté. En outre, pour la fusion au niveau décisionnel, des 2D CNN individuels sont formés pour chaque tâche de la modalité, où deux modèles MLP sont appliqués pour combiner les vecteurs de probabilité obtenus par toutes les tâches de chaque modalité. À un stade ultérieur, un autre modèle MLP est utilisé pour combiner les deux modalités. Pour la modalité vocale, deux cas sont étudiés: le premier cas est celui où le spectrogramme est obtenu pour l'ensemble du signal audio, et le second cas est celui où le signal audio est découpé en segments de longueur fixe (ici la longueur est de 4s) et où les modèles 2D CNN sont suivis par des BLSTM pour faire la prédiction finale (voir Figure 7.4), où l'on s'assure qu'il n'y a pas de tranches de fenêtre se référant au même participant dans la formation et le test.

#### 10.7.2.2 1D CNN-BLSTM/ 1D CNN-MLP

Le deuxième modèle profond appliqué est le modèle 1D CNN-BLSTM avec des séries temporelles brutes en entrée (présenté dans la Figure 5.54), où il n'est appliqué ici qu'en modalité écriture, et remplacé par le modèle 1D CNN-MLP (présenté dans la Figure 7.5 où chaque signal audio est découpé en segments de 4s) en modalité voix pour surmonter le problème d'utilisation de la mémoire puisque nous traitons des signaux très longs et que les BLSTM consomment plus de mémoire que les MLP. Ici aussi, nous avons veillé à ce qu'il n'y ait pas de tranches de fenêtre se référant au même participant dans la formation et le test. La fusion des modalités manuscrites et vocales n'est effectuée ici qu'au niveau décisionnel, car les caractéristiques des deux modalités sont diffèrent. Les 1D CNN-BLSTM individuels 1D sont formés pour chaque tâche en modalité écriture, où les 1D CNN-MLP individuels 1D suivis par les BLSTM pour faire la prédiction finale sont formés pour chaque tâche en

modalité voix comme le montre la Figure 7.6. Également Trois modèles de MLP sont appliqués pour la combinaison comme décrit dans la section 10.7.2.1.

### 10.7.2.3 1D CNN-BLSTM/ 2D CNN

Les modèles 1D CNN-BLSTMs et 2D CNNs sont combinés dans cette section. Pour la voix, les 2D CNNs sont appliqués avec la transformée de Fourier à temps court (STFT) en entrée, tandis que pour l'écriture manuscrite, les 1D CNN-BLSTMs sont utilisés avec les signaux bruts en entrée. La fusion des modalités manuscrites et vocales est également effectuée ici au niveau décisionnel pour la même raison que celle mentionnée dans la section 10.7.2.2. Pour la modalité vocale, les deux cas discutés dans la section 10.7.2.1 sont également étudiés ici (spectrogramme du signal entier ou spectrogramme pour chaque segment 4s (où nous avons veillé à ce qu'il n'y ait pas de tranches de fenêtre se référant au même participant dans la formation et le test)). Les 1D CNN-BLSTMs et les 2D CNNs sont formés pour chaque tâche aux modalités d'écriture et de voix respectivement, comme le montre la Figure 7.7, où les MLP sont appliqués pour la combinaison. En cas de segmentation audio, les 2D CNNs sont suivis par les BLSTMs pour la prédiction finale.

### 10.7.2.4 Expériences et conclusions

Dans ce travail également, une validation croisée à 3 plis a été appliquée où nous avons décidé de ne pas utiliser un ensemble de tests séparé. En commençant par les résultats obtenus avec le classificateur SVM, une comparaison entre la précision obtenue avec les méthodes de sélection des caractéristiques en une et deux étapes pour chaque méthode combinée se trouve dans le Tableau 7.2. Une précision de classification allant jusqu'à 100 % a été atteinte avec les deux méthodes combinées, où la plupart des caractéristiques sélectionnées offrant les meilleures performances dans les deux cas comprennent des caractéristiques cinématiques, de pression et de corrélation entre les caractéristiques cinématiques et de pression pour la modalité d'écriture manuscrite, et des coefficients MFCC pour la modalité vocale; en accord avec la conclusion figurant dans les sections 10.5.1.1 et 10.6.

En passant à l'évaluation multimodale par l'apprentissage profond, où tous les modèles décrits ont été étudiés et comparés et le nombre de signaux dynamiques d'écriture k est un hyper-paramètre variant entre 1 et 7. Aussi la fonction de coût de l'entropie croisée et l'optimiseur Adam sont appliqués aux modèles profonds proposés dans ce travail, où la procédure d'arrêt précoce et la technique d'abandon (où une couche d'abandon est ajoutée juste après la couche cachée avec un taux d'abandon de 0.4), et une formation par mini-batch combinée à un remaniement à chaque époque sont appliquées. Aucune batch-normalisation n'a été appliquée. Les résultats obtenus sont résumés dans le Tableau 7.5.

Sur la base des résultats obtenus, nous avons constaté que la fusion au niveau de la décision est plus efficace que la fusion au niveau des caractéristiques en cas de travail avec des signaux non synchronisés, et nous avons également constaté que la combinaison des meilleures caractéristiques d'écriture manuscrite est la même que celle trouvée dans la section 10.5.2.1.

Après avoir sélectionné les meilleurs modèles profonds multimodaux selon les résultats présentés dans le Tableau 7.6, nous avons appliqué les techniques d'augmentation des données sélectionnées dans la section 10.5.2.2, où pour la gigue, plusieurs valeurs d'intensité du bruit sont étudiées et les données de formation sont augmentées deux fois. Les performances "toutes tâches" des deux modalités à côté du système multimodal sont indiquées dans le Tableau 7.8. Nous avons découvert que les caractéristiques audio apprises en profondeur n'ont aucun effet sur la détection de la MP, et qu'il est difficile d'apprendre des modèles acoustiques en profondeur à partir de signaux bruts, en particulier lorsque nous travaillons avec de petits ensembles de données. Nous avons également découvert que le 2D CNN avec analyse à court terme est plus efficace que l'analyse globale, et que le log-spectrogramme améliore les performances des modèles acoustiques en profondeur. Des précisions sur les tâches ont également été obtenues et présentées dans le Tableau 7.9. Une analyse rapide des résultats nous a permis de constater que la tâche de lecture de texte est plus efficace que la voyelle de phonation soutenue "a" dans la détection de la MP, car elle est plus riche en termes d'informations acoustiques et prosodiques.

Une précision de classification allant jusqu'à 100 %, qui doit être confirmée sur des données à plus grande échelle, est obtenue lorsque les informations des deux modalités sont

combinées et utilisées avec le SVM. Malgré l'amélioration observée lors de l'application du log-spectrogramme, mais les résultats sont encore insatisfaisants par rapport aux résultats obtenus avec l'écriture manuscrite. Une explication possible de ce comportement est que dans les représentations par spectrogramme, il est difficile de séparer des sons simultanés car ils s'additionnent tous en un tout distinct. De plus, le déplacement vertical d'un son dans un spectrogramme peut en influencer la signification. Par conséquent, l'invariance spatiale fournie par les 2D CNNs pourrait ne pas être aussi performante pour cette forme de données. Enfin, la recherche de caractéristiques locales dans les spectrogrammes utilisant des convolutions 2D sera compliquée dans ce cas [Lonce, 2017]. Enfin, en cas de travail avec une petite base de données, il peut être plus efficace de construire un modèle profond en utilisant certains descripteurs acoustiques de bas niveau au lieu d'utiliser un signal vocal, ce qui peut permettre de construire des fonctions de parole profonde sans avoir besoin d'agrandir le modèle.

#### 10.7.2.5 Corrélation entre les caractéristiques artisanales et ceux apprises en profondeur

Pour améliorer l'interprétabilité des caractéristiques profondes extraites pour l'écriture et l'acoustique, nous avons effectué une analyse de corrélation entre les caractéristiques obtenues par les couches convolutives du modèle 2D CNN (illustré à la Figure 5.49 avec k images d'entrée du spectrogramme, où k est égal à 7 pour l'écriture et à 1 pour la parole), et les caractéristiques artisanales définies aux sections 5.1 et 6.5. Les caractéristiques apprises et artisanales ont été obtenues dans chaque modalité pour chaque tâche séparément, formant deux matrices de taille (294, 114688) et (294, 189) pour l'écriture manuscrite et deux autres matrices de taille (84, 16384) et (84, 220) pour la parole. Dans cette section, pour simplifier, nous avons décidé d'échantillonner toutes les tâches vocales à 8 KHz. Deux techniques ont été appliquées pour déterminer la force de la relation entre les caractéristiques profondes et les caractéristiques artisanales. La première technique consiste à obtenir la corrélation entre chaque caractéristique artisanales et chaque caractéristique apprise en profondeur (ce qui donne une matrice de corrélation de taille (189, 114688) pour l'écriture manuscrite et une autre de taille (220, 16384) pour la parole). Le plan de corrélation pour les deux modalités sont présentées respectivement dans la Figure 7.8-a et la Figure 7.9-a, où nous ne montrons

que les 189 caractéristiques profondes d'écriture les plus corrélées et les 220 caractéristiques profondes acoustiques les plus corrélées. Dans ces plans thermiques, les couleurs bleue et rouge indiquent respectivement une corrélation positive et négative. Plus la couleur est foncée, plus la relation est forte. Nous pouvons considérer que les caractéristiques profondes et artisanales sont fortement corrélées pour l'écriture manuscrite, alors que pour la parole, la corrélation n'est pas aussi forte que pour l'écriture manuscrite. La deuxième technique appliquée est la régression linéaire multiple (MLR), où nous avons fait l'hypothèse que: Y=AX (Y: caractéristiques artisanales et X: caractéristiques apprise par apprentissage profond). Chaque caractéristique artisanale est régressée séparément sur les caractéristiques apprises en profondeur. Les caractéristiques artisanales et profondes ont été normalisées afin d'obtenir des caractéristiques de moyenne nulle et de STD de un. Ensuite, l'analyse en composantes principales (PCA) a été appliquée aux caractéristiques apprises en profondeur pour réduire la taille. Les erreurs MLR obtenues sur chaque caractéristique artisanale (écriture ou parole) pour différentes réductions de la taille PCA sont présentées dans les Figure 7.8-b et la Figure 7.9-b respectivement, où nous avons veillé à ce que le nombre d'échantillons soit au moins la moitié plus que le nombre de paramètres (réduction de la taille PCA). Les meilleurs résultats ont été obtenus lorsque nous avons réduit la taille des caractéristiques profondes d'écriture à 189, et des caractéristiques profondes acoustiques à 50. Pour l'écriture manuscrite, nous pouvons voir que dans l'ensemble, les résultats obtenus sont acceptables, où certaines caractéristiques artisanales sont fortement corrélées avec les caractéristiques profondes que d'autres (celles avec des erreurs inférieures à 0.4). En revanche, pour la parole, nous pouvons constater que très peu de caractéristiques artisanales sont fortement corrélées avec les caractéristiques profondes, alors que la plupart d'entre elles ne le sont pas. D'après les résultats des deux techniques, nous pouvons dire que les caractéristiques de l'écriture manuelle et celles de l'apprentissage profond sont fortement corrélées, et que parmi les caractéristiques de l'écriture manuelle fortement corrélées, certaines font référence aux caractéristiques de pression et de cinématique, ce qui confirme l'importance de ces caractéristiques dans la détection de la MP. En ce qui concerne la parole, nous avons constaté que les caractéristiques acoustiques artisanales et apprises en profondeur ne sont pas fortement corrélées comme les caractéristiques de l'écriture (confirmé avec la conclusion trouvée dans la section 10.7.2.4 concernant la représentation du spectrogramme audio avec CNN).

# 10.8 Conclusions et travaux futurs

L'objectif de cette thèse est de construire un système multimodal indépendant du langage pour évaluer les troubles moteurs chez les patients atteints de la maladie de Parkinson à un stade précoce, basé sur des signaux combinés d'écriture et de parole, en utilisant des techniques d'apprentissage automatique. Même s'il est difficile de diagnostiquer la maladie de Parkinson à un stade précoce, de petites différences dans l'écriture et la voix sont détectables par la machine.

Les modèles classique et profond construits dans cette thèse pour la détection de la maladie de Parkinson sont considérés comme indépendants du langage puisque le signal étudié (écriture ou parole) peut être considéré comme la somme de trois composantes de base comprenant l'information linguistique, l'information sur les canaux et l'information sur les maladies. Dans le cas des modèles classiques, les caractéristiques globales ont été obtenues en appliquant certaines fonctions statistiques (telles que la moyenne, la médiane, STD, etc.) aux caractéristiques à court terme. La moyenne des informations sur les canaux est la même pour tous les sujets; la moyenne des informations linguistiques peut être considérée comme très similaire entre les sujets puisque le calcul de la moyenne peut supprimer toute la spécificité linguistique existant dans un discours. En conclusion, la moyenne du signal observé peut être considérée comme la somme d'un certain bruit avec les caractéristiques de la maladie; où le bruit n'interfère pas dans la classification. La même conclusion peut également être appliquée aux modèles profonds, car il est bien connu que de tels modèles peuvent obtenir la moyenne d'une meilleure manière que le modèle linéaire.

Cette thèse a abordé de nombreux sujets et a soulevé un certain nombre d'idées importantes qui peuvent être utilisées comme point de départ pour des études ultérieures:

- Les observations et les conclusions obtenues, ainsi que la pertinence de notre système, doivent être validées sur une base de données à plus grande échelle.

- Il existe de nombreux facteurs qui influencent l'écriture, la voix ou les mouvements des yeux et qui peuvent influer sur la décision de classification, comme le médicament L-dopa. D'autres études sont nécessaires pour

approuver les conclusions tirées de ce travail dans les cas " sous médication" et " hors médication".

- Les études futures devraient explorer les associations entre les aspects cognitifs et moteurs de la MP et l'écriture, la voix et les mouvements des yeux, car les troubles cognitifs peuvent apparaître dans certains cas avant les symptômes moteurs.

- Effectuer notre mode de prédiction par étapes décrit à la section 10.5.1.2 sur un ensemble de données important et équilibré dans les cas "sous médication" et "hors médication".

- La construction de modèles profonds avec quelques descripteurs acoustiques de bas niveau comme entrées au lieu de signaux vocaux bruts est nécessaire pour prouver son efficacité dans la détection précoce de la MP.

- Construire un système automatique de diagnostic précoce de la MP basé sur les mouvements des yeux, et un autre basé sur la combinaison des signaux de l'écriture, de la voix et des mouvements des yeux.

# 11 References

**[Abadi et al., 2016]** Abadi et al. (2016). Tensorflow: Large- scale machine learning on heterogeneous distributed systems. arXiv**2016**, arXiv:1603.04467.

**[Abou-Abbas et al., 2017]** Abou-Abbas, L., Tadj, C., Gargour, C., & Montazeri, L. (2017). Expiratory and Inspiratory Cries Detection Using Different Signals' Decomposition Techniques. *Journal of Voice, 31*(2). doi:10.1016/j.jvoice.2016.05.015

**[Ackermann and Hughes, 2003]** Ackermann, H., & Hughes, T. (2003). Dysarthria and Dysphonia. *Neurological Disorders*, 245–248. doi: 10.1016/b978-012125831-3/50217-3

**[Alsheikh et al., 2016]** Alsheikh, M. A., Selim, A., Niyato, D., Dayle, L., Lin, S., & Tan, H-P. (2016). *Deep Activity Recognition Models with Triaxial Accelerometers*. AAAI Workshop: Artificial Intelligence Applied to Assistive Technologies and Smart Environments, Phoenix, Arizona, USA.

**[Alty et al., 2017]** Alty, J., Cosgrove, J., Thorpe, D., & Kempster, P. (2017). How to use pen and paper tasks to aid tremor diagnosis in the clinic. *Practical Neurology*, *17*(6), 456–463. doi: 10.1136/practneurol-2017-001719

**[Amador et al., 2006]** Amador, SC., Hood, AJ., Schiess, MC., Izor, R., Sereno, AB. (2006). Dissociating cognitive deficits involved in voluntary eye movement dysfunctions in Parkinson's disease patients. *Neuropsychologia ,44*(8),1475e82.

**[Atienza, 2017]** Atienza, R. (2017, May 22). LSTM by Example using Tensorflow. Retrieved from https://towardsdatascience.com/lstm-by-example-using-tensorflow-feb0c1968537

**[Bachu et al., 2009]** Bachu, R., Kopparthi, S., Adapa, B., & Barkana, B. (2009). Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy. *Advanced Techniques in Computing Sciences and Software Engineering*, 279–282. doi: 10.1007/978-90-481-3660-5_47

**[Bandini et al., 2015]** Bandini et al. (2015). Automatic identification of dysprosody in idiopathic Parkinsons disease. *Biomedical Signal Processing and Control*, *17*, 47–54. doi: 10.1016/j.bspc.2014.07.006

**[Bang et al., 2013]** Bang, Y.-I., Min, K., Sohn, Y. H., & Cho, S.-R. (2013). Acoustic characteristics of vowel sounds in patients with Parkinson disease. *NeuroRehabilitation*, *32*(3), 649–654. doi: 10.3233/nre-130887

**[Barth et al., 2012]** Barth et al. (2012). Combined analysis of sensor data from hand and gait motor function improves automatic recognition of Parkinsons disease. *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. doi: 10.1109/embc.2012.6347146

**[Benba et al., 2015]** Benba, A., Jilbab, A., & Hammouch, A. (2015). Detecting Patients with Parkin son's disease using Mel Frequency Cepstral Coefficients and Support Vector Ma-

chines. *International Journal on Electrical Engineering and Informatics, 7*(2), 297-307. doi:10.15676/ijeei.2015.7.2.10

**[Benecke et al., 1986]** Benecke, R., Rothwell, J. C., Dick, J. P. R., Day, B. L., & Marsden, C. D. (1986). Performance of Simultaneous Movements in Patients with Parkinsons Disease. *Brain*, *109*(4), 739–757. doi: 10.1093/brain/109.4.739.

**[Benesty et al., 2017]** Benesty, J., Sondhi, M. M., & Huang, Y. (2017). *Springer handbook of speech processing*. Berlin: Springer.

**[Bengio, 2012]** Bengio, Y. (2012). Practical Recommendations for Gradient-Based Training of Deep Architectures. *Lecture Notes in Computer Science Neural Networks: Tricks of the Trade,* 437-478. doi:10.1007/978-3-642-35289-8_26

**[Benke et al., 2000]** Benke, T., Hohenstein, C., Poewe, W., & Butterworth, B. (2000, September). Repetitive speech phenomena in Parkinson's disease. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1737094/

**[Berwick and Idiot, 2016]** Berwick, R., & Idiot, V. (2016). An Idiot's guide to Support vector machines (SVMs). Retrieved from http://web.mit.edu/6.034/wwwbob/svm.pdf

**[Bharadi and Kekre, 2009]** Bharadi, V. K., & Kekre, B. H. (2009). Using Component Object Model for Interfacing Biometrics Sensors to Capture Multidimensional Features. *International Journal of Intelligent Information and Database Systems*, *2*(6), 279-285.

**[Blanchet and Snyder, 2009]** Blanchet, P., & Snyder, G. (2009). Speech Rate Deficits in Individuals with Parkinson's Disease: A Review of the Literature. *Journal of Medical Speech Language Pathology*, *17* (1), 1-7.

**[Boisseau et al., 1987]** Boisseau, M., Chamberland, G., & Gauthier, S. (1987). Handwriting Analysis of Several Extrapyramidal Disorders. *Canadian Society of Forensic Science Journal*, *20*(4), 139–146. doi: 10.1080/00085030.1987.10756952

**[Boudra and Salzenstein, 2018]** Boudraa, A. O., & Salzenstein, F. (2018). Teager–Kaiser energy methods for signal and image analysis: A review. *Digital Signal Processing*, *78*, 338–375. doi: 10.1016/j.dsp.2018.03.010

**[Braak and Tredici, 2017]** Braak, H., & Tredici, K. D. (2017). Neuropathological Staging of Brain Pathology in Sporadic Parkinson's disease: Separating the Wheat from the Chaff. *Journal of Parkinsons Disease*, *7*(s1). doi: 10.3233/jpd-179001

**[Brereton, 2015]** Brereton, R. G. (2015). The t-distribution and its relationship to the normal distribution. *Journal of Chemometrics*, *29*(9), 481–483. doi: 10.1002/cem.2713

**[Brownlee, 2016]** Brownlee, J. (2016). 8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset. Retrieved from https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/

**[Brownlee, 2019]** Brownlee, J. (2019). Train Neural Networks with Noise to Reduce Overfitting. Retrieved from https://machinelearningmastery.com/train-neural-networks-with-noise-to-reduce-overfitting/

**[Burakhimmetoglu, 2017]** Burakhimmetoglu. (2017, September 19). Time series classifica-

tion with Tensorflow. Retrieved from https://burakhimmetoglu.com/2017/08/22/time-series-classification-with-tensorflow/

**[Caliskan et al., 2017]** Caliskan, A., Badem, H., Başturk, A., Yüksel, M. E. (2017). Diagnosis of the Parkinson Disease by using Deep Neural Network Classifier. *Journal of Electrical And Electronics Engineering*, *17*(2), 3311-3318.

**[Castiello, et al., 1999]** Castiello, U., Bennett, K., Bonfiglioli, C., Lim, S., & Peppard, R. (1999). The reach-to-grasp movement in Parkinson's disease: Response to a simultaneous perturbation of object position and object size. *Experimental brain research*, *125*(4). 453-462. doi: 10.1007/s002210050703

**[Chammas et al., 2018]** Chammas, E., Mokbel, C. and Likforman-Sulem, L. (2018). *Handwriting Recognition of Historical Documents with Few Labeled Data*. 13th IAPR International Workshop on Document Analysis Systems (DAS), Vienna, Austria.

**[Chenausky et al., 2011]** Chenausky, K., Macauslan, J., & Goldhor, R. (2011). Acoustic Analysis of PD Speech. *Parkinsons Disease*, 1–13. doi: 10.4061/2011/435232

**[Cheung and Klotz, 1997]** Cheung, Y. K., & Klotz, J. H. (1997). The Mann Whitney Wilcoxon Distribution using Linked Lists. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.620.8195

**[Coe and Munoz, 2017]** Coe, B. C., & Munoz, D. P. (2017). Mechanisms of saccade suppression revealed in the anti-saccade task. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1718). doi:10.1098/rstb.2016.0192

**[Dai et al., 2017]** Dai, W., Dai, C., Qu, S., Li, J., & Das, S. (2017). *Very deep convolutional neural networks for raw waveforms*. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA.

**[Darley et al., 1969]** Darley, F. L., Aronson, A. E., & Brown, J. R. (1969). Differential Diagnostic Patterns of Dysarthria. *Journal of Speech and Hearing Research*, *12*(2), 246–269. doi: 10.1044/jshr.1202.246

**[Dey, 2019]** Dey, N. (2019). *Classification techniques for medical image analysis and computer aided diagnosis*. London: Elsevier/Academic Press.

**[Doermann et al., 2010]** Doermann, D., Zotkina, E., & Li, H. (2010). GEDI-A groundtruthing environment for document images. Retrieved from https://www.researchgate.net/publication/228918668_GEDI
A_Groundtruthing_Environment_for_Document_Images

**[Dounskaia et al., 2009]** Dounskaia, N., Gemmert, A. W. V., Leis, B. C., & Stelmach, G. E. (2009). Biased wrist and finger coordination in Parkinsonian patients during performance of graphical tasks. *Neuropsychologia*, *47*(12), 2504–2514. doi:10.1016/j.neuropsychologia.2009.04.020

**[Drotar et al., 2013]** Drotar, P., Mekyska, J., Rektorova, I., Masarova, L., Smekal, Z., & Faundez-Zanuy, M. (2013). *A new modality for quantitative evaluation of Parkinsons disease: In-air movement*. 13th IEEE International Conference on BioInformatics and BioEngineering, Chania, Greece.

**[Drotár et al., 2014]** Drotár, P., Mekyska, J., Rektorová, I., Masarová, L., Smékal, Z., & Faundez-Zanuy, M. (2014). Analysis of in-air movement in handwriting: A novel marker for Parkinsons disease. *Computer Methods and Programs in Biomedicine*, *117*(3), 405–411. doi: 10.1016/j.cmpb.2014.08.007

**[Drotar et al., 2015a]** Drotar, P., Mekyska, J., Smekal, Z., Rektorova, I., Masarova, L., & Faundez-Zanuy, M. (2015a). *Contribution of different handwriting modalities to differential diagnosis of Parkinsons Disease*. 2015 IEEE International Symposium on Medical Measurements and Applications (MeMeA), Torino, Italy.

**[Drotar et al., 2015b]** Drotar, P., Mekyska, J., Rektorova, I., Masarova, L., Smekal, Z., & Faundez-Zanuy, M. (2015b). Decision support framework for Parkinsons disease based on novel handwriting markers. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *23*(3). doi: 10.1109/tnsre.2014.2359997

**[Drotár et al., 2016]** Drotár, P., Mekyska, J., Rektorová, I., Masarová, L., Smékal, Z., & Faundez-Zanuy, M. (2016). Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinsons disease. *Artificial Intelligence in Medicine*, *67*, 39–46. doi: 10.1016/j.artmed.2016.01.004

**[Du and Swamy, 2013]** Du, K.-L., & Swamy, M. N. S. (2013). Multilayer Perceptrons: Architecture and Error Backpropagation. *Neural Networks and Statistical Learning*, 83–126. doi: 10.1007/978-1-4471-5571-3_4

**[Dubey et al., 2014]** Dubey, R., Zhou, J., Wang, Y., Thompson, PM., & Ye, J. (2014). Analysis of sampling techniques for imbalanced data: An n = 648 ADNI study. *Neuroimage, 87*, 220-241.

**[Duffy, 2000]** Duffy, J. (2000). Motor Speech Disorders: Clues to Neurologic Diagnosis, *Parkinson's Disease and Movement Disorders: Diagnosis and Treatment Guidelines for the Practicing Physician*, 35–53. doi:10.1385/1-59259-410-7:35

**[Duffy, 2013]** Duffy, J. (2013). *Motor speech disorders: Substrates, differential diagnosis, and management*. St. Louis, MO: Elsevier.

**[Dumas et al., 2009]** Dumas, B., Lalanne, D., & Oviatt, S. (2009). Multimodal Interfaces: A Survey of Principles, Models and Frameworks. *Lecture Notes in Computer Science Human Machine Interaction,* 3-26. doi:10.1007/978-3-642-00437-7_1

**[Eichhorn et al., 1996]** Eichhorn et al. (1996). Computational analysis of open loop handwriting movements in Parkinson's disease: a rapid method to detect dopamimetic effects. *Movement disorders, 11*(3), 289-297. doi:10.1002/mds.870110313

**[Etter, 2018]** Etter, D. M. (2018). *Introduction to MATLAB*. Hoboken, NJ: Pearson Education.

**[Eyben et al., 2010]** Eyben, F., Wöllmer, M., & Schuller, B. (2010). openSMILE -- The Munich Versatile and Fast Open-Source Audio Feature Extractor. *Proceedings of the International Conference on Multimedia - MM 10*. doi:10.1145/1873951.1874246

**[Fant, 1960 ]** Fant, G. (1960). Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations. *Mouton: The Hague*.

**[Fawaz et al., 2018]** Fawaz, I. H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P-A. (2018). Data augmentation using synthetic data for time series classification with deep residual networks. Retrieved from https://arxiv.org/abs/1808.02455

**[Flash et al., 1992]** Flash, T., Inzelberg, R., Korczyn, AD. (1992b). Quantitative methods for the assessment of motor performance in Parkinson's disease. In Rose, CF. (ed). *Parkinson's disease and problems in clinical trials* (pp. 87-106). Smith-Gordon, London.

**[Fox and Ramig, 1997]** Fox, C. M., & Ramig, L. O. (1997). Vocal Sound Pressure Level and Self-Perception of Speech and Voice in Men and Women with Idiopathic Parkinson Disease. *American Journal of Speech-Language Pathology*, *6*(2), 85–94. doi: 10.1044/1058-0360.0602.85

**[Fraser et al., 2017]** Fraser, K. C., Fors, K. L., Kokkinakis, D., & Nordlund, A. (2017). *An analysis of eye-movements during reading for the detection of mild cognitive impairment*. 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark.

**[Frid, et al., 2016]** Frid, A., Kantor, A., Svechin, D., & Manevitz, L. M. (2016). *Diagnosis of Parkinsons disease from continuous speech using deep convolutional networks without manual selection of features*. 2016 IEEE International Conference on the Science of Electrical Engineering (ICSEE), US.

**[Gallicchio et al., 2018]** Gallicchio, C., Micheli, A., & Pedrelli, L. Deep Echo State Networks for Diagnosis of Parkinson's Disease. arXiv**2018**, arXiv:1802.06708.

**[Gamboa and Borges, 2017]** Gamboa & Borges, J. C. (2017). Deep Learning for Time-Series Analysis. Kaiserslauter Univ., Kaiserslautern, Germany.

**[Glorot and Bengio, 2010]** Glorot, X., & Bengio, Y. (2010). *Understanding the difficulty of training deep feedforward neural networks*. 13th International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Sardinia, Italy.

**[Goberman et al., 2002]** Goberman, A. M., & Coelho, C. (2002). Acoustic analysis of Parkinsonian speech I: Speech characteristics and L-Dopa therapy. *NeuroRehabilitation*, *17*(3), 237–246. doi: 10.3233/nre-2002-17310

**[Goel et al., 2013]** Goel, G., Maguire, L., Li, Y., & Mcloone, S. (2013). Evaluation of Sampling Methods for Learning from Imbalanced Data. *Intelligent Computing Theories Lecture Notes in Computer Science*, 392–401. doi: 10.1007/978-3-642-39479-9_47

**[Golbe et al., 2012]** Golbe, L. I., Mark, M. H., & Sage, J. (2012). *Parkinsons disease handbook*. New York: American Parkinson Disease Association.

**[Gómez-Vilda et al., 2017]** Gómez-Vilda et al. (2017). Parkinson Disease Detection from Speech Articulation Neuromechanics. *Frontiers in Neuroinformatics*, *11*(56). doi: 10.3389/fninf.2017.00056

**[Goyal et al., 2014]** Goyal, V., Behari, M., Srivastava, A., Sood, S., Shukla, G. & Sharma, R. (2014). Saccadic eye movements in Parkinson′s disease. *Indian J Ophthalmol*, *62*, 538–544.

**[Gracia, 2010]** Garcia, D. (2010). Robust smoothing of gridded data in one and higher dimensions with missing values. *Computational Statistics & Data Analysis*, *54*(4), 1167–1178.

doi: 10.1016/j.csda.2009.09.020

**[Gunduz, 2019]** Gunduz, H. (2019). Deep Learning-Based Parkinson's Disease Classification Using Vocal Feature Sets. *IEEE Access*, *7*, 115540–115551. doi: 10.1109/access.2019.2936564

**[Gupta and Dishashree, 2020]** Gupta, D., & Dishashree. (2020, February 20). Fundamentals of Deep Learning - Activation Functions and their use. Retrieved from https://www.analyticsvidhya.com/blog/2020/01/fundamentals-deep-learning-activation-functions-when-to-use-them/

**[Guttag, 2012]** Guttag, J. V. (2012). *Introduction to computation and programming in Python*. Cambridge, MA: J.V. Guttag.

**[Hadeel, 2016]** Hadeel, A. (2016). 3-STFT - B3 Short Time Fourier Transform (STFT Objectives Understand the concept of a time varying frequency spectrum and the spectrogram Understand the: Course Hero. Retrieved from https://www.coursehero.com/file/16448070/3-STFT/

**[Hariharan et al., 2014]** Hariharan, H., Polat, K., & Sindhu, R. (2014). A new hybrid intelligent system for accurate detection of Parkinson's disease. *Computer Methods and Programs in Biomedicine*, *113*(3), 904–913. doi: 10.1016/j.cmpb.2014.01.004PMID:24485390

**[He et al., 2008]** He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). *ADASYN: Adaptive synthetic sampling approach for imbalanced learning*. 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong.

**[He et al., 2015]** He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*. 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile.

**[Helmich et al., 2011]** Helmich, R., Janssen, M., Oyen, W., Bloem, B. & Toni, I. (2011). Pallidal dysfunction drives a cerebellothalamic circuit into Parkinson tremor. *Annals of neurology*, *69*, 269-281. doi: 10.1002/ana.22361

**[Hemmerling et al., 2016]** Hemmerling, D., Orozco-Arroyave, JR., Skalski, A., Gajda, J., & No¨th, E. (2016). *Automatic Detection of Parkinson's Disease Based on Modulated Vowels*. 17th Annual Conference of the International Speech Communication Association (INTERSPEECH), San Francisco, USA.

**[Henchcliffe and Parmar, 2018]** Henchcliffe, C., & Parmar, M. (2018). Repairing the Brain: Cell Replacement Using Stem Cell-Based Technologies. *Journal of Parkinsons Disease*, *8*(s1). doi: 10.3233/jpd-181488.

**[Hermes, 1988]** Hermes, D. J. (1988). Measurement of pitch by subharmonic summation. *The Journal of the Acoustical Society of America*, *83*(1), 257–264. doi: 10.1121/1.396427

**[Holmes et al., 2000]** Holmes, R. J., Oates, J. M., Phyland, D. J., & Hughes, A. J. (2000). Voice characteristics in the progression of Parkinson's disease. *International Journal of Language & Communication Disorders*, *35*(3), 407–418. doi: 10.1080/136828200410654

**[Hsu et al., 2003]** Hsu, CW., Chang, CC., & Lin, CJ. (2003). A practical guide to support vector classification. Department of Computer Science and Information Engineering, National Taiwan University.

**[Hunker et al., 1982]** Hunker, C. J., Abbs, J. H., & Barlow, S. M. (1982). The relationship between parkinsonian rigidity and hypokinesia in the orofacial system: A quantitative analysis. *Neurology*, *32*(7), 749–749. doi: 10.1212/wnl.32.7.749

**[Illes et al., 1988]** Illes, J., Metter, E., Hanson, W., & Iritani, S. (1988). Language production in Parkinsons disease: Acoustic and linguistic considerations. *Brain and Language*, *33*(1), 146–160. doi: 10.1016/0093-934x(88)90059-4

**[Isenberg and Conrad, 1994]** Isenberg, C., & Conrad, B. (1994). Kinematic properties of slow arm movements in Parkinsons disease. *Journal of Neurology*, *241*(5), 323–330. doi: 10.1007/bf00868441.

**[Isenkul et al., 2014]** Isenkul, M., Sakar, B., & Kursun, O. (2014). *Improved Spiral Test Using Digitized Graphics Tablet for Monitoring Parkinson's Disease*. 2nd International Conference on E-Health and TeleMedicine-ICEHTM 2014, Istanbul, Turkey.

**[Jeancolas et al., 2020]** Jeancolas et al. (2020). X-vectors: New Quantitative Biomarkers for Early Parkinson's Disease Detection from Speech. Retrieved from https://arxiv.org/pdf/2007.03599

**[Jehangir et al., 2018]** Jehangir et al. (2018). Slower saccadic reading in Parkinson's disease. *Plos One*, *13*(1). doi: 10.1371/journal.pone.0191005

**[Jeni et al., 2013]** Jeni, L. A., Cohn, J. F., & Torre, F. D. L. (2013). Facing Imbalanced Data--Recommendations for the Use of Performance Metrics. *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. doi: 10.1109/acii.2013.47

**[Jeremy, 2018-a]** Jeremy, J. (2018, October 20). Common architectures in convolutional neural networks. Retrieved from https://www.jeremyjordan.me/convnet-architectures/

**[Jeremy, 2018-b]** Jeremy, J. (2018, November 5). Setting the learning rate of your neural network. Retrieved from https://www.jeremyjordan.me/nn-learning-rate/

**[Jiménez et al., 1997]** Jiménez-Jiménez et al. (1997). Acoustic voice analysis in untreated patients with Parkinsons disease. *Parkinsonism & Related Disorders*, *3*(2), 111–116. doi: 10.1016/s1353-8020(97)00007-2

**[Jiménez-Monsalve et al., 2017]** Jiménez-Monsalve, J. C., Vásquez-Correa, J. C., Orozco-Arroyave, J. R., & Gomez-Vilda, P. (2017). Phonation and Articulation Analyses in Laryngeal Pathologies, Cleft Lip and Palate, and Parkinson's Disease. *Biomedical Applications Based on Natural and Artificial Computing Lecture Notes in Computer Science*, 424–434. doi: 10.1007/978-3-319-59773-7_43

**[Joshi et al., 2016]** Joshi, P., Vinh, & Wolf. (2016, April 7). Understanding Xavier Initialization In Deep Neural Networks. Retrieved from https://prateekvjoshi.com/2016/03/29/understanding-xavier-initialization-in-deep-neural-networks/

**[Karim, 2020]** Karim, R. (2020, January 1). Animated RNN, LSTM and GRU. Retrieved from https://towardsdatascience.com/animated-rnn-lstm-and-gru-ef124d06cf45

**[Kaslovsky and Meyer, 2010]** Kaslovsky, D. N., & Meyer, F. G. (2010). Noise Corruption Of Empirical Mode Decomposition And Its Effect On Instantaneous Frequency. *Advances in Adaptive Data Analysis*, *02*(03), 373–396. doi: 10.1142/s1793536910000537

**[Kataoka et al., 1996]** Kataoka, R., Michi, K.-I., Okabe, K., Miura, T., & Yoshida, H. (1996). Spectral Properties and Quantitative Evaluation of Hypernasality in Vowels. *The Cleft Palate-Craniofacial Journal*, *33*(1), 43–50. doi: 10.1597/1545-1569(1996)033<0043:spaqeo>2.3.co;2

**[Kekre and Bharadi, 2010]** Kekre, H. B., & Bharadi, V. A. (2010). Gabor Filter Based Feature Vector for Dynamic Signature Recognition. *International Journal of Computer Applications*, *2*(3), 74–80. https://doi.org/10.5120/639-895

**[Khan et al., 2011]** Khan, N. A., Jafri, M. N., & Qazi, S. A. (2011). *Improved resolution short time Fourier transform*. 7th International Conference on Emerging Technologies, Busan, Korea.

**[Khan, 2014]** Khan, T. (2014, March 14). Running-speech MFCC are better markers of Parkinsonian speech deficits than vowel phonation and diadochokinetic. Retrieved from http://www.diva-portal.org/smash/record.jsf?pid=diva2:705196

**[Khatamino et al., 2018]** Khatamino, P., Cantürk, I., & Özyılmaz, L. (2018). *A Deep Learning-CNN Based System for Medical Diagnosis: An Application on Parkinson's Disease Handwriting Drawings*. 2018 6th International Conference Control Engineering Information Technology, Istanbul, Turkey.

**[Kim, 2017]** Kim, P. (2017). *Matlab deep learning: with machine learning, neural networks and artificial intelligence*. New York (NY): Apress.

**[Kimming et al., 2002]** Kimmig, H., Haußmann, K., Mergner, T., & Lücking, C. H. (2002). What is pathological with gaze shift fragmentation in Parkinsons disease? *Journal of Neurology*, *249*(6), 683-692. doi:10.1007/s00415-002-0691-7

**[Kingma and Ba, 2015]** Kingma, D. P., & Ba, J. (2015). Adam: a method for stochastic optimization. 3rd International Conference for Learning Representations, San Diego,

**[Koepf and Masjed-Jamei, 2006]** Koepf, W., & Masjed-Jamei, M. (2006). A generalization of Students t-distribution from the viewpoint of special functions. *Integral Transforms and Special Functions*, *17*(12), 863–875. doi: 10.1080/10652460600856419

**[Kording et al., 2017]** Kording, K. P., Benjamin, A. S., & Farhoodi, R. (2017). The roles of machine learning in biomedical science. *Frontiers of Engineering Reports on Leading-Edge Engineering from the 2017 Symposium* (pp. 61-71). Washington, DC: The National Academic Press.

**[Kurlowicz and Wallace, 1999]** Kurlowicz, L. & Wallace, M. (1999). The Mini-Mental State Examination (MMSE). *Journal of Gerontological Nursing*, 25(5), 8-9.

**[Kyaagba, 2018]** Kyaagba, S. (2018, September 7). Dynamic Time Warping with Time Series. Retrieved from https://medium.com/@shachiakyaagba_41915/dynamic-time-warping-

with-time-series-1f5c05fb8950

**[Ladefoged, 2003]** Ladefoged, P. (2003). *Phonetic Data Analysis: An Introduction to Fieldwork and Instrumental Techniques*. Malden, MA: Blackwell.

**[Lazarus and Stelmach, 1992]** Lazarus, J.-A. C., & Stelmach, G. E. (1992). Interlimb coordination in Parkinsons disease. *Movement Disorders*, *7*(2), 159–170. doi: 10.1002/mds.870070211

**[Letanneux et al., 2014]** Letanneux, A., Danna, J., Velay, J.-L., Viallet, F., & Pinto, S. (2014). From micrographia to Parkinsons disease dysgraphia. *Movement Disorders*, *29*(12), 1467–1475. doi: 10.1002/mds.25990

**[Liou et al., 1997]** Liou et al. (1997). Environmental risk factors and Parkinsons disease: A case-control study in Taiwan. *Neurology*, *48*(6), 1583–1588. doi: 10.1212/wnl.48.6.1583

**[Little et al., 2008]** Little, M., Mcsharry, P., Hunter, E., Spielman, J., & Ramig, L. (2009). Suitability of Dysphonia Measurements for Telemonitoring of Parkinsons Disease. *IEEE Transactions on Biomedical Engineering*, *56*(4), 1015–1022. doi: 10.1109/tbme.2008.2005954

**[Liutkus, 2015]** Liutkus, A. (2015). Scale-space peak picking. *Speech Processing Team*, Inria Nancy - Grand Est: Villers-les-Nancy, France.

**[Loffe and Szegedy, 2015]** Loffe, S., & Szegedy, C. (2015). *Batch normalization: Accelerating deep network training by reducing internal covariate shift*. 32nd International Conference on Machine Learning, Lille, France.

**[Logemann et al., 1978]** Logemann, J. A., Fisher, H. B., Boshes, B., & Blonsky, E. R. (1978). Frequency and Cooccurrence of Vocal Tract Dysfunctions in the Speech of a Large Sample of Parkinson Patients. *Journal of Speech and Hearing Disorders*, *43*(1), 47–57. doi: 10.1044/jshd.4301.47

**[Lonce, 2017]** Lonce, W. (2017). *Audio spectrogram representations for processing with convolutional neural networks*. 1st International Workshop on Deep Learning for Music, Anchorage, AK, USA.

**[Ma et al., 2016]** Ma, X., Yang, H., Chen, Q., & Huang, D. (2016). DepAudioNet: An Efficient Deep Model for Audio based Depression Classification. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, AVEC '16*, 35-42. doi: 10.1145/2988257.2988267.

**[Magrinelli et al., 2016]** Magrinelli et al. (2016). Pathophysiology of Motor Dysfunction in Parkinson's Disease as the Rationale for Drug Treatment and Rehabilitation. *Parkinson's disease*, 1-18. doi: 10.1155/2016/9832839

**[Manliguez, 2016]** Manliguez, C. (2016). Generalized Confusion Matrix for Multiple Classes. doi: 10.13140/RG.2.2.31150.51523.

**[Marcelino, 2019]** Marcelino, P. (2019). Transfer learning from pre-trained models. Retrieved from https://towardsdatascience.com/transfer-learning-from-pre-trained-models-f2393f124751

**[Margolin and Wing, 1983]** Margolin, DI., & Wing, AM. (1983) Agraphia and micrographia: clinical man-ifestations of motor programming and performance disorders. *ActaPsychol*, *54*, 263-283.

**[May, 2006]** May, P. J. (2006). The mammalian superior colliculus: laminar structure and connections. *Progress in Brain Research Neuroanatomy of the Oculomotor System*, 321–378. doi: 10.1016/s0079-6123(05)51011-2

**[McKell, 2016]** McKell, K. M. (2016). The Association between Articulator Movement and Formant Trajectories in Diphthongs. Retrieved from https://scholarsarchive.byu.edu/cgi/viewcontent.cgi?article=7006&context=etd

**[McLennan et al., 1972]** McLennan, JE., Nakano, K., Tyler, HR., & Schwab, RS. (1972). Micrographia in Parkinson's disease. *J Neurol Sci*, *15*, 141–152.

**[Midi et al., 2007]** Midi, I., Dogan, M., Koseoglu, M., Can, G., Sehitoglu, M. A., & Gunal, D. I. (2007). Voice abnormalities and their relation with motor dysfunction in Parkinson's disease. *Acta Neurologica Scandinavica*, *0*(0). doi: 10.1111/j.1600-0404.2007.00965.x

**[Misra, 2019]** Misra, R. (2019). Support Vector Machines-Soft Margin Formulation and Kernel Trick. Retrieved from https://towardsdatascience.com/support-vector-machines-soft-margin-formulation-and-kernel-trick-4c9729dc8efe

**[Moetesum et al., 2019]** Moetesum, M., Siddiqi, I., Vincent, N., & Cloppet, F. (2019). Assessing visual attributes of handwriting for prediction of neurological disorders—A case study on Parkinson's disease. *Pattern Recognition Letters, 121*, 19-27. doi:10.1016/j.patrec.2018.04.008

**[Mormont et al., 2018]** Mormont, R., Geurts, P., & Marée, R. (2018). *Comparison of Deep Transfer Learning Strategies for Digital pathology*. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT.

**[Mucha et al., 2018]** Mucha et al. (2018). *Identification and Monitoring of Parkinson's Disease Dysgraphia Based on Fractional-Order Derivatives of Online Handwriting*. 41st International Conference on Telecommunications and Signal Processing (TSP), Athens, Greece.

**[Nachar, 2008]** Nachar, N. (2008). The Mann-Whitney U: A Test for Assessing Whether Two Independent Samples Come from the Same Distribution. *Tutorials in Quantitative Methods for Psychology*, *4*(1), 13–20. doi: 10.20982/tqmp.04.1.p013

**[Naik, 2012]** Naik, P. (2012, March 5). An In-Depth Look At The Brand New Wacom Intuos5. Retrieved from https://fstoppers.com/product/review-depth-look-brand-new-wacom-intuos5-6734

**[Nakano et al., 1973]** Nakano, K. K., Zubick, H., & Tyler, H. R. (1973). Speech defects of parkinsonian patients: Effects of levodopa therapy on speech intelligibility. *Neurology*, *23*(8), 865–865. doi: 10.1212/wnl.23.8.865

**[Naranjo et al., 2016]** Naranjo, L., Pérez, C. J., Campos-Roca, Y., & Martín, J. (2016). Addressing voice recording replications for Parkinson's disease detection. Expert Systems with Applications, 46, 286-292. doi:10.1016/j.eswa.2015.10.034

**[Nilashi et al., 2016]** Nilashi, M., Ibrahim, O., & Ahani, A. (2016). Accuracy Improvement for Predicting Parkinson's Disease Progression. *Scientific Reports*, *6*(1). doi:10.1038/srep34181

**[Olive, 2017]** Olive, D. J. (2017). Multiple Linear Regression. *Linear Regression*, 17-83. doi:10.1007/978-3-319-55252-1_2

**[Ordóñez et al., 2016]** Ordóñez, F., & Roggen, D. (2016). Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition. *Sensors*, *16*(1), 115. doi: 10.3390/s16010115

**[Orozco-Arroyave et al., 2015]** Orozco-Arroyave et al. (2015). Characterization Methods for the Detection of Multiple Voice Disorders: Neurological, Functional, and Laryngeal Diseases. *IEEE Journal of Biomedical and Health Informatics*, *19*(6), 1820–1828. doi: 10.1109/jbhi.2015.2467375

**[Orozco-Arroyave, 2016]** Orozco-Arroyave, J. R. (2016). *Analysis of Speech of People with Parkinsons Disease*. Berlin: Logos Berlin.

**[Pal, 2020]** Pal, A. (2020, December 15). Tutorial: Understanding dimension reduction with principal component analysis. Retrieved from https://blog.paperspace.com/dimension-reduction-with-principal-component-analysis/

**[Paliwal et al., 2011]** Paliwal, K., & Lyons, J., & Wojcicki, K. (2011). *Preference for 20-40 ms window duration in speech analysis*. 4th International Conference on Signal Processing and Communication Systems, ICSPCS'2010 – Proceedings, Australia.

**[Panday, 2018]** Pandey, P. (2018, March 05). Deep Generative Models. Retrieved from https://towardsdatascience.com/deep-generative-models-25ab2821afd3

**[Parkinson, 1969]** Parkinson, J. (1969). An Essay on The Shaking Palsy. *Archives of Neurology*, *20*(4), 441–445. doi: 10.1001/archneur.1969.00480100117017

**[Pattanayak, 2017]** Pattanayak, S. (2017). *Pro deep learning with TensorFlow: a mathematical approach to advanced artificial intelligence in Python*. Berkeley, CA: Apress.

**[Pereira et al., 2016]** Pereira, C. R., Pereira, D. R., Silva, F. A., Masieiro, J. P., Weber, S. A., Hook, C., & Papa, J. P. (2016). A new computer vision-based approach to aid the diagnosis of Parkinson's disease. *Computer Methods and Programs in Biomedicine*, *136*, 79-88. doi:10.1016/j.cmpb.2016.08.005

**[Pereira et al., 2016]** Pereira, C. R., Weber, S. A. T., Hook, C., Rosa, G. H., & Papa, J. P. (2016). *Deep Learning-Aided Parkinsons Disease Diagnosis from Handwritten Dynamics*. 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Sao Paulo, Brazil.

**[Pereira et al., 2018]** Pereira et al. (2018). Handwritten dynamics assessment through convolutional neural networks: An application to Parkinson's disease identification. *Artificial Intelligence in Medicine*, 87, 67-77.

**[Perlmutter, 2009]** Perlmutter, J. S. (2009). Assessment of Parkinson Disease Manifestations. *Current Protocols in Neuroscience*, *49*(1). doi: 10.1002/0471142301.ns1001s49

**[Petersen, 2017]** Petersen, M. (2017). *Tractography and Neurosurgical Targeting in Deep Brain Stimulation for Parkinson's Disease*. Aarhus University, Denmark.

**[Pham et al., 2019]** Pham et al. (2019). *Multimodal Detection of Parkinson Disease based on Vocal and Improved Spiral Test*. 2019 International Conference on System Science and Engineering (ICSSE), Dong Hoi.

**[Phillips et al., 1991]** Phillips, JG., Stelmach, GE., & Teasdale. N. (1991). What can indices of hand-writing quality tell us about Parkinsonian handwriting? *Hum MovSci, 10*, 301-314.

**[Pinho et al., 2018]** Pinho, P., Monteiro, L., Soares, M. F. D. P., Tourinho, L., Melo, A., & Nóbrega, A. C. (2018). Impact of levodopa treatment in the voice pattern of Parkinson's disease patients: a systematic review and meta-analysis. *CoDAS*, *30*(5). doi: 10.1590/2317-1782/20182017200

**[Pinto et al., 2017]** Pinto, S., Chan, A., Guimarães, I., Rothe-Neves, R., & Sadat, J. (2017). A cross-linguistic perspective to the study of dysarthria in Parkinson's disease. *Journal of Phonetics*, *64*, 156–167. doi: 10.1016/j.wocn.2017.01.009

**[Poluha et al., 1998]** Poluha, P., Teulings, H.-L., & Brookshire, R. (1998). Handwriting and speech changes across the levodopa cycle in Parkinson's disease. *Acta psychologica*, *100*(1), 71-84.

**[Pompili et al., 2017**] Pompili et al. (2017). Automatic Detection of Parkinson's Disease: An Experimental Analysis of Common Speech Production Tasks Used for Diagnosis. *Text, Speech, and Dialogue Lecture Notes in Computer Science*, 411-419. doi:10.1007/978-3-319-64206-2_46

**[Port, 2019]** Port, B. (2019, September 9). What brain areas are affected by Parkinson's? Retrieved from https://medium.com/parkinsons-uk/what-brain-areas-are-affected-by-parkinsons-8c14dbf30954.

**[Prashanth et al., 2016]** Prashanth, R., Roy, S. D., Mandal, P. K., & Ghosh, S. (2016). High-Accuracy Detection of Early Parkinsons Disease through Multimodal Features and Machine Learning. *International Journal of Medical Informatics*, *90*, 13–21. doi: 10.1016/j.ijmedinf.2016.03.001

**[Prechelt, 1998]** Prechelt, L. (1998). Early Stopping - But When? *Lecture Notes in Computer Science Neural Networks: Tricks of the Trade*, 55-69. doi:10.1007/3-540-49430-8_3

**[Qawaqneh et al., 2017]** Qawaqneh, Z., Mallouh, AA., & Barkana, BD. (2017). Deep Convolutional Neural Network for Age Estimation based on VGG-Face Model. arXiv preprint arXiv:1709.01664.

**[Rahn et al., 2007]** Rahn, D. A., Chou, M., Jiang, J. J., & Zhang, Y. (2007). Phonatory Impairment in Parkinsons Disease: Evidence from Nonlinear Dynamic Analysis and Perturbation Analysis. *Journal of Voice*, *21*(1), 64–71. doi: 10.1016/j.jvoice.2005.08.011

**[Rakotomamonjy, 2004]** Rakotomamonjy, A. (2004). Support vector machines and area under ROC curve. Technical Report, PSI-INSA de Rouen.

**[Raudmann et al., 2014]** Raudmann, M., Taba, P., & Medijainen, K. (2014). Handwriting speed and size in individuals with Parkinson's disease compared to healthy controls: the pos-

sible effect of cueing. Acta Kinesiologiae Universitatis Tartuensis. doi: 10.12697/akut.2014.20.04

**[Rizwan, 2018]** Rizwan, M. (2018, May 9). How to Select Activation Function for Deep Neural Network. Retrieved from https://engmrk.com/activation-function-for-dnn/

**[Rodriguez-Oroz et al., 2009]** Rodriguez-Oroz et al. (2009). Initial clinical manifestations of Parkinsons disease: features and pathophysiological mechanisms. *The Lancet Neurology*, *8*(12), 1128–1139. doi: 10.1016/s1474-4422(09)70293-5

**[Rosenblum et al., 2013]** Rosenblum, S., Samuel, M., Zlotnik, S., Erikh, I., & Schlesinger, I. (2013). Handwriting as an objective tool for Parkinson's disease diagnosis. *Journal of Neurology*, *260*(9), 2357–2361. doi: 10.1007/s00415-013-6996-x

**[Rousseeuw et al., 1993]** Rousseeuw, P. J., & Croux, C. (1993). Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association*, *88*(424), 1273–1283. doi: 10.1080/01621459.1993.10476408

**[Royston, 1993]** Royston, P. (1993). A Toolkit for Testing for Non-Normality in Complete and Censored Samples. *The Statistician*, *42*(1), 37. doi: 10.2307/2348109

**[Rusz et al., 2011-a]** Rusz et al. (2011). Acoustic assessment of voice and speech disorders in Parkinsons disease through quick vocal test. *Movement Disorders*, *26*(10), 1951–1952. doi: 10.1002/mds.23680

**[Rusz et al., 2011-b]** Rusz, J., Cmejla, R., Ruzickova, H., & Ruzicka, E. (2011). Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. *The Journal of the Acoustical Society of America*, *129*(1), 350–367. doi: 10.1121/1.3514381

**[Rusz et al., 2012]** Rusz et al. (2012). Evaluation of speech impairment in early stages of Parkinson's disease: a prospective study with the role of pharmacotherapy. *Journal of Neural Transmission*, *120*(2), 319–329. doi: 10.1007/s00702-012-0853-4

**[Rusz et al., 2013]** Rusz et al. (2013). Imprecise vowel articulation as a potential early marker of Parkinsons disease: Effect of speaking task. *The Journal of the Acoustical Society of America*, *134*(3), 2171–2181. doi: 10.1121/1.4816541

**[Sadouk, 2019]** Sadouk, L. (2019). CNN Approaches for Time Series Classification. *Time Series Analysis - Data, Methods, and Applications*. doi:10.5772/intechopen.81170

**[Sakar et al., 2013]** Sakar et al. (2013). Collection and Analysis of a Parkinson Speech Dataset With Multiple Types of Sound Recordings. *IEEE Journal of Biomedical and Health Informatics*, *17*(4), 828–834. doi: 10.1109/jbhi.2013.2245674

**[Sakar et al., 2019]** Sakar et al. (2019). A comparative analysis of speech signal processing algorithms for Parkinson's disease classification and the use of the tunable Q-factor wavelet transform. *Applied Soft Computing*, *74*, 255-263. doi:10.1016/j.asoc.2018.10.022

**[Schuller et al., 2006]** Schuller, B., Arsi, D., Wallhoff, F., & Rigoll, G. (2006). *Emotion recognition in the noise applying large acoustic feature sets*. 3rd International Conference on Speech Prosody (ISCA), Dresden, Germany.

**[Schuller et al., 2015]** Schuller et al. (2015). *The INTERSPEECH 2015 Computational Paralinguistics Challenge: Degree of Nativeness, Parkinson's & Eating Condition*. INTERSPEECH 2015, Dresden, Germany.

**[Schuller, 2013]** Schuller, B. (2013). *Intelligent Audio Analysis.* Signals and communication Technology. Springer.

**[Siegel, 2013]** Siegel, J. (2013). Handwriting assessment can be used for early detection of Parkinson's disease. Retrieved from https://www.jpost.com/Health-and-Science/Handwriting-assessment-can-be-used-for-early-detection-of-Parkinsons-disease-325798

**[Singh, 2020]** Singh, G. (2020, March 4). Demystifying LSTM Weights and Biases Dimensions. Retrieved from https://medium.com/analytics-vidhya/demystifying-lstm-weights-and-biases-dimensions-c47dbd39b30a

**[Skodda et al., 2008]** Skodda, S., & Schlegel, U. (2008). Speech rate and rhythm in Parkinsons disease. *Movement Disorders*, *23*(7), 985–992. doi: 10.1002/mds.21996

**[Skodda et al., 2011]** Skodda, S., Visser, W., & Schlegel, U. (2011). Vowel Articulation in Parkinsons Disease. *Journal of Voice*, *25*(4), 467–472. doi: 10.1016/j.jvoice.2010.01.009

**[Skodda et al., 2012]** Skodda, S., Grönheit, W., & Schlegel, U. (2012). Impairment of Vowel Articulation as a Possible Marker of Disease Progression in Parkinsons Disease. *PLoS ONE*, *7*(2). doi: 10.1371/journal.pone.0032132

**[Skodda et al., 2013]** Skodda, S., Grönheit, W., Mancinelli, N., & Schlegel, U. (2013). Progression of Voice and Speech Impairment in the Course of Parkinsons Disease: A Longitudinal Study. *Parkinsons Disease*, *2013*, 1–8. doi: 10.1155/2013/389195

**[Smite et al., 2014]** Smits et al. (2014). Standardized Handwriting to Assess Bradykinesia, Micrographia and Tremor in Parkinsons Disease. *PLoS ONE*, *9*(5). doi: 10.1371/journal.pone.0097614

**[Smith et al., 1984]** Smith, J., & Gossett, P. (1984). *A flexible sampling-rate conversion method*. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 84), San Diego, CA, USA, USA.

**[Srivastava et al., 2014]** Srivastava, A., Sharma, R., Sood, S., Shukla, G., Goyal, V., & Behari, M. (2014). Saccadic eye movements in Parkinson's disease. *Indian journal of ophthalmology, 62*, 538-44.

**[Srivastava et al., 2014]** Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929-1958.

**[Stemple et al., 2020]** Stemple, J. C., Roy, N., & Klaben, B. (2020). *Clinical voice pathology: theory and management*. San Diego, CA: Plural Publishing Inc.

**[Stevens, 1998]** Stevens, K. N. (1998). Acoustic Phonetics. Cambridge, MA: MIT Press.

**[Sudarsen et al., 2017]** Sudarsen, S., & Ravindran, B. (2017). Class Imbalance Learning – A Quarterly Publication of ACCS. *A Quarterly Publication of ACCS*. Retrieved from http://acc.digital/class-imbalance-learning-4/

**[Taleb et al., 2017]** Taleb, C., Likforman, L., Khachab, M., and Mokbel, C. (2017). *Feature Selection for an Improved Parkinson's Disease Identification Based on Handwriting*. Arabic Script Analysis and Recognition (ASAR), 2017 1st International Workshop, Nancy, France.

**[Taleb et al., 2018]** Taleb, C., Likforman, L., Khachab, M., & Mokbel, C. (2018). *A Reliable Method to Predict Parkinson's Disease Stage and Progression based on Handwriting and Re-sampling Approaches*. Arabic Script Analysis and Recognition (ASAR), 2018 2nd International Workshop, The Alan Turing Institute, London-UK.

**[Taleb et al., 2019-a]** Taleb, C., Likforman, L., Khachab, M., and Mokbel, C. (2019). *Visual Representation of Online Handwriting Time Series for Deep Learning Parkinson's Disease Detection*. Arabic Script Analysis and Recognition (ASAR), 2019 3rd International Workshop, Sydney, Australia.

**[Taleb et al., 2019-b]** Taleb C., Likforman-Sulem L., Mokbel C. (2020). Improving Deep Learning Parkinson's Disease Detection through Data Augmentation Training. In: Djeddi C., Jamil A., Siddiqi I. (eds), *Pattern Recognition and Artificial Intelligence* (pp. 79–93). Cham: Springer International Publishing.

**[Terao et al., 2013]** Terao, Y., Fukuda, H., Ugawa, Y., & Hikosaka, O. (2013). New perspectives on the pathophysiology of Parkinson's disease as assessed by saccade performance: A clinical review. *Clinical Neurophysiology*, *124*(8), 1491–1506. doi: 10.1016/j.clinph.2013.01.021

**[Teulings and Stelmach, 1991]** Teulings, HL., & Stelmach, GE. Control of stroke size, peak acceleration, and stroke duration in Parkinsonian handwriting. *Hum MovSci*, *10*, 315-334.

**[Teulings et al., 1989]** Teulings, H.-L., Thomassen, A. J. W. M., & Maarse, F. J. (1989). A Description of Handwriting in Terms of Main Axes. *Computer Recognition and Human Production of Handwriting*, 193–211. doi: 10.1142/9789814434195_0014

**[Teulings et al., 1997]** Teulings, H.-L., Contreras-Vidal, J. L., Stelmach, G. E., & Adler, C. H. (1997). Parkinsonism Reduces Coordination of Fingers, Wrist, and Arm in Fine Motor Control. *Experimental Neurology*, *146*(1), 159–170. doi: 10.1006/exnr.1997.6507

**[Theodoros et al., 1995]** Theodoros, DG., Murdoch, BE., & Thompson, EC. (1995). Hypernasality in Parkinson disease: a perceptual and physiological analysis. *Journal of Medical Speech-Language Pathology*, *3*, 73-84.

**[Titze, 2000]** Titze, I. R. (2000). *Principles of voice production*. Iowa City, IA: National Center for Voice & Speech.

**[Tjaden and Wilding, 2013]** Tjaden, K., Lam, J., & Wilding, G. (2013). Vowel Acoustics in Parkinsons Disease and Multiple Sclerosis: Comparison of Clear, Loud, and Slow Speaking Conditions. *Journal of Speech, Language, and Hearing Research, 56*(5), 1485-1502. doi:10.1044/1092-4388(2013/12-0259)

**[Tran et al., 2020]** Tran, J., Anastacio, H., & Bardy, C. (2020). Genetic predispositions of Parkinson's disease revealed in patient-derived brain cells. *Npj Parkinson's Disease*, *6*(1). doi:10.1038/s41531-020-0110-8

**[Tsanas et al., 2012]** Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2012). Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning. *IEEE Transactions on Biomedical Engineering, 57*, 884-893

**[Tseng et al., 2013]** Tseng, P., Cameron, I., Pari, G., Reynolds, J., Munoz, D., & Itti, L. (2013). High-throughput classification of clinical populations from natural viewing eye movements. *Journal of neurology, 260*, 275-284. doi:10.1007/s00415-012-6631-2

**[Tucha et al., 2006]** Tucha, O., Mecklinger, L., Thome, J., Reiter, A., Alders, G., Sartor, H., et al. (2006). Kinematic analysis of dopaminergic effects on skilled handwriting movements in Parkinson's disease. *Journal of neural transmission, 113*(5), 609-623.

**[Um et al., 2017]** Um et al. (2017). *Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks*. 19th ACM International Conference on Multimodal Interaction - ICMI 2017, Glasgow, Scotland.

**[Vapnik, 2010]** Vapnik, V. N. (2010). *The nature of statistical learning theory*. New York, NY: Springer.

**[Vasquez-Correa et al., 2015]** Vasquez-Correa, J. C., Arias-Vergara, T., Orozco-Arroyave, JR., Vargas-Bonilla, JF., Arias-Londoño, JD., & Noth, E. (2015). *Automatic Detection of Parkinson's Disease from Continuous Speech Recorded in Non-Con-trolled Noise Conditions*. 16th Annual Conference of the International Speech Communication Association (INTERSPEECH), Dresden, Germany.

**[Vásquez-Correa et al., 2019]** Vásquez-Correa, J. C., Arias-Vergara, T., Orozco-Arroyave, J. R., Eskofier, B., Klucken, J., & Nöth, E. (2019). Multimodal assessment of Parkinson's disease: A deep learning approach. *IEEE J. Biomed. Health Informat.*, 23(4), 1618–1630.

**[Velazquez, 2018]** Velazquez, L. M. (2018). Towards the differential evaluation of Parkinson's disease by means of voice and speech processing. Retrieved from http://oa.upm.es/51278/1/LAUREANO_MORO_VELAZQUEZ.pdf

**[Vidailhet et al. 1994]** Vidailhet, M., Rivaud, S., Gouider-Khouja, N., Pillon, B., Bonnet, AM., Gaymard, B. (1994). Eye movements in parkinsonian syndromes. *Annals of Neurology, 35*(4), 420-426. doi:10.1002/ana.410350408

**[Vogel et al., 2009]** Vogel, AP., Ibrahim, HM., Reilly, S., Kilpatrick, N. (2009). A comparative study of two acoustic measures of hypernasality. *Journal of Speech Language and Hearing Research, 52*, 1640-1651.

**[Voytek, 2006]** Voytek, B. (2006). Emergent Basal Ganglia Pathology within Computational Models. *Journal of Neuroscience*, *26*(28), 7317–7318. doi: 10.1523/jneurosci.2255-06.2006

**[Waldthaler et al., 2018]** Waldthaler, J., Tsitsi, P., Seimyr, G. Ö., Benfatto, M. N., & Svenningsson, P. (2018). Eye movements during reading in Parkinson's disease: A pilot study. *Movement Disorders*, *33*(10), 1661–1662. doi: 10.1002/mds.105

**[Walter, 2003]** Walter, Z. (2003). Kernel density estimation. Retrieved from http://staff.ustc.edu.cn/~zwp/teach/Math-Stat/kernel.pdf

**[Wang et al., 2014]** Wang, Z., & Xue, X. (2014). Multi-Class Support Vector Machine. *Support Vector Machines Applications*, 23–48. doi: 10.1007/978-3-319-02300-7_2

**[Wang et al., 2015]** Wang, Z., & Oates, T. (2015). *Imaging time-series to improve classification and imputation*. 24th International Join Conference on Artificial Intelligence(IJCAI), Buenos Aires, Argentina.

**[Wang et al., 2018]** Wang, F., Zhong, S., Peng, J., Jiang, J. and Liu, Y. (2018). Data Augmentation for EEG-Based Emotion Recognition with Deep Convolutional Neural Networks. In Schoeffmann, K., Chalidabhongse, T.H., Ngo, C.W., Aramvith, S., O'Connor, N.E., Ho, Y.-S., Gabbouj, M., Elgamma, A. (Eds), *MultiMedia Modeling* (pp. 82-93). Springer: Cham.

**[Weiner et al., 2013]** Weiner, W. J., Shulman, L. M., & Lang, A. E. (2013). *Parkinsons disease: a complete guide for patients and families*. Baltimore: The Johns Hopkins University Press.

**[Weismer et al., 1998]** Weismer, G., & Wildermuth, J. (1998). Formant trajectory characteristics in persons with Parkinson, cerebellar, and upper motor neuron disease. *The Journal of the Acoustical Society of America*, *103*(5), 2892–2892. doi: 10.1121/1.421814

**[Weismer, 2006]** Weismer, G. (2006). Philosophy of research in motor speech disorders. *Clinical Linguistics & Phonetics*, *20*(5), 315–349. doi: 10.1080/02699200400024806

**[Wolfe et al., 1975]** Wolfe, V., Garvin, J., Bacon, M., & Waldrop, W. (1975). Speech changes in Parkinsons disease during treatment with L-DOPA. *Journal of Communication Disorders*, *8*(3), 271–279. doi: 10.1016/0021-9924(75)90019-2

**[Yadav, 2020]** Yadav, S. (2020, January 17). Weight Initialization Techniques in Neural Networks. Retrieved from https://towardsdatascience.com/weight-initialization-techniques-in-neural-networks-26c649eb3b78

**[Ye et al., 2005]** Ye, Z., Suri, J., Sun, Y., & Janer, R. (2005). *Four image interpolation techniques for ultrasound breast phantom data acquired using Fischers full field digital mammography and ultrasound system (FFDMUS): a comparative approach*. IEEE International Conference on Image Processing 2005, Genoa, Italy.

**[Young, 1997]** Young, S. (1997). *The Htk book*. Cambridge: Entropic Cambridge Research Laboratory.

**[Yumoto et al., 1982]** Yumoto, E., Gould, W. J., & Baer, T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *The Journal of the Acoustical Society of America*, *71*(6), 1544–1550. doi: 10.1121/1.387808

**[Zafari et al., 2019]** Zafari, A., Zurita-Milla, R., & Izquierdo-Verdiguier, E. (2019). Evaluating the Performance of a Random Forest Kernel for Land Cover Classification. *Remote Sensing*, *11*(5), 575. doi: 10.3390/rs11050575

**[Zeiler and Fergus, 2013]** Zeiler, D, M., & Fergus, R. (2013, November 28). Visualizing and Understanding Convolutional Networks. Retrieved from https://arxiv.org/abs/1311.2901