TÉCNICO LISBOA

RNNs for detecting depression in Huntington's Disease

Mariana de Avelino Geraldo Dias

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisor: Prof. Maria Margarida Campos da Silveira

Examination Committee

Chairperson: Prof. Mário Jorge Costa Gaspar da Silva
Supervisor: Prof. Maria Margarida Campos da Silveira
Members of the Committee: Prof. Susana de Almeida Mendes Vinga Martins

April 2020

The work presented in this thesis was performed at the Institute of Systems and Robotics of Instituto Superior Técnico (Lisbon, Portugal), during the period March 2019-January 2020, under the supervision of Prof. Margarida Silveira.

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

First of all, I want to thank Professor Margarida Silveira for giving me the opportunity of developing this enthusiastic project and for providing orientation and all the necessary tools to make it possible.

I would also like to thank *Enroll-HD* for providing the data and Institute of Systems and Robotics (ISR-Lisboa) for welcoming this project.

A huge thanks to my parents, to my brother Afonso, to my sister Camila and to all my closest friends, without whom this journey would not have been the same.

Abstract

Huntington's Disease (HD) is a neurodegenerative disorder characterized by motor, cognitive and psychiatric progressive dysfunctions, caused by a genetic mutation on a protein whose function remains incompletely understood. The evolution of HD through time is marked by great variability, which makes it of difficult management. One highly incident psychiatric impairment in HD is depression. While, unlike other symptoms of the disease, it is not correlated to disease progression, it has been linked to greater functional damage and worse cognitive performance. Furthermore, it has an extreme impact on the quality of life of both the patient and family.

In the present study, a Deep Learning model for detecting if depression was ever a part of a patient's medical history, based on sequential clinical data, was developed. For that, longitudinal data of 9474 HD patients and 1481 controls from the Enroll-HD database was used. The gathered data comprises information from annual clinical visits where several questionnaires are answered and exams are performed, regarding the evaluation of all clinical aspects of HD.

With the main objective of understanding if it was possible to distinguish, from the evolution of the disease, cases where depression had been present from those that did not suffer from it, several Recurrent Neural Network architectures were tested. It was also observed that adding "profile" data about the patient and family contributed to an enhanced detection ability. With the implementation of a GRU model an accuracy of 80% was achieved, with a sensitivity of 85% and a specificity of 69%.

Keywords

Huntington's Disease, Depression, Deep Learning, Recurrent Neural Networks

Resumo

A doença de Huntington é uma doença neurodegenerativa caracterizada por disfunções progressivas do foro motor, cognitivo e psicológico, causadas por uma mutação genética numa proteína cuja função não é totalmente compreendida. A evolução desta doença ao longo do tempo pauta-se de uma grande variabilidade, o que a torna difícil de controlar e prever. Um distúrbio psicológico muito incidente nesta doença é a depressão. Embora, ao contrário de outros sintomas, não esteja correlacionada com a progressão da doença, tem sido associada a um pior desempenho funcional e cognitivo. Ademais, tem um impacto extremo na qualidade de vida, tanto do paciente como da família.

No presente estudo, foi desenvolvido um modelo de Aprendizagem Profunda para a deteção de um historial de depressão, baseado em dados clínicos. Para tal, foram usados dados longitudinais, provenientes da base de dados "Enroll-HD", de 9474 pessoas com HD e 1481 controlos. Os dados recolhidos abrangem informação de visitas clínicas anuais nas quais são realizados exames e vários questionários são preenchidos, no sentido de avaliar as diferentes componentes da doença.

Com o principal objetivo de perceber se seria possível distinguir, a partir da evolução da doença, casos de depressão, testaram-se várias arquiteturas de Redes Neuronais Recorrentes. Observou-se ainda que adicionar dados de perfil do paciente e respetiva família contribuia para uma capacidade de deteção melhorada. Com a implementação de um modelo de Unidades Recorrentes de Porta, conseguiu-se uma exatidão de 80%, uma taxa de verdadeiros positivos de 85% e uma taxa de verdadeiros negativos de 69%.

Palavras Chave

Doença de Huntington, Depressão, Aprendizagem Profunda, Redes Neuronais Recorrentes

Contents

List of Figures

List of Tables

Acronyms

Adam Adaptive Moment Estimation

AI Artificial Intelligence

CAG cytosine, adenine and guanine

GRU Gated Recurrent Unit

HD Huntington's Disease

HTT Huntingtin

JHD Juvenile Huntington's Disease

LSTM Long Short-Term Memory

ML Machine Learning

RNN Recurrent Neural Network

SGD Stochastic Gradient Descent

TNR True Negative Rate

TPR True Positive Rate

UHDRS Unified Huntington's Disease Rating Scale

1

Introduction

Contents

1.1 Motivation

Huntington's Disease is a neurodegenerative terminal disease for which there exists no cure. The evolution of the disease is extremely heterogeneous and is still very poorly understood. The most frequently occurring psychiatric sign is depression but no relation to disease progress has been evidenced [1], [2]. Very often the neuropsychiatric symptoms are described as one of the most distressing aspect of Huntington's disease, having a great impact in quality of life and contributing to functional decline. Suicide is estimated to be the cause of 5-10% of the deaths in HD [3].

Depression, despite being one of the most common mental disorders worldwide, after decades of research is also still incompletely demystified and many different physiological mechanisms have been linked to it [4]. Consequently, it is difficult to localize the anomalies and to make a diagnosis based on objective parameters, being usually made using standardized questionnaires and interviews which are often of subjective interpretability. Like HD, it is characterized by a heterogeneous symptomatology. For all these reasons, the clinical treatment approach is usually a trial and error approach, which is extremely unadvantageous as antidepressants may have very adverse secondary effects [5]. Moreover, depression has consistently been linked to cognitive impairments [6], [7].

In the presence of HD, depression is even more difficult for a clinician to diagnose as apathy, lack of initiative and weight loss are also frequent signs of HD alone [8]. Many hypothesis have been formulated for the prevalence of this psychiatric disorder in HD but no conclusions have been found. There exists, this way, the necessity to understand if there are specific patterns in the disease that are linked to depression and to develop objective mechanisms for this purpose. Machine learning offers the ability to recognize these patterns in what is, for the human perspective, simply heterogeneous information and model it, creating high-level abstractions, and finally giving useful outputs [9].

1.2 Objectives

The main objective of this dissertation was to develop a model able to detect HD cases where there is a medical history of depression (with or without a formally stated diagnosis), from sequential clinical data. While most studies regarding this issue focus on statistically associating specific phases of the disease and/or specific symptoms and signs to depression, this work aims to be a "proof of concept" and a starting point in the use of Deep Learning for processing longitudinal clinical data (with no *a priori* patient stratification) to obtain conclusions about the presence of depression in Huntington's Disease. In a near future, similar methods may be applied to the diagnosis process.

More concretely, these are the main objectives of the present dissertation:

- To develop an RNN model that best suits the task of detecting depression based on longitudinal clinical data. For that, standard RNNs, LSTMs and GRUs models will be tested.
- To use information regarding the patient's family history, demographics, first noted symptoms and age at which they appeared, to provide additional information to the model.

- To compare the predictability of this condition in the HD and control participants separately, using the same method.

1.3 Thesis Outline

This dissertation is composed of 6 chapters. Chapters 2 and 3 are introductory theoretical chapters covering the state of the art of the topics upon which this dissertation relies. Chapter 2 is about Huntington's Disease, providing the necessary information for the comprehension of the developed work regarding its general description and relation with depression. Chapter 3 introduces the theoretical concepts behind sequence classification with Deep Learning, including introductory notions about Machine Learning and Recurrent Neural Networks. In Chapter 4 the followed methods are described and the results are presented and discussed in Chapter 5. The last chapter summarizes the outcomes of the study and main conclusions that can be drawn, describes its limitations and suggests future steps to take following this work.

2

Huntington's Disease

Contents

2.1 Huntington's Disease

Huntington's Disease is a rare neurodegenerative disorder characterized by involuntary choreatic movements, behavioural and psychiatric disturbances and dementia [10]. It is a terminal illness with an autosomal dominant inheritance (50% probability of being transmitted through generations), described by a progressive course of a combination of motor, cognitive and behavioural impairments, which make it devastating both to patients and their families [11].

HD is caused by an elongated CAG trinucleotide repeat in the gene coding for the protein Huntingtin [12]. Prevalence rates vary geografically: the overall worldwide prevalence is 2.7 per 100,000 but considering the two subgroups (1) North America, Europe and Australia and (2) Asia, in the first the estimated prevalence is 5.7 per 100,000 and in the second it is only 0.4 per 100,000 [13]. The reason for this great difference is thought to be in part due to differences in CAG tract size (in populations with lowered prevalence rates of HD, CAG sequence size in the wild-type HTT gene is shorter) [14].

Genetic predictive testing can inform whether, but not precisely when, the disorder will manifest itself [15]. The onset of symptoms is in most cases between the ages of 30-50 years old, but in some rare cases (around 5% of all HD cases) the symptoms start before the age of 20 and it is called Juvenile Huntington's Disease (JHD) [16].

There is no cure for this illness: management is multidisciplinary and based on treating symptoms with the aim of improving the patients' quality of life. Medication and non-medical care for depression and aggressive behavior and also to improve motor anomalies is used. The evolution of the disease leads to an increasing dependency in daily life, which results at some point in patients requiring full-time care, and finally death [10]. The most common cause of death is pneumonia [17].

2.1.1 Etiology of the disease: the Huntingtin (HTT) gene and protein

As previously mentioned, HD is an autosomal dominant inherited disease caused by an expanded CAG repeat (36 repeats or more) on the short arm of chromosome 4p16.3 in the Huntingtin gene (gene IT-15) [10]. CAG is a trinucleotide (composed of citosine, adenine and guanine) that codes for the amino acid glutamine. There is a negative relation between the number of CAG repeats and the age of onset [18] but it gives no indication about the initial symptoms, the course or the duration of illness [10].

The wild-type HTT gene contains a DNA segment of 6 to 26 CAG repeats. From 27 to 35 CAG repeats it is unstable and has the potential to expand into the abnormal range in future generations; from 36 to 39 it is abnormal but there may be reduced penetrance; over 40 repeats, there will unequivocally be clinical manifestations. In cases of JHD the number of repeats often exceeds 60 [19]. There's also evidence that about 1% of suspected Huntington's Disease cases emerge as phenocopy syndromes: patients presenting with the features of HD but lacking the genetic mutation [20].

Although the exact function of the Huntingtin protein remains uncompletely understood, it appears to have an important role in the neurons functioning and it is essential for the normal development before birth [21].

Huntingtin is found in many of the body's tissues, and expressed at its highest levels in the brain

and testes. Within the brain, it is found in all neurons, as well as glial cells. Evidences have shown that it interacts with a large number of effector proteins to mediate many physiological processes, such as in axonal trafficking (transport of cellular organelles and molecules through the neuron's cytoplasm - from the cell body to the axon), regulation of gene transcription, and cell survival (protecting it from self-destruction, called aptoptosis). HTT has also been suggested to have both pre- and post-synaptic roles [22]. Figure 2.1 illustrates the role of this protein within a neuron.

The presence of the expanded CAG segment leads to the production of an abnormally long version of the huntingtin protein [11], resulting in a cascade of cell death and cerebral degeneration. Although other parts of the brain are also affected, the basal ganglia appears to be the most heavily damaged [8]. Figure 2.2 shows the damage caused by HD in the brain volume. The elongated protein is cut into smaller, toxic fragments that bind together and accumulate in neurons and this process particularly affects regions of the brain that help coordinate movement and control thinking and emotions (the striatum and cerebral cortex) [22]. Summarily, HD is caused both by a toxic gain-of-function due to the expanded protein and the loss of normal HTT function [11].

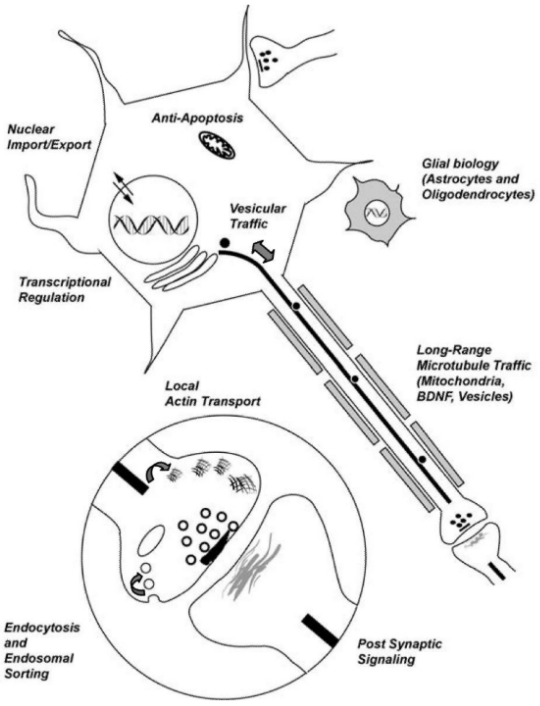


Figure 2.1: Schematic illustration of the functioning of the HTT protein [22].

2.1.2 Symptoms, clinical onset and evolution of the disease

A person with one parent with HD is at risk of also being a gene mutation carrier: if a genetic test confirms this, the person is at the pre-clinical or pre-manifest stage of the disease until the symptoms appear and a clinical diagnosis is done. Note that not only those with a parent suffering from HD could develop the disease, as a *de novo* mutation could also lead to the expanded form of the HTT gene.

The progression of symptoms in HD is not well understood but the overall clinical course of the

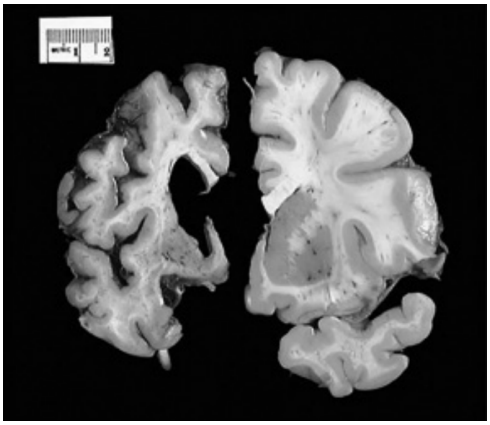


Figure 2.2: Damage in brain volume caused by HD (left); comparison with a healthy brain (right) [8].

disease is characterized by a decrease in independence and an increase of severity in motor, cognitive and behavioural impairments, as illustrated in Figure 2.3.

In the past, the diagnosis was only suggested after the first motor signs had started. However, it has become clear that psychiatric and cognitive changes can be the first signs, even many years before motor impairments become visible [15], [23]. Many patients mention a gradual change in behaviour and performance at work (for example, staying home due to experiencing the symptoms of a burn-out or a depression). Although these signs are non-specific and may have a different plausible explanation, it has become clear that these signs can be the first manifestation of HD [10].

As already mentioned, the age at onset is usually between 30 and 50 years, but it can happen any time between the ages of 2 to 85. The mean duration of the disease is around 17-20 years [10].

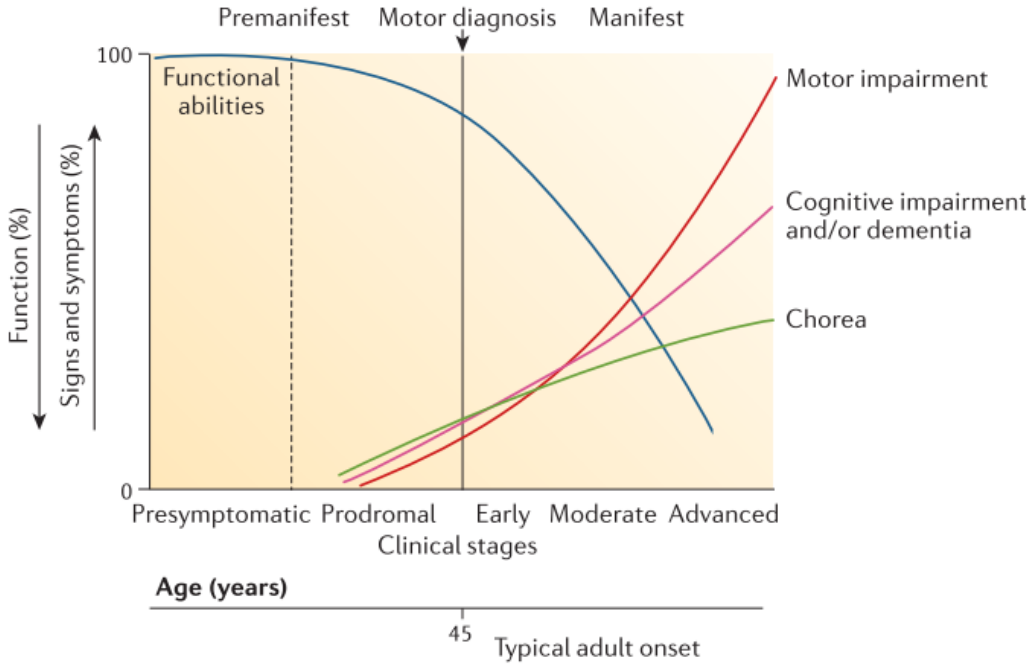


Figure 2.3: Simplified illustration of the evolution of the different symptomatic domains of HD through time [11].

Although the nuclear features of HD consist of motor, cognitive and psychiatric disturbances, there

are other less known, but prevalent and debilitating symptoms which include unintended weight loss, sleep disturbance and autonomic nervous system impairment.

2.1.2.A Motor symptoms and signs

One of the main signs of the motor impairment in patients with HD is the emerging of involuntary movements. Initially, the movements occur in the distal extremities (like fingers and toes) and in small facial muscles; walking becomes unstable and the person may look as if inebriated. No pattern exists and facial choreatic movements lead to a continuous movement of facial muscles [10].

Alterations in oculomotor performance are also among the first observable physical alterations during initial stages of HD. In fact, quantifiable measurements of oculomotor performance have been used to distinguish controls, pre-symptomatic and manifest patients [24]. Dysarthria (speech impairments resulting from neurologic disease or damage [25]) and dysphagia (difficulty in swallowing) become prominent during the course of the disease. Talking and swallowing gradually become difficult tasks (frequently leading to choking) and often in later stages of the disease the patient becomes mute [10].

All patients develop hypokinesia, a primary motor control impairment characterized by slow movement (bradykinesia) or no movement due to difficulty in starting it (akinesia) [26]. Also, falling becomes frequent. The involuntary choreiform movements tend to decrease in later stages of the disease, while rigidity and bradykinesia increase [27].

2.1.2.B Cognitive symptoms and signs - dementia

As previously mentioned, cognitive impairments very often emerge before the motor ones. It has been shown that neurocognitive tests are robust clinical indicators of the disease process prior to reaching criteria for motor diagnosis of HD [28]. Like motor disability, cognitive decline progresses gradually. The features of cognitive damage in Huntington's Disease are similar to disorders associated with striatal–subcortical brain pathology (such as Parkinson's disease) but are dissimilar to Alzheimer disease - for example, rapid forgetting is not a pronounced characteristic of HD [29].

Common cognitive decline features in HD include deficits in attention, emotion recognition, visuomotor processing, decreased ability to learn and retrieve new information [23], [28], [30]. These cognitive impairments with simultaneous psychiatric problems result many times in lack of initiative, social disengagement, impulsivity and lack of awareness [31].

2.1.2.C Behavioural and psychiatric symptoms and signs

Along with motor and cognitive changes, psychiatric problems complete the triad of signs and symptoms that characterize HD. Due to the burden they cause on the daily life, the psychiatric issues have a highly negative impact on functioning and on the family of the patients [32]. Psychiatric symptoms, particularly depression, can develop during the prodromal stage or when the disease is manifest. It has been suggested as a significant prognostic component in HD patients (these symptoms frequently arise before motor onset, like cognitive impairments) and may present several years prior to disease onset [33], [34].

The most frequently occurring psychiatric sign is depression [1]. The diagnosis is often difficult because apathy, lack of initiative and weight loss also occur in HD alone (a correlation between weight loss and the length of the cytosine, adenine and guanine (CAG) repeat has been described [35]). Usually there is low self-esteem, feelings of guilt and anxiety [36]. Apathy is correlated to disease progress (cognitive deterioration and functional decline), whereas anxiety and depression are not [2].

Since the original description of HD in 1872, a tendency to suicide has been reported as a peculiarity of the disease [37]. Suicidal ideation is most common when patients start experiencing symptoms (before formal clinical diagnosis) and in the stage of HD following diagnosis when patients become less independent [38].

Obsessions and compulsions are a very common sign and also lead to irritability and aggressive behaviour [10]. Psychosis (presence of delusions and/or hallucinations) is also a psychiatric manifestation for some patients, mainly in the later stages of the disease [39]. The prevalence of psychotic symptoms in HD patients varies between 3 and 11% [40].

Overall, due to this combination of motor, cognitive and psychiatric deficits, for patients with HD, the activities of daily living (ADL) such as getting out of bed, taking a shower, getting dressed, cooking or eating become increasingly difficult.

2.1.3 Standard clinical assessment procedures and severity measures

The Unified Huntington's Disease Rating Scale (UHDRS) is a standardized clinical rating scale to assess four domains of clinical performance and capacity in HD: motor function, cognitive function, behavioral abnormalities, and functional capacity [41].

The motor scale assesses eye movements, motor control, rigidity, bradykinesia, dystonia, chorea, and gait. The cognitive section is composed of a test of verbal fluency, the Symbol Digit Modalities Test and the Stroop Test. The behavior section assesses the frequency and severity of psychiatric symptoms (like depression, delusions). Frequency and severity of these symptoms are scored on a scale from 0 to 4 with lower numbers indicating less frequent and less severe psychiatric symptoms. A brief health history is obtained which asks whether treatment has been sought for depression and whether any suicide attempts have been made [38], [42].

The Total Functional Capacity (TFC) scale is a standard measure of functional capacity employed in HD research. The TFC scale consists of five items assessing occupation, capacity to handle financial affairs, to manage domestic responsibilities, to perform activities of daily living, etc. Scores range from 0 to 13, with higher scores indicative of higher functioning and greater independence. Some studies use TFC as the basis in the determination of the stage of illness of the patient [38].

2.2 HD and Depression

2.2.1 What is depression?

Depression is one of the most common mental disorders in the general population and it is the leading cause of disability worldwide. It is a major contributor to the overall global burden of disease

with more than 264 million people affected [43]. Its presence is linked to diminished physical health and quality of life [44] and higher risk of suicide [45].

Also called Major Depressive Disorder (MDD), depression is a prevalent heterogeneous illness characterized by depressed mood, anhedonia (lack of interest or pleasure) and altered cognitive function [4]. Risk factors for the development of depression include biological (like genetics, neurological alterations - both functional and structural alterations), cognitive (such as beliefs, information processing, personality) and social factors (life experiences, stress) [46].

2.2.1.A Clinical assessment and treatment

The diagnosis of MDD is largely based on application of criteria from the Diagnostic and Statistical Manual of Mental Disorders (DSM) and clinician judgment; upon diagnosis most patients are started on first-line antidepressant agents (such as selective serotonin reuptake inhibitors (SSRIs) and tricyclic antidepressants (TCAs)) which is largely a trial and error process [47]. Although there is a high rate of efficacy in treating MDD with antidepressants, the neurobiological mechanisms of their efficacy are not well understood and one problem with this approach relies on the secondary effects that are adjacent to these drugs, such as gastrointestinal disturbances, hepatotoxicity and hypersensitivity reactions, metabolic and sexual dysfunctions [5].

Health-care providers may also offer non-pharmacological treatments, such as cognitive behavioural therapy, naturopathic interventions, psychotherapy and exercise-based interventions, which do not have secondary adverse effects but are usually not as efficacious as the use of antidepressants, so in most cases both kinds of treatment should be considered [48].

2.2.1.B Cognitive changes in depression

As previously mentioned, it is known that depression is associated with changes in cognitive abilities. Depression is characterized by general cognitive deficits - for example, impairments in executive functioning, attention and memory - and negative cognitive biases, such as in the processing of emotional information. Also, patients with this disorder tend to increase the use of maladaptive emotion regulation strategies, like rumination, and not to use adaptive ones, like reappraisal [49].

More concretely, considering the general cognitive impairments, one study showed that individuals with MDD scored significantly lower than an age and IQ-matched group of control subjects in the areas of verbal fluency, visual memory tasks, spatial span tasks, working memory tasks, and tasks involving executive function [6].

2.2.1.C Biological etiology: Where is depression?

Despite decades of research, to date, the biological bases for the presence and heterogeneity of depression remain poorly understood [4]. Many different regions of the brain have been linked to depression. Decreased metabolism in the prefrontal cortex (especially dorsolateral and dorsoventral brain regions), structural and functional impairments in limbic brain regions (amygdala, hippocampus and

dorso-medial thalamus) and an abnormal metabolism in the brain stem and basal ganglia are only some examples of brain activity abnormalities found in different studies [50], [51], [52].

As in other disorders of higher mental functions, it is difficult to localize the anomalies. The abnormality in depression may also lie at a molecular level affecting neurotransmission and consequently neurons in several brain regions simultaneously. Another possibility is that the abnormality lies in a brain system that influences the functioning of multiple brain regions or even different symptoms of depression may involve different brain regions in varying degrees [50]. Furthermore, recent studies have shown evidences of a link between depression and functional connectivity abnormalities [53], [54].

2.2.1.D Depression in the presence of a medical diagnosis

Depression is a frequent feature of many terminal illnesses. One study showed that patients with any medical diagnosis were more than twice as likely to have depression than patients without a medical diagnosis [55]. In neurological diseases, the rates of depression are recognized as higher than in the general population, however, there is much debate about the etiology of this depression: if on one hand, the impact of a medical diagnosis in the patient's life may increase the risk for depression, on the other, there may be physiological impairments related to the disease that would be on its origin [38].

2.2.2 Depression in Huntington's Disease

As previously mentioned in section 2.1.2.C, depression is the most common psychiatric comorbidity in HD. Moreover, depressive mood has been considered one of the most impacting factors in the quality of life of patients with this illness [56].

Estimates of the prevalence of depressed mood in Huntington's Disease vary widely, ranging from 33% [57] to 69% [36]. One reason for this discrepancy is the fact that the rate of depression in HD has been measured using different assessments, which makes the interpretation across studies more difficult.

Many of the symptoms of HD resemble and may potentially disguise the symptoms of depression (for example, changes in appetite, fatigue, changes in sleep). It can be difficult to tell whether a person's symptoms are depression, HD or a combination of both [8].

Depressive symptoms in HD do not correlate well with motor or cognitive measures of the disease progression, which suggests that different neuropathological processes may be involved [38]. Nonetheless, it has been observed to have higher prevalence in specific periods of the illness and to be less frequent in late stages, possibly due to greater cognitive impairments and, consequently, decreasing illness awareness [40].

As explained in section 2.2.1 depression (not in the presence of HD) has been associated with some changes in cognitive abilities. In light of the observed cognitive decline in HD, the high incidence of depression in HD, and the association of cognitive decline with depression, questions emerge about a possible impact of depression on cognition in individuals who are HD-presymptomatic, which would lead to the exacerbation of cognitive impairments during the illness development [58].

The etiology of HD depression is unclear and may be due to a number of factors: the development of depressive symptoms in Huntington's disease could be a direct result of cerebral degeneration, for which several neuropathological mechanisms have been proposed [1], [59], it could be related to the disease associated alterations in the neurotransmitters in the brain that regulate mood [8] or it could be a psychological reaction to being at risk for Huntington's disease, having grown up in an insecure and harmful environment, and/or the awareness of disease onset [40].

3

Deep Learning for sequential data classification

Contents

3.1 Theoretical Concepts behind sequence classification

3.1.1 Machine Learning (ML)

Machine Learning is a branch of the wide field of Artificial Intelligence (AI). François Chollet defines AI as "the effort to automate intellectual tasks normally performed by humans" [60]. ML is an approach of AI to solve complex problems which could not be determined by explicit rules, making the use of computers to statistically estimate complicated functions [61].

As the name suggests, a Machine Learning algorithm should be able to learn. In *Machine Learning*, Tom M. Mitchell defines learning as follows: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E" [62].

ML algorithms can be broadly divided in two categories: supervised and unsupervised. A supervised learning algorithm uses a dataset containing features and each example is associated with a label; unsupervised learning algorithms use datasets containing features and from them learn characteristics of the structure of the dataset, without using explicitly-provided labels [61].

A wide variety of tasks can be solved with Machine Learning. Classification is one of them and in this type of task, the computer program is asked to specify which of k categories some input belongs to.

3.1.1.A Training a Machine Learning model

The training phase is when the learning takes place. In order to train a machine learning model, it's necessary to have access to a training set. The ML system is presented with many examples relevant to a task and finds the statistical structure that determines the rules behind this task [60].

For this, it is necessary to have for each training sample an associated expected output (in the case of supervised learning) and some error measure (loss function) to compute the training error (which compares the algorithm's output to the expected one) - the idea is to reduce this error. Neverthless, the aim is to make the algorithms perform well on previously unseen inputs - in other words, the aim is to achieve good generalization. This way, after the training phase, the performance of the algorithm is measured on a test set of new examples. If the training error is small but the test error is large it is called overfitting [61].

Most machine learning algorithms have several settings that control the behavior of the learning model. These settings are called hyperparameters.

While the loss function is the error measure that will be minimized during training, the optimizer is the method that determines how the model will be updated, based on that measure.

3.1.2 Artificial Neural Networks

An Artificial Neural Network (ANN) is a software reproduction of the neuronal structure of the human brain [63]. Biological neurons spread electrochemical signals along neural pathways through synapses - some of the transmitted signals tend to excite the reached neurons, others to inhibit them. For each neuron, if the cumulative effect exceeds a defined threshold, the neuron fires and sends a signal to other

neurons. An artificial neuron mimics this biological functioning: it receives a set of inputs and each input is multiplied by a weight, equivalent to the synaptic strenght [64], and then maps it to an output, through an activation function. An ANN is represented by a set of nodes (which represent the neurons) and connections with coefficients (weights) [65]. Figure 3.1 illustrates the basic structure of an ANN.

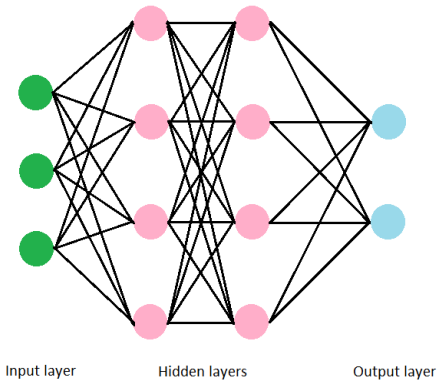


Figure 3.1: Basic structure of an ANN.

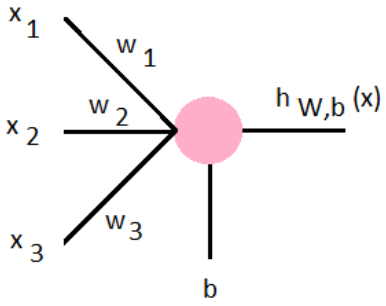


Figure 3.2: Schematic representation of an artificial node.

Figure 3.2 shows the structure of a single artificial neuron that receives input from sources that are either other neurons or data input. Each input of the node is multiplied by the correspondent weight and the sum passes through the activation function to compute the output of the node (as in expression 3.1), which will be used as an input for the next layer's nodes. The output of the network is the output of the final layer. The number of successive layers is what conceives "depth" to the network and it is the reason why this field of Machine Learning is called Deep Learning.

$$h_{W,b}(X) = f(w_1 \cdot x_1 + w_2 \cdot x_2 + w_3 \cdot x_3 + b) = f(W \cdot X + b) \tag{3.1}$$

where f is the chosen activation function of the node and b is the bias. The activation function serves to introduce non-linearity in the modeling capabilities of the network and it is usually a sigmoid, hyperbolic tangent (tanh), rectified linear (ReLU) or max-pooling function [60]. The sigmoid function converts the node's inputs to simple probabilities between 0 and 1 and most of its output will be very close to the extremes of 0 or 1. The tanh function is also a sigmoid-like curve but unlike the sigmoid function, the normalized range of tanh is between -1 and 1, which has the advantage of including negative values. In figure 3.3, the linear, sigmoidal and hyperbolic tangent functions are represented.

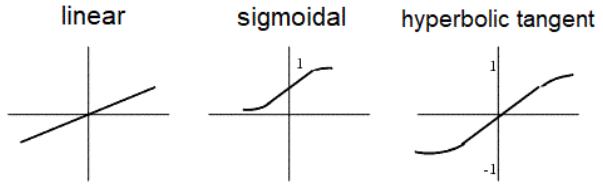


Figure 3.3: Graphical representation of the linear, logistic sigmoidal and hyperbolic tangent activation functions.

3.1.2.A Gradient descent

Gradient descent is an optimization technique used to minimize the training error, by iteratively changing the parameters of the model in steps with opposite sign of the partial derivatives of the loss function, in order to find the values for these parameters that make the loss reach its minimum. It is often necessary to optimize complex functions that may have many local minima or that have multidimensional inputs - this makes optimization difficult and, therefore, it is usual to define a stopping criteria different from finding the global minimum.

A problem in machine learning is that large training sets are necessary for good generalization, but large training sets are also more computationally expensive. Hence, instead of computing the gradient of the loss function for every training example on each iteration of the algorithm (Stochastic Gradient Descent (SGD)), it is possible to use a random subset of examples from the training set (called minibatch) or even a single sample. The first method is called mini-batch gradient descent and the second is called stochastic gradient descent and they are extensions of the gradient descent algorithm [66].

When the algorithm has seen the entire training set it is called an epoch. Having a dataset with N samples, with the gradient descent method the parameters are only updated after each epoch while using the stochastic gradient descent, in one epoch, the parameters are updated N times; choosing a batch-size n, the parameters are updated N/n times per epoch.

The gradient descent method uses the gradient of the loss function (with respect to the weights) to make a step change in w to lead it towards the minimum of the error curve. This is an iterative method: each time, every weight and every bias of the network are updated according to the expressions 3.2 and 3.3 [66].

$$w_{ij}^{(l)} = w_{ij}^{(l)} - \alpha \frac{\partial L(w,b)}{\partial w_{ij}^{(l)}} \tag{3.2}$$

$$b_{ij}^{(l)} = b_{ij}^{(l)} - \alpha \frac{\partial L(w,b)}{\partial b_{ij}^{(l)}} \tag{3.3}$$

Where α is the learning rate and L is the loss function. The learning rate, also called step, is a hyperparameter which controls how much to change the value of the weight in the opposite direction from the gradient. It is difficult to choose an adequate value for it, as if it is too small, it will take many iterations to reach the minimum loss and it might stop at a local minimum and if it is too large, the optimization will diverge [60].

There are many algorithms developed to address this problem. Adding a momentum term [67] helps to accelerate SGD in the relevant direction by adding a fraction of the update vector of the past time-step to the current update vector, as in expression 3.4. This allows having a faster convergence with reduced oscillation.

$$v_t = \gamma v_{t-1} + \alpha \nabla_\theta J(\theta) \tag{3.4}$$

$$\theta = \theta - v_t \tag{3.5}$$

Adaptive Moment Estimation (Adam) [68] is a method that computes adaptive learning rates for each parameter. It stores an exponentially decaying average of past squared gradients and also keeps an exponentially decaying average of past gradients m_t, similar to momentum.

When working with neural networks, the most commonly used method to compute these gradients is the backpropagation method [69], an optimizer which uses the chain rule of differentiation for applying the gradient descent to the equations that define the network.

3.1.2.B Backpropagation

Considering a neural network that only has forward connections between neurons, i.e, a feedforward neural network (FFNN) [70] (notice that when the networks are extended to include feedback connections, they are called recurrent neural networks - these will be discussed in section 3.2), the expression 3.6 represents the total input of a node from a hidden layer.

$$a_i^{(l)} = b_i^{(l)} + \sum_{j=1}^{r_{l-1}} w_{ji}^l o_j^{l-1} \tag{3.6}$$

Where $a_j^{(l)}$ is the weighted sum of the inputs and bias for node j in layer l, r_l is the number of nodes in layer l and o_j^{l-1} is the output of the node j in layer $l-1$.

In order to compute the partial derivatives from expressions 3.2 and 3.3, the chain rule of differentiation is used as in expression 3.7 [69].

$$\frac{\partial L}{\partial w_{ij}^{(l)}} = \frac{\partial L}{\partial a_j^{(l)}} \times \frac{\partial a_j^{(l)}}{\partial w_{ij}^{(l)}} = \delta_j^l \times \frac{\partial a_j^{(l)}}{\partial w_{ij}^{(l)}} \tag{3.7}$$

Where δ_j^k is the partial derivative of the loss function with respect to the node's input, called the error term. From **??**,

$$\frac{\partial a_j^{(l)}}{\partial w_{ij}^{(l)}} = \frac{\partial}{\partial w_{ij}^{(l)}}(b_j^{(l)} \sum_{k=0}^{r_{l-1}} w_{kj}^l o_k^{l-1}) = o_i^{l-1} \tag{3.8}$$

Assuming the network as m layers, the last layer only has 1 node (i.e, there is only 1 output node) and the loss function is computed as in expression 3.9, the error term is computed as shown in expressions 3.11 and 3.12.

$$L = \frac{1}{2}(\hat{y} - y)^2 \tag{3.9}$$

For the last layer ($l = m$)

$$\hat{y} = f(a_1^{(m)}) \tag{3.10}$$

$$\delta_1^m = \frac{\partial L}{\partial \hat{y}} \times \frac{\partial \hat{y}}{\partial a_1^{(m)}} = (\hat{y} - y) \cdot f'(a_1^{(m)}) \tag{3.11}$$

And for the rest of the layers $1 \leq l < m$:

$$\delta_j^l = \sum \frac{\partial L}{\partial a_k^{(l+1)}} \times \frac{\partial a_k^{(l+1)}}{\partial a_j^{(l)}} = \sum_{k=1}^{r_{l-1}} \delta_k^{l+1} \times \frac{\partial a_k^{(l+1)}}{\partial a_j^{(l)}} = f'(a_j^{(l)}) \sum_{k=1}^{r_{l-1}} \delta_k^{l+1} w_{jk}^{l+1} \tag{3.12}$$

3.2 Recurrent Neural Networks

Recurrent Neural Networks or RNNs are a family of neural networks for processing sequential data [61]. RNNs are neural nets which include feedback connections among hidden units, associated with a time delay [71]. The key point is that the recurrent connections allow a “memory” of previous inputs to persist in the network’s internal state. RNNs are also trained with backpropagation and the forward pass of an RNN is identical to that of a feedforward network, except that the hidden layers receive as inputs both the current external input and the output from the previous timestep [72], as represented in figure 3.4.

Considering an RNN with I input nodes, a single hidden layer (with H hidden nodes) and an output layer (with K output nodes), which receives as input a sequence x, the forward pass is as represented in expressions 3.13 and 3.14 [72].

$$a_h^t = \sum_{i=1}^{I} w_{ih} x_i^t + \sum_{h'=1}^{H} w_{h'h} o_{h'}^{t-1} \tag{3.13}$$

$$o_h^t = f(a_h^t) \tag{3.14}$$

Similarly to what is done in the learning phase when working with FFNN, it is necessary to compute the partial derivatives of the loss function with respect to the weights, as in 3.7. The most used algorithm for efficiently calculate these derivatives is backpropagation through time (BPTT) [73], which equations are represented in expressions 3.15 and 3.16 [72]. It is important to notice that the same weights are used every timestep.

$$\frac{\partial L}{\partial w_{ij}} = \sum_{t=1}^{T} \delta_j^t \times \frac{\partial a_j^t}{\partial w_{ij}} = \sum_{t=1}^{T} \delta_j^t o_i^t \tag{3.15}$$

$$\text{with } \delta_h^t = f'(a_h^t)\left(\sum_{k=1}^{K} \delta_k^t w_{hk} + \sum_{h'=1}^{H} \delta_{h'}^{t+1} w_{hh'}\right) \tag{3.16}$$

The problem with RNNs is that the inputs cycle around the network’s recurrent connections, which leads us to the problem of learning long-term dependencies [74].

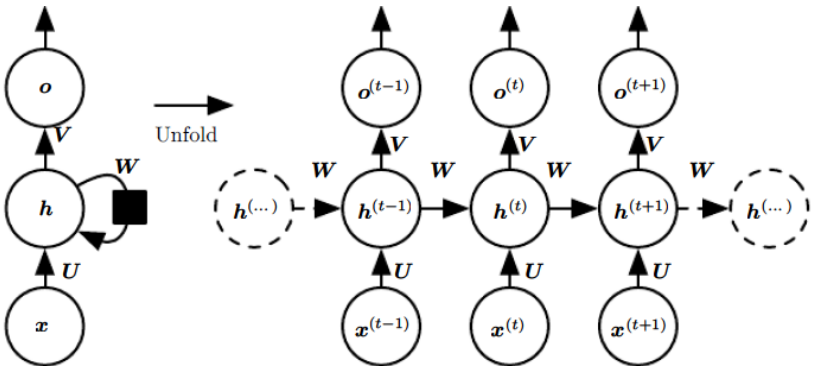


Figure 3.4: Schematic representation of an RNN [61].

3.2.1 The problem of long-term dependencies

The mathematical challenge of learning long-term dependencies is that gradients propagated over many timesteps tend to either vanish or explode (being the first phenomenon much more probable than the second).

Recurrent networks involve the chain multiplication of the derivative of the activation function, once per time step. It is possible to think of the recurrence relation $h^{(t)} = W * h^{(t-1)}$ as a very simple recurrent neural network without a nonlinear activation function and inputs. This recurrence relation essentially describes the power method and it may be simplified to $h^{(t)} = W^t h^{(0)}$ and if W admits an eigendecomposition of the form $W = Q \bigwedge Q^T$ with orthogonal Q, the recurrence may be simplified to $h^{(t)} = Q \bigwedge^t Q^T h^{(0)}$. The eigenvalues are raised to the power of t causing eigenvalues with magnitude less than one to decay to zero and eigenvalues with magnitude greater than one to explode [61].

3.2.2 Long Short-Term Memory (LSTM) Networks

The Long Short-Term Memory (LSTM) method was first introduced by Hochreiter and Schmidhuber in 1997 [75], with the aim of addressing the problem of long-term dependencies, described in section 3.2.1. It is a variant of an RNN with a gated structure, which enables it to handle long input sequences.

The central idea behind the LSTM architecture is a memory cell which can maintain its state over time, and non-linear gating units which regulate the information flow into and out of the cell [76]: the input, output and forget gates.

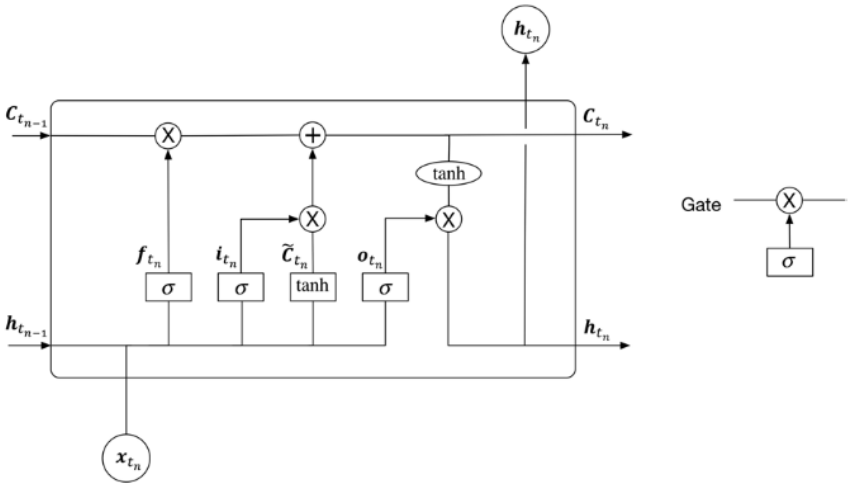


Figure 3.5: Schematic representation of an LSTM cell [77].

Figure 3.5 illustrates an LSTM cell and the expressions from 3.17 to 3.22 are the equations for the forward pass. Having as input a sequence x, x_t represents the input at the current timestep t and h_{t-1} and C_{t-1} represent, respectively, the output and the cell state from the previous timestep. W, b are the weights and biases and the indexes f, i and o correspond to the forget, input and output gates, respectively. These multiplicative gates are sigmoid layers and each has a different task. The forget gate (3.17) takes as input x_t and h_{t-1} and it is what enables the cell state to be reset, as its output will

multiply the previous cell state C_{t-1} (see 3.20). The input gate (3.18) "decides" which values will be updated and from 3.19 the new values to add to the cell state are computed. Finally, in the output gate (3.21) it is decided what part of the current cell state is going to be output (3.22).

$$f_t = \sigma(\, W_f \cdot [\, h_{t-1},\, x_t] + \, b_f) \quad (forget\ gate) \tag{3.17}$$

$$i_t = \sigma(\, W_i \cdot [\, h_{t-1},\, x_t] + \, b_i) \quad (input\ gate) \tag{3.18}$$

$$\tilde{C}_t = tanh(\, W_C \cdot [\, h_{t-1},\, x_t] + \, b_C) \quad (candidate\ values) \tag{3.19}$$

$$C_t = \, f_t * \, C_{t-1} + \, i_t * \, \tilde{C}_t \quad (cell\ state) \tag{3.20}$$

$$o_t = \sigma(\, W_o \cdot [\, h_{t-1},\, x_t] + \, b_o) \quad (output\ gate) \tag{3.21}$$

$$h_t = \, o_t * tanh(\, C_t) \quad (output) \tag{3.22}$$

The gates allow LSTM cells to store and access information over long periods of time, mitigating the vanishing gradient problem. For example, as long as the input gate remains closed (i.e. has an activation near 0), the activation of the cell will not be overwritten by the new inputs arriving in the network, and can therefore be made available to the net much later in the sequence, by opening the output gate [72]. The preservation over time of gradient information by LSTM is illustrated in figure 3.7.

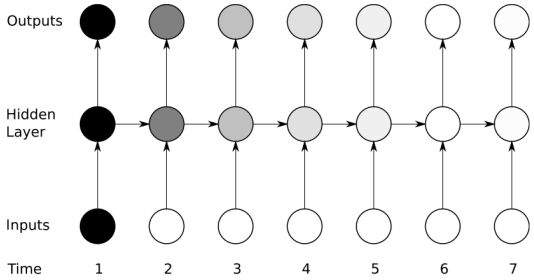


Figure 3.6: Schematic representation of the vanishing gradient problem [72].

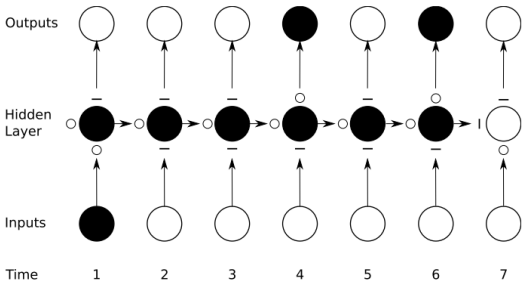


Figure 3.7: Schematic representation of the preservation of the information over time with LSTM [72].

3.2.3 Gated Recurrent Units

Another type of RNN with a special gated architecture is the GRU, which was developed in 2014 [78]. The main differences from the LSTMs are that it uses two gates (the update and reset gates) instead of three and that it doesn't use the cell state to transfer information, but rather the hidden state [79]. It can be thought of as a modification of the LSTM with a less complex architecture and, consequently, more computationally efficient [60].

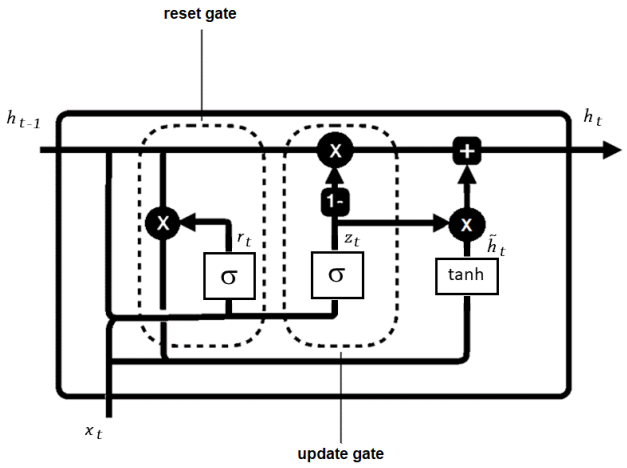


Figure 3.8: Schematic representation of a GRU (adapted from [80]).

Figure 3.8 illustrates the structure of the GRU and its components. The update gate, represented by z_t, "couples" the forget and input gates from the LSTM architecture into one, which simultaneously controls how much of the previous memory content (h_{t-1}) to forget and how much of the new content (x_t) is to be added, through the computation of expression 3.23. The reset gate, r_t, allows the unit to forget the previous hidden states and it is computed as in expression 3.24. The candidate hidden state ($\tilde{h}_t$) is done similarly to that of the the LSTM (expression 3.25) [81]. Finally, at timestep t the state of the GRU is the linear interpolation between the previous activation (h_{t-1}) and the candidate hidden state ($\tilde{h}_t$) [79].

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \quad (update\ gate) \tag{3.23}$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \quad (reset\ gate) \tag{3.24}$$

$$\tilde{h}_t = tanh(W \cdot [h_{t-1} * r_t, x_t] + b_h) \quad (candidate\ activation) \tag{3.25}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \quad (output) \tag{3.26}$$

Despite its simpler architecture and the fact that it may not have as much representational power as LSTM, it has been evidenced that it can outperform LSTM networks [79], particularly on smaller datasets.

3.2.4 State of the Art applications

Recurrent Neural Networks, and in particular LSTM, achieve state-of-the-art results for several real-world problems which require modeling sequential data, such as those covering natural language processing (like speech or handwriting recognition or generation [82], [83]), genomic analysis [84] and music generation [85].

Although neural networks have long been applied to medical data, the ability of LSTM to recognize patterns in multivariate time series of clinical measurements was first empirically evaluated in 2016, when these networks were successfully used for multilabel classification of diagnoses, using sensor

data and lab test results from patients' Electronic Health Record [86]. It is important to mention that one great advantage of Deep Learning (not only RNNs, but of all types of neural networks) is that these algorithms can identify the input features that are most important in mapping to the desired output. This is very beneficial in the medical context, as no *a priori* information is required concerning which features are relevant towards making a diagnostic or treatment classification decision.

Also regarding biomedical applications, GRU models have been used for predicting Alzheimer's disease progression [87] and LSTMs have been used in the same task but in the context of Amyotrophic Lateral Sclerosis, using longitudinal data [88]. Being both neurodegenerative diseases (and ALS an also very heterogeneous disease), from the achieved results in these studies it is plausible to think that similar methods could be applied to HD.

4

Methods

Contents

4.1 Technological Materials

Data analysis was executed using Python language [89] with the code interpreter Jupyter Notebook 5.7.8 [90]. For data preprocessing and analysis, the python packages used were Pandas, Numpy and Scikit-learn. The networks were developed using keras with Tensorflow backend. Matplotlib [91] and Seaborn [92] were used for data visualization purposes as these libraries provide interfaces for drawing informative statistical graphics.

4.2 Data

The data used to develop this study was the Enroll-HD Periodic Dataset (Version 2018-10-R1) provided to the research community. The dataset includes data from the studies ENROLL and REGISTRY and also Adhoc data. It represents a data extract of the database of the Enroll-HD multistudy system from October of 2018.

4.2.1 Studies brief description

Enroll-HD is a global, longitudinal observational study of Huntington's Disease that started in 2011 and includes participants from North America, Europe, Australasia and Latin America [93].

REGISTRY is a longitudinal observational study of HD whose participants are from Europe [94]. The study started in 2004 and as Enroll-HD was created there was a transition of the REGISTRY participants into Enroll-HD. Therefore, this dataset includes a subset of participants who initially joined REGISTRY and have consented to Enroll-HD and to have their REGISTRY data integrated in the Enroll-HD study. Study procedures include annual assessments conducted during study visits and performed by trained clinical personnel [95].

Ad hoc is a subset of assessment data that was gathered at routine clinical visits pre-dating the participant's enrolment into REGISTRY [93].

4.2.2 Participants

Subjects of this study include (1) 11582 individuals who are carriers of the HD gene expansion mutation, independently of phenotypical manifestation (*i.e.* pre-manifest or manifest) or of the stage of the disease and (2) 3719 controls who do not carry the HD expansion mutation and who comprise the comparator study population. The second group includes genotype negative subjects who are family members of carriers of the Huntington's Disease mutation.

From the total 15301 participants of Enroll-HD, 3798 were previously part of the REGISTRY study. In the dataset, there is also Ad hoc information about 258 participants.

4.2.3 Assessments

The provided dataset comprises several data files and those that were used in the present study were "profile", "adhoc", "registry" and "enroll". The first contains general and updated information about

each participant and the others contain the assessments from the respective studies' visits.

These files contain data items defined by variables, which were used as features of the developed classifier. Two examples of these items are: age at the moment of the visit (from the visits data, as it varies from visit to visit) and ethnicity of the participant (found in "profile", as it is a "static" feature). Also, these items may represent both numerical and categorical variables. It is important to mention that all categorical features where given in a numeric form (for example, in this case, Caucasian corresponds to 1, American Black to 2, Hispanic to 3, etc) instead of the in the written form.

4.2.3.A Visits assessments - sequential data

The sequential data used in this study comprises all the three visits files ("adhoc", "registry" and "enroll"). Some items were excluded:

1) Those in which there was no variability between participants - all given answers were the same. From this criterion, 12 features were deleted - for example, the frequency of abuse of the drug ritalin.

2) Those which concerned assessments that were not performed on the Enroll-HD visits (i.e, the variables assessed in the REGISTRY visits but not in the Enrol-HD visits).

The visits files' items come from the forms that are filled during those visits. Some are filled by the participant, others require performing clinical examinations. A brief description of each is given below and table 4.1 contains the number of items that compose them.

- ***Medical History***: comprises questions regarding the history of drug (for non-medical reasons), tobacco and alcohol abuse. It includes a long list of different drugs (for example, heroin, cocaine, amphetamines, opium, tranquilizers) and keeps information about frequency of abuse.
- ***Variable Items***: general items, like age, anthropometric measurements, habits of drug use for medical and non-medical reasons, marital status, education and employment.
- ***UHDRS Motor Diagnostic Confidence***: motor section of the UHDRS - assesses motor features of HD with standardized ratings of oculomotor function, dysarthria, chorea, dystonia, gait, and postural stability.
- ***UHDRS Total Functional Capacity (TFC) and Functional Assessment Independence Scale***: functional section of the UHDRS, used to assess participants' functional status. The Total Functional Capacity scale includes 5 items in the domains of occupation, finances, domestic chores, activities of daily living and level of care required by the participant. The Functional Assessment Independence Scale comprises an extensive list of "yes or no" questions about daily life activities, such as "could the subject walk/drive/handle finances without help?".

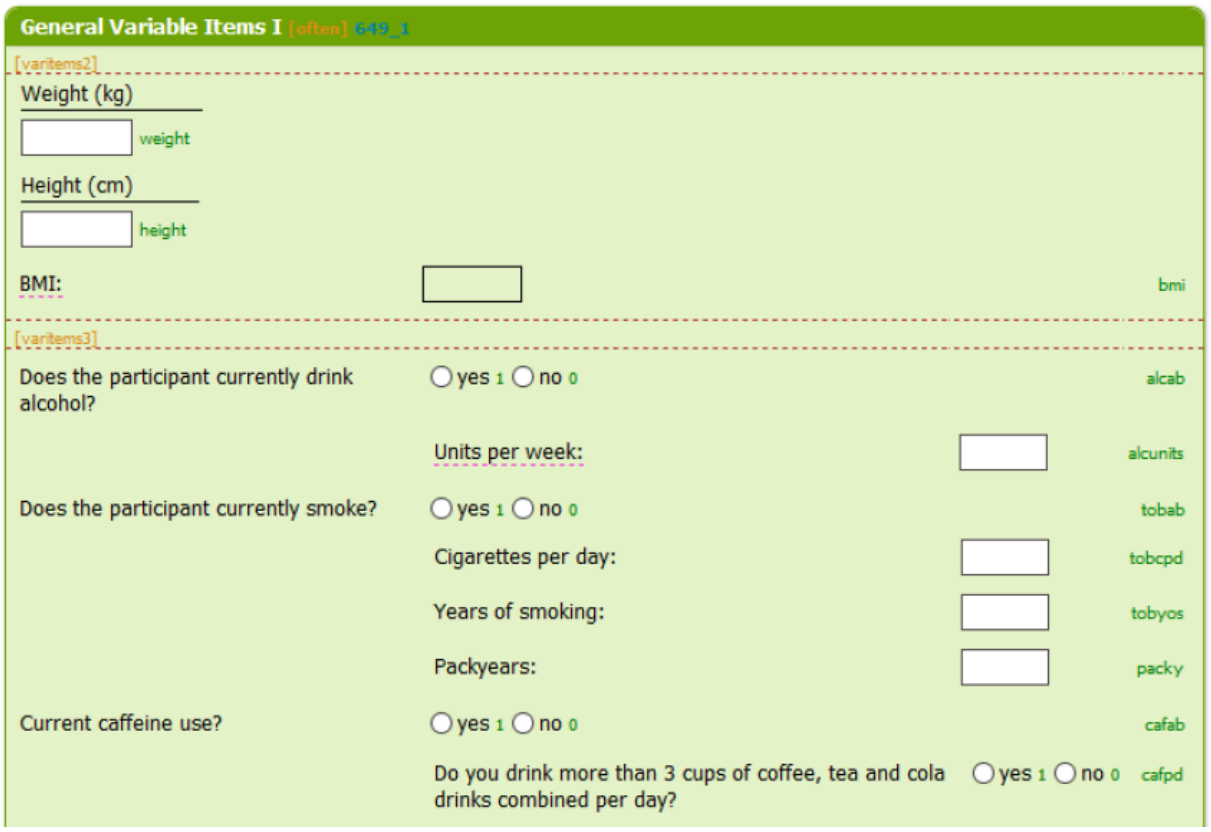


Figure 4.1: Part of the "Variable Items" form.

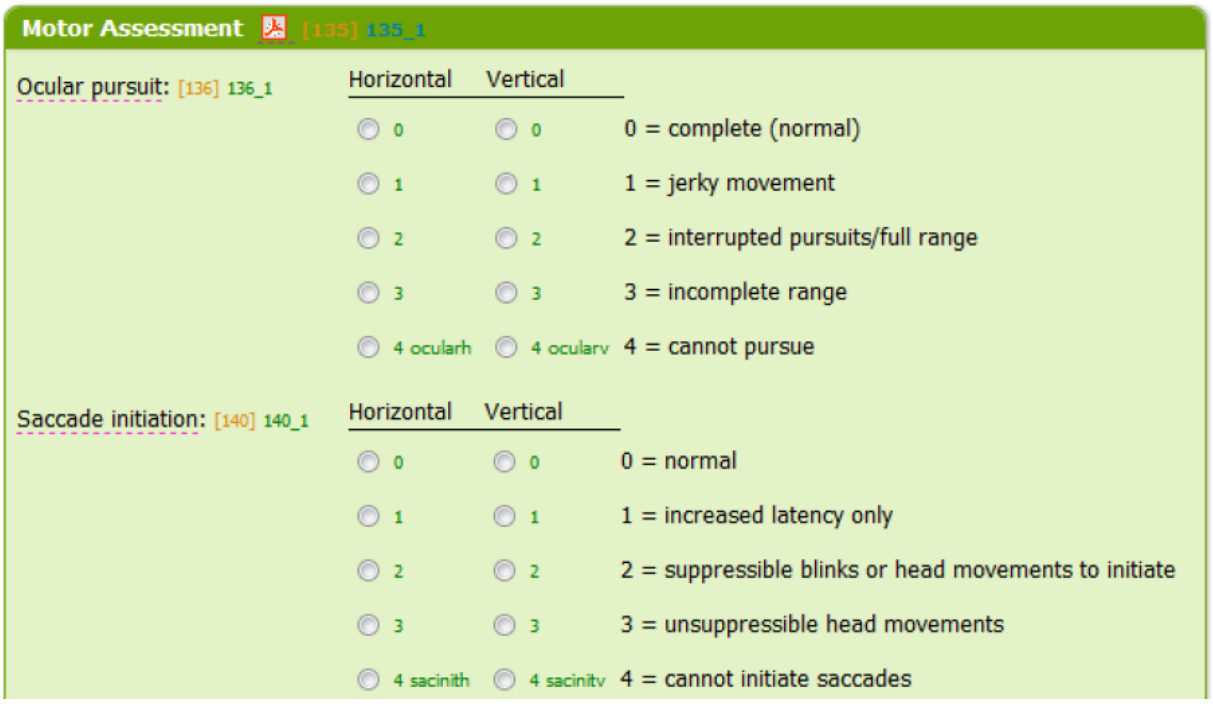


Figure 4.2: Part of the motor section of the UHDRS form.

- ***Cognitive Assessments***: Regarding the assessment of the cognitive functioning, the participants perform three tests: 1) the Categorical Verbal Fluency Test (neuropsychological test which examines the ability to spontaneously produce words from a category within a fixed time [96]), 2) the

Symbol Digit Modality Test (using a reference key, the examinee has 90 seconds to pair specific numbers with given geometric figures) [97] and 3) the Stroop Color and Word Reading Test (which involves reading names of colors and naming the colors of the ink of color words - for example, if the word "red" is printed in green ink, the examinee has to say "green" [98]) [95]. The variables from this form include, for example, the total of correct given answers and errors of each test.

- ***Mini Mental State Examination (MMSE)***: screening test to identify individuals with cognitive deterioration and dementia [99].
- ***Physiotherapy Outcomes Measures***: form regarding the performance of two tests: the "Timed Up and Go" (TUG) and the "30 Second Chair Stand Test". The first is a measurement of mobility, which includes performing tasks such as walking, turning, stopping, and sitting down. The second provides a measurement of one's lower body strength.
- ***Problem Behaviours Assessment - Short (PBA-s)***: used to perform assessments related to behavior symptoms relevant to HD [100], it consists of a short version of the Problem Behaviors Assessment for HD, a 40-item semistructured interview [57]. This instrument measures frequency and severity of symptoms related to altered emotions, thought content and coping strategies. It is an interview that should be performed in the presence of a companion (like the caregiver or spouse) and it includes items covering an extensive range of behaviors such as depressed mood, anxiety, suicidal thought, aggressive behavior, irritability, hallucinations and apathy.
- ***Short Form Health Survey-12***: 12 items that measure the overall health status [101].
- ***Hospital Anxiety and Depression Scale/Snaith Irritability Scale (HADS-SIS)***: The Hospital Anxiety and Depression Scale (HADS) is a self-report rating scale of depression and anxiety symptoms whereas the Snaith Irritability Scale (SIS) is a self-report rating scale of irritability. The items comprise anxiety, depression and irritability scores.
- ***Work Productivity and Activity Impairment-Specific Health Problem Questionnaire (WPAI-SHP)***: questionnaire that measures the effect of the disease on the number of hours missed from work and on productivity.
- ***Columbia Suicide Severity Rating Scale (C-SSRS)***: questionnaire aimed to assess severity and monitor suicidal events [102]. It can be administered by a rater following a structured interview. The items from this form include the answers to questions such as "Have you wished you were dead or wished you could go to sleep and not wake up?" or "Has there been a time when you started to do something to try to end your life but you stopped yourself before you actually did anything?" and others to assess if the participant has suicidal behaviour/thoughts, the severity and even if there has ever been attempts and how many.
- ***Missed Visit***: answered by phone contact, when participant misses visit - reason for missed visit.

Some of the above mentioned assessments are directly related to the diagnosis of depression and/or depressive behaviour in the participant (such as depressed mood scores, presence and frequency of

suicidal thoughts). These items will be referred to as "depression features" or "DEP" throughout this dissertation. Table 4.1 indicates the number of items (and DEP items) of each form.

Form	# Items	# DEP
Medical History	29	-
Variable Items	49	-
UHDRS Motor Diagnostic Confidence (Motor)	34	-
UHDRS Total Functional Capacity (TFC)	6	-
UHDRS Functional Assessment Independence Scale (Function)	28	-
Cognitive Assessments (Cognitive)	40	-
Mini Mental State Examination (MMSE)	1	-
Physiotherapy Outcomes Measures (Physiotherapy)	4	-
Problem Behaviours Assessment – Short (PBA-s)	35	6
Short Form Health Survey-12	10	-
Hospital Anxiety and Depression Scale Snaith/ Irritability Scale (HADS-SIS)	5	1
Work Productivity and Activity Impairment-Specific Health Problem Questionnaire (WPAI-SHP)	4	-
Columbia Suicide Severity Rating Scale (C-SSRS)	30	30
Missed Visit	4	-
Total	284	

Table 4.1: Visits' forms and respective number of items and number depression diagnosis related items (DEP).

4.2.3.B "Profile" data

The data found in "profile" regards general information about the participant, like demographic characteristics, the CAG repeat length, whether the mother/father were affected and at what age they had the first symptoms, if applicable. Also some more disease-related aspects are detailed (if applicable, evidently): age at onset of HD, first symptoms, whether the participant has a medical history of apathy, irritability, psychosis, significant cognitive impairment or motor symptoms compatible with HD and at what age have these symptoms been first noted.

Also, it was from this data file, that the item "ccdep" ("Has depression (includes treatment with antidepressants with or without a formally-stated diagnosis of depression) ever been a part of the participant's medical history?") was selected. It is the binary variable that assumes the value 0 if depression was never part of the person's medical history and 1 otherwise. It is important to mention that there is the item "at what age did depression begin?" but it was deleted from the dataset, as having any answer for this question would mean that depression had been a part of the participant's medical history.

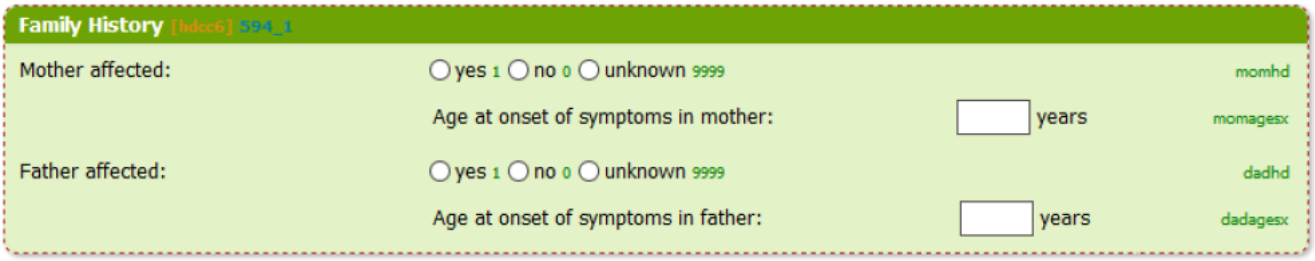


Figure 4.3: Part of the "Profile" questionnaire.

4.3 Data Pre-processing

First, it was necessary to select the data that would be used. From the whole dataset, the following steps were taken (see Figure 4.4):

1. The entries (corresponding to visits) of the dataset in which the age of the participant by the moment of the visit wasn't specified were deleted. This happens when the participant is under 18 years old: in the column *age* it is only written "<18" - this way, it is not possible to know the exact age of the participant.

2. The entries regarding participants who only attended one visit were removed, since this study is intended to be based on longitudinal data.

3. All the participants whose information about the medical history of depression was missing were excluded.

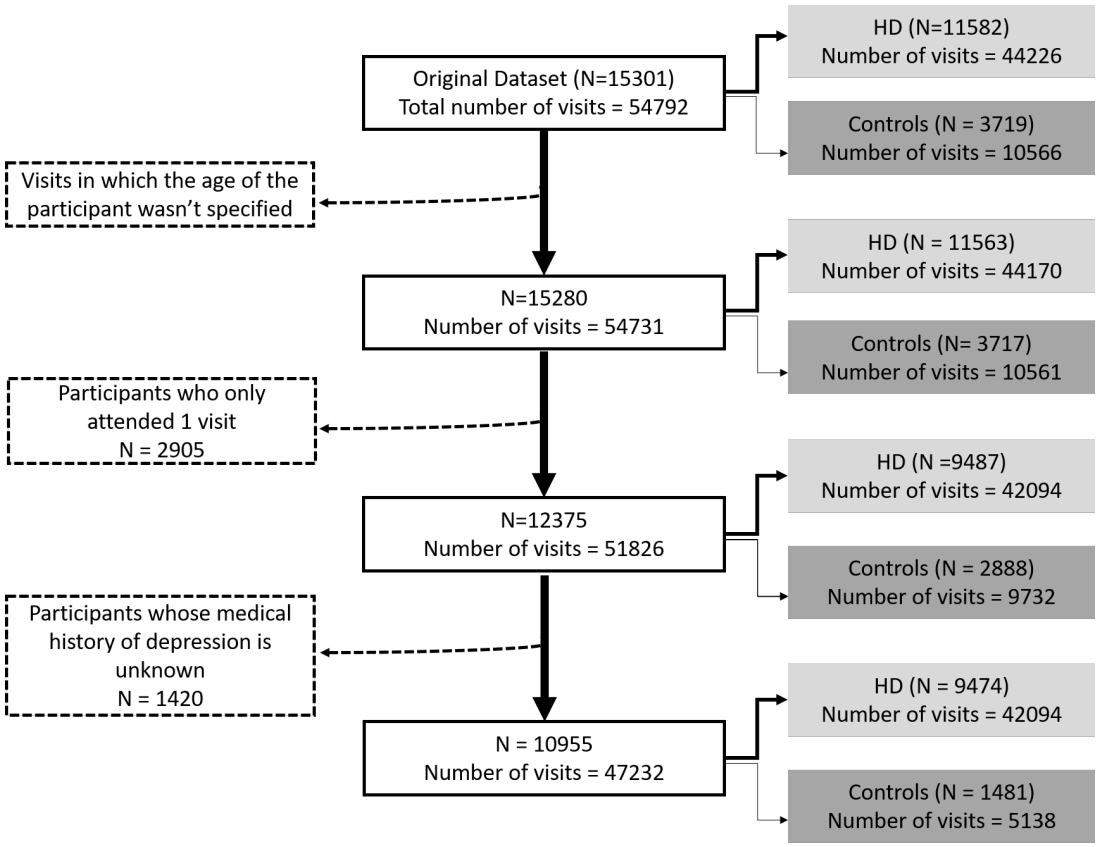


Figure 4.4: Schematic representation of the data removed from the original Dataset.

4.3.1 Feature Standardization

Once different features range between different values, it was necessary to perform standardization in order to set a common a scale [103]. This way, each column was separately standardized using the formula presented in 4.1, where z_i represents the sample x_i after normalization and $\overline{x}$ and σ are the mean and standard deviation of the feature, respectively.

$$z_i = \frac{x_i - \overline{x}}{\sigma} \tag{4.1}$$

After applying this transformation, also called z-normalization or z-score [104], each feature has a mean of 0 and a standard deviation of 1.

A second approach that was used and tested was applying the one-hot encoding method to the categorical features. As mentioned in section 4.2.3, all categorical variables were given in a numerical form, *i.e.* integer encoded, so that the network could read the information - nonetheless, this approach is not optimal, especially when there is no ordinal relationship between the values that the variable assumes (for example, the ethnicity of the participant), because it allows the model to assume a natural order between categories and it may result in poor performance.

One-hot encoding consists of converting each categorical feature in n binary features, with n being the number of possible categories [105]. Regarding the example of the handedness, this feature generates 4 binary features: "left", "right", "mixed" and "NaN" (the last one is for when the value of the variable is missing). Figure 4.5 illustrates this encoding method.

handed
1
1
2
2
1
1
1
1

→

NaN	right	left	mixed
0.0	1.0	0.0	0.0
0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0
0.0	0.0	1.0	0.0
0.0	1.0	0.0	0.0
0.0	1.0	0.0	0.0
0.0	1.0	0.0	0.0
0.0	1.0	0.0	0.0

Figure 4.5: Illustration of the one hot encoding method applied to the categorical features.

Both methods were applied to the categorical features and compared because the one-hot encoding approach, despite being more adequate, implies increased computational complexity, as the feature dimensionality expands considerably.

4.3.2 Handling Missing Data

The provided dataset contains many missing values. Some of the reasons for a certain value of the data to be missing are the following:

1. The value was forgotten or refused to be entered.

2. The value could not have been answered by the participant because of certain characteristics (for example, a question about listening for a deaf participant).

3. The value was marked as incorrect by the data entry person (for example, due to broken measuring instruments).

4. As the different visit files were concatenated and the rows concerning the visits associated to the REGISTRY study and Ad hoc data did not contemplate all the items of the Enroll-HD visits, all the values corresponding to those not assessed variables are missing.

5. There are variables whose answer depends on the answer of another one: for example, when assessing drug abuse and frequency, for each type of drug there is a binary variable for indicating if the patient abuses the drug and another for frequency of abuse - the last is only answered if the first was responded positively.

6. There are variables that correspond to items which are not assessed in all visits. For example, the "Medical History" form is only filled in Baseline visits (first visit after the enrollment into the study) and the form "Missed Visit" in phone contact visits, to document the reason for a missed visit.

The mean value of missing observations per visit is 150 and more than 90% of the information about 82 from the 284 features used are missing. Missing values represent, this way, a large portion of the dataset entries. Hence, it was not viable to remove all the rows nor columns that contained missing observations and it was necessary to handle them. For categorical features, when the one hot encoding method was applied, the solution was to add the category "missing value", as mentioned in the previous section 4.3.1.

For non-categorical features and for the categorical when only integer encoding was used, two things were done: filling the *NaN* with previous valid observation when applicable and then replace the remaining missing entries with "0's". The filling with previous valid observation was only used in the columns in which it was reasonable: for example, as it was mentioned before, "Missed Visit" is a subset of the features that would not make sense to have valid values when the participant attended the visit. The choice of replacing with "0's" the remaining *NaN* was due not only to the fact that it is the column's mean but also because it "marks" them: 0 is a value that otherwise does not appear in the dataset after feature standardization, becoming an indicator of the missing data.

4.3.3 Final Dataset

In order to use the dataset as input of the RNNs, it was necessary to transform it into a 3D matrix with fixed dimensions (# samples, # time-steps, # features), as represented in figure 4.13.

Once the differences among participants regarding the number of time-steps were large, two different methods were used, in order to understand which worked better. Firstly, with the aim of avoiding very long padding sequences and, at the same time, increasing the number of training samples, it was stipulated that the number of time-steps of each sample would be three. This way, from each participant, with n representing the number of visits attended in consecutive years, $n - 2$ samples are originated: for example, a participant who attended to 5 visits originates 3 samples: one from the first to the third

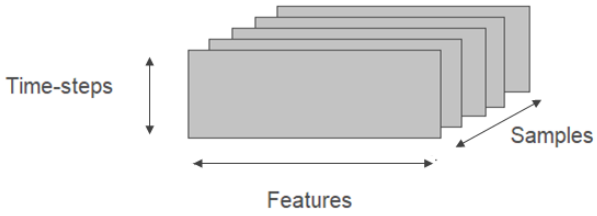


Figure 4.6: Representation of the dataset as a 3D tensor.

visits, another from the second to the fourth and the last from the third to the fifth. Alternatively, samples of 15 time-steps were used, each with the information of one participant (originating, this way, as many samples as participants); those corresponding to who had attended less than 15 visits were pre-padded with 0's [106]; for the ones who attended more than 15 visits (which are only 31), the last 15 were used. It was decided to use samples of 15 time-steps based on the distribution of the number of visits attended per patient (figure 4.7). Figure 4.8 illustrates the two methods used: in the image, the information of a participant who attended to 5 visits originates (1) 3 samples of 3 time-steps and (2) one sample of 15 time-steps, pre-padded with 0's.

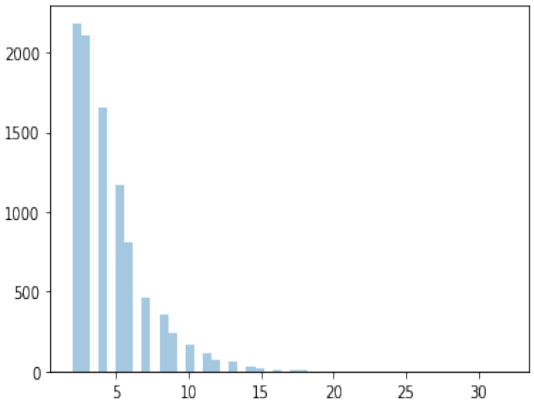


Figure 4.7: Histogram of the distribution of the number of visits attended per patient.

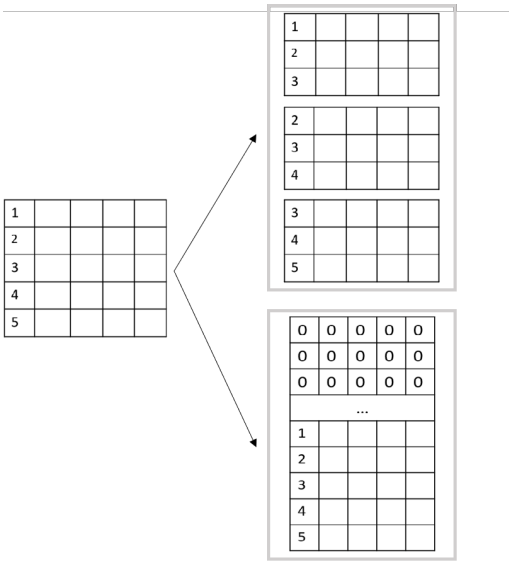


Figure 4.8: Schematic representation of the two methods applied for transforming the participants' data into fixed length samples.

Also, although the visits are supposedly annual, the time-interval between visits was not always of 1 year precisely - so, this value was added as a feature of the visits data. In the first visit of each patient this value corresponds to 0 and for the rest it was calculated from the subtraction of the "visit day" of the previous visit from the "visit day" of the current one. This approach has been used in other studies [77]. Table 4.2 describes the dataset composition after applying the described pre-processing steps.

Table 4.2: Number of participants of each group and correspondent percentage of female participants, total number of visits and mean number of visits per participant.

	N	% female	total # visits	# visits/participant
HD	9474	46%	42094	4.44 ± 2.57
control	1481	40%	5138	3.47 ± 1.48

4.4 Building the Deep Learning model

4.4.1 Model Architecture

During the development of this dissertation, in order to better understand how informative the available clinical data through time is for the purpose of detecting depression, different RNN architectures were tested. The idea was to use LSTMs which would have as input the patient's visits. Then, the "profile" data about each participant was converged to this LSTM network and, finally, 2 other types of RNNs (GRU and standard RNN) were tested. The objective was not to compare the models' efficacy but rather to find the best one for, with the available data, detecting the medical history of depression.

4.4.1.A LSTM network for processing sequential data

The first architecture was built using keras "Sequential" model, which consists in a stack of LSTM layers, receiving as input the 3D tensor with the visits' data.

The first layer of the network is a masking layer for masking the time-steps entirely filled with "0's" (see section 4.3.3).

In order to stack multiple RNN layers on top of each other, all intermediate layers return their full sequence of outputs (a 3D tensor, with the same number of time-steps as the input) and the last layer returns its output at the last time-step (which contains information about the entire sequence), as represented in figure 4.9. This type of architecture, where one single value is output from a sequential input is called "many-to-one".

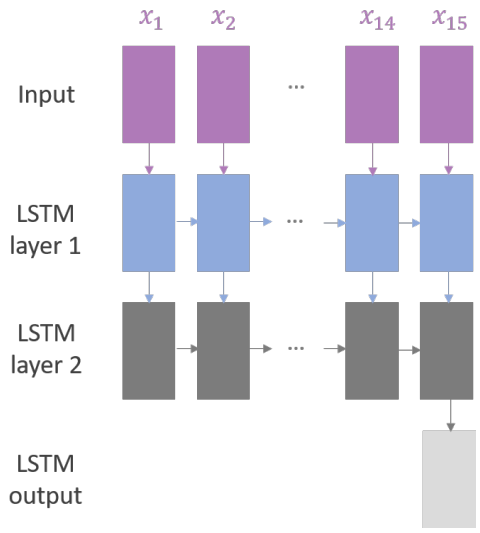


Figure 4.9: LSTM network "many-to-one" architecture.

For all the tested networks, the last layer is a Dense layer: a regular feed-forward layer, which outputs the computation of an activation function to its inputs [107]. The used activation function for the last layer is the sigmoid function, as it will be explained in 4.4.2.B. Figure 4.10 is a plot of the architecture of a Sequential LSTM model, representing all the layers as well as the shape of the input and output of each layer (*None* appears in the place of the number of samples of the tensors shape, as it is not part of the model architecture).

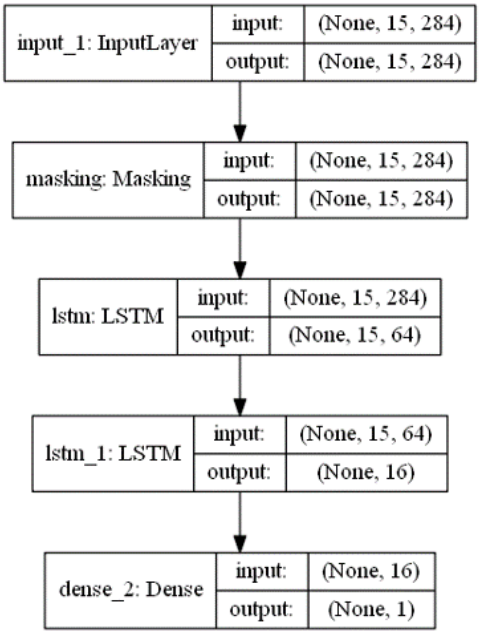


Figure 4.10: Schematic illustration of an LSTM Sequential model structure.

4.4.1.B Combining sequential and non-sequential data

As already metioned, it was aimed to also use the "profile" information of each participant. However, as this is a different type of data (it is static, not sequential like the visits assessments), it didn't make sense to be used as an input of the LSTM network and a different kind of neural layer would be necessary to process it. So far the networks were built using the keras Sequential model [108], which makes the assumption that the network has exactly one input and exactly one output and that it consists of a linear stack of layers, making it impossible to have multiple inputs [109], [60]. For this, there is Keras functional API [110], a more flexible way to use Keras, where layers are used as functions and are connected pairwise [111]. Figure 4.11 illustrates the difference between the Sequential and functional API.

Hence, in parallel to the LSTM layers processing the sequential data, a Dense layer with a linear activation was added to process the profile data. As represented in figure 4.11 (b), the outputs of the last LSTM layer and of the Dense layer are concatenated.

4.4.1.C Trying other RNNs: the GRU and "SimpleRNN"

As discussed in chapter 3, GRUs work using the same principle as LSTM, but have a less complex architecture, being less computationally expensive, achieving, nonetheless, comparable performance in many tasks. Therefore, these two types of RNN were used and compared for the present task.

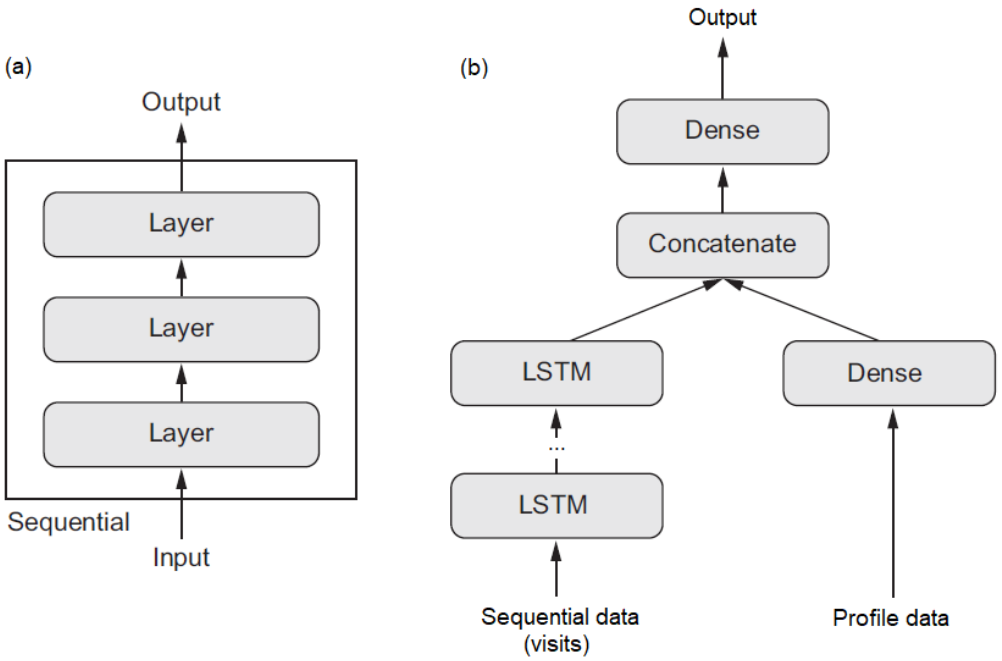


Figure 4.11: Schematic representations of the keras (a) sequential and (b) funcional API models (adapted from [60]).

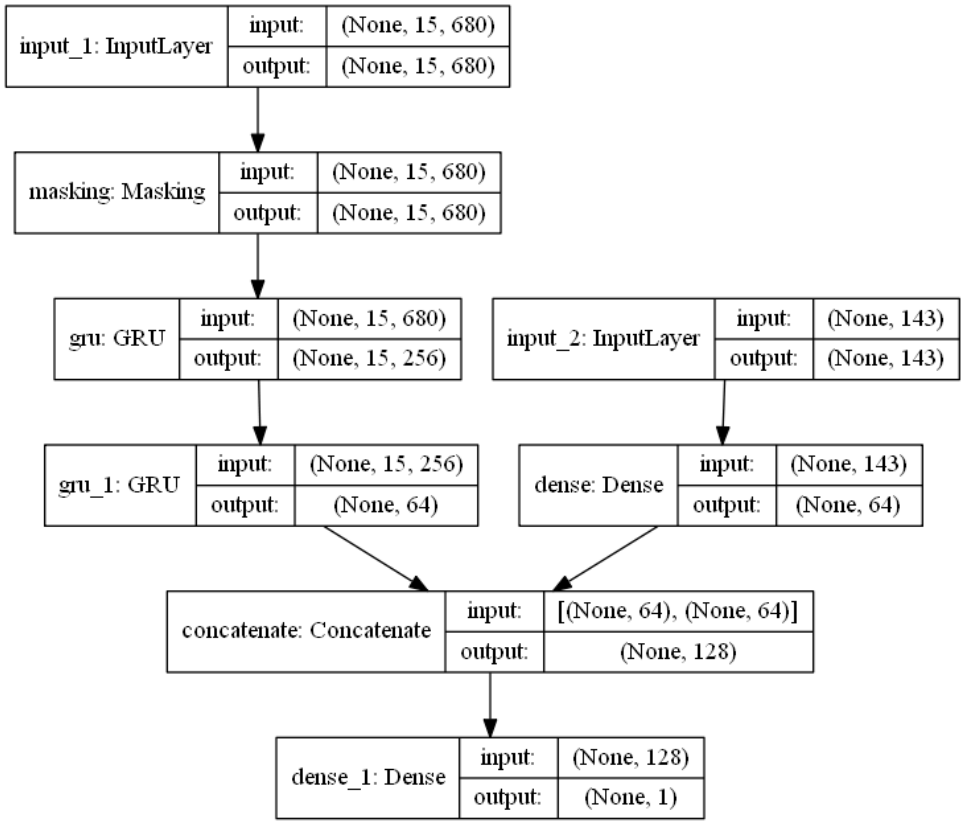


Figure 4.12: Schematic representation of the architecture of a GRU network, built with the keras Functional API.

Finally, once the input sequences have a small number of time-steps, it was also tested to use a standard RNN with the tanh activation (as discussed in 3.2.1, the major problem of these networks was the vanishing gradient, when learning long term dependencies).

Figure 4.12 is a plot of the architecture of a multi-input model, representing all of its layers as well as the shape of the input and output of each layer.

4.4.2 Training and Testing

The input of the model is the 3D matrix described in 4.3.3 and the output a binary variable (*ccdep*). The data was randomly divided into training and testing sets, in a proportion of 80 to 20 percent, respectively.

4.4.2.A Learning, Validation and Early Stopping

In order to prevent overfitting, 20% of the training data is used to create a validation set. During the training phase, after each epoch, the model is tested on this set of data and the loss is measured, so that when the loss value starts increasing, the training stops. This is done using callback functions: *early stopping*, that grants that the training phase stops when the validation loss starts increasing (more precisely, after waiting 12 epochs without any improvement), *model checkpoint* that saves the network's parameters each time the validation loss improves so that the best model (*i.e.*, the parameters that lead to the lowest validation loss) is used (and not the last one). Also, *ReduceLROnPlateau* monitors the validation loss and if it decreases and no improvement is seen after 6 epochs, the learning rate is reduced by a factor of 0.9. After each epoch, the training data is shuffled as it is to be avoided to provide the training examples in a meaningful order, once this may bias the optimization algorithm [66].

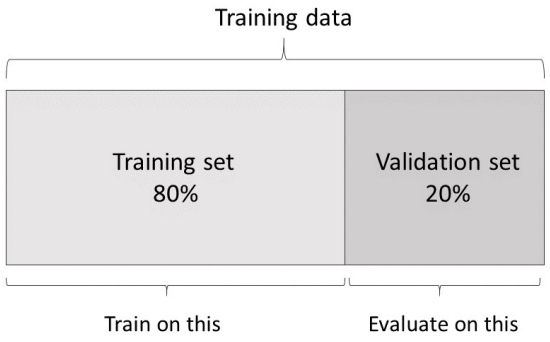


Figure 4.13: Representation of the training data division into training and validation sets.

4.4.2.B Hyper-Parameters

In the previous section the different types of RNNs that were used through the development of this dissertation were enumerated. However, besides the RNN algorithm there are different parameters of the architecture of the network that have to be tuned to optimize our model.

First, it was necessary to find an appropriate model size, *i.e.* an adequate number of layers and nodes. The approach carried out was to start with a simple network, with few units and parameters, and increase its complexity until the addition of parameters no longer adds representational power and the performance stops improving. The reason for this is the fact that having more parameters, despite allowing to learn more complex representations, it is not only more computationally expensive but also

may lead to learning patterns specific from the training data, leading to overfitting, instead of greater generalization power. Regarding the nodes, each layer had a number of nodes corresponding to a power of 2 (32, 64, 128,...), with deeper layers having fewer nodes (in a "triangular" shape), as the output is 1 single value (and, so, the last layer has only one node). This way, for each number of layers, the model was trained with different sets of nodes and the corresponding training and validation curves were evaluated: the set of nodes that lead to the lowest validation loss value would be selected.

Since we are predicting a binary outcome, the loss function used was binary cross entropy (expression 4.2) and the activation function of the last layer of the neural network was the sigmoid function (as, from expression 4.3, it results in a value between 0 and 1, encoding a probability of the sample belonging to class "1", *i.e.*, to have a medical history of depression) [60]. The used optimizer was the Adam optimizer.

$$\textbf{Binary Cross Entropy} : L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(\hat{y_i}) + (1 - y_i) \cdot log(1 - \hat{y_i}) \tag{4.2}$$

$$\textbf{Logistic Sigmoid} : \sigma(x) = \frac{1}{1 + e^{-x}} \tag{4.3}$$

Where y_i is the true class (0 or 1) and $\hat{y_i}$ is the predicted value for the sample i.

4.4.2.C Class imbalance problem

Once the data being used to develop the classifier is imbalanced - around 60% of the total number of participants (who were part of the present study) have a medical history of depression (regarding the HD patients, it was even more imbalanced) - it was necessary to balance the classes. For that, each class was mapped to a value (based on the representativity of the class) used for weighting the loss function (during training only). This way, samples from an under-represented class had a greater impact on the loss function.

Table 4.3: Classes representativity in each group of participants.

	N	*ccdep*=1 (%)	*ccdep*=0 (%)
HD	9474	66.6	33.4
control	1481	35.4	64.6
HD + control	10955	62.3	37.7

4.4.2.D Dropout

Another regularization technique that was used was adding dropout to the network. Complex models, *i.e.*, models that have many parameters, are very likely to overfit to the training data. Dropout consists of randomly and temporarily dropping out (by setting to zero) units along with their connections from the network during training [112]. As the weights of the network will be larger than normal because of dropout, the weights have to be scaled by the chosen dropout rate (fraction of units to drop). There are different ways to do this rescaling, the way Keras implement dropout the weights are multiplied by the dropout rate after each weight update at the end of the mini-batch.

It is important to point out that the dropout was applied to the input connections (*i.e.*, to the non-recurrent connections), as standard dropout does not work well on RNNs [113].

4.4.3 Model Performance Evaluation

In binary classification, the data is classified with one of two labels - in this case, 0 or 1, for having a medical history of depression or not, respectively. Another nomenclature we can use to distinguish the outputs in binary classification is between "Positive" and "Negative": this way, the possible outcomes are "True Positive", "True Negative", "False Positive" and "False Negative", as illustrated in table 4.4. In this study, the positives represent those with a medical history of depression (or "1's").

Table 4.4: Confusion Matrix

		Predicted class	
		Positive	Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

The first metric used for evaluating the developed model was *Accuracy*, the most common metric for classifier evaluation. It assesses the overall effectiveness of the algorithm by computing the probability of the correct prediction [114], using the formula presented in expression 4.4 [115].

$$\mathbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.4}$$

Since we are evaluating a binary classifier on an imbalanced dataset, *accuracy* alone can be misleading [114]. In clinical practice, a False Positive occurs when someone is diagnosed with a disease without having it and a False Negative when a person who has the disease is given a negative diagnosis result.

Although it is not the purpose of this dissertation to "make a diagnosis" based on the presented data (what is being predicted is not if the person is currently suffering from depression, but whether if he/she has had it at any moment in the past), in the context of depression, particularly in HD, a False Positive would mean that someone who does not have depression is diagnosed as having it, which would probably lead to an unnecessary prescription of antidepressants - as mentioned in chapter 2, this should be avoided as these drugs are linked to adverse secondary effects; on the other hand, a False Negative occurs when someone who has depression is nonetheless given a negative diagnosis - obviously, this is also dangerous, as someone with depression should follow the adequate treatment.

So, besides accuracy, two other metrics were used: the True Positive Rate (TPR) and the True Negative Rate (TNR). The TPR or sensitivity is the rate of participants belonging to the positive class who were correctly predicted as positive (computed as in expression 4.5) whereas the TNR or specificity is the rate of participants belonging to the negative class who were correctly predicted as negative (4.6). Balanced accuracy is the mean value of these two metrics (4.7), so we can summarily say that the objective is to achieve the highest balanced accuracy with the smallest difference between the TPR and the TNR.

$$\textbf{TPR or Sensitivity} = \frac{TP}{TP + FN} \tag{4.5}$$

$$\textbf{TNR or Specificity} = \frac{TN}{TN + FP} \tag{4.6}$$

$$\textbf{Balanced Accuracy} = \frac{TPR + TNR}{2} \tag{4.7}$$

Finally, as the focus of this dissertation is studying the detection of medical history of depression specifically in HD, 2 datasets (from the two main groups of participants: HD and controls) were separated. The different types of architectures described in the present chapter were trained and tested using the data regarding HD participants. Afterwards, the obtained results are compared to applying the RNN model to the whole dataset (HD + controls) and to the control dataset.

The model that is being developed aims to predict, based on the evolution of the disease in each participant, if depression was ever a part of their medical history. Accordingly, in order to better understand the impact of the clinical observations on the predictability of this condition, some experiments were conducted using different combinations of features (for example, the model performance was assessed both with and without the information regarding the features directly related to the diagnosis of depression).

5

Results and Discussion

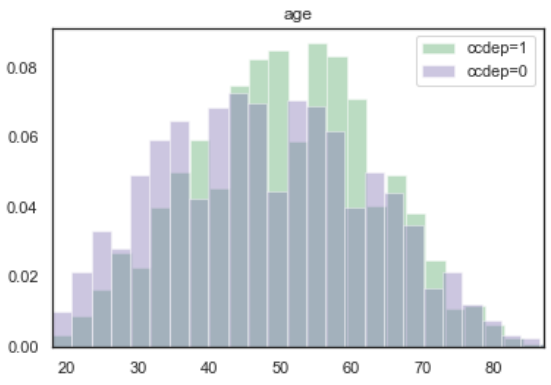
Contents

5.1 Feature Analysis

Some interesting variations between features distributions from the participants with *ccdep*=0 and with *ccdep*=1 were found (important to remember that *ccdep* is the variable behind the distinction of the two classes and it is the answer to the question "Has depression (includes treatment with antidepressants with or without a formally-stated diagnosis of depression) ever been a part of the participant's medical history?").

HD is a progressive disease and, as such, the symptoms and impairments get worse with time. This way, firstly, the distribution of the patients' age from the 2 classes is shown in figure 5.1. As it is possible to see in the image, the ages range between roughly the same values (from 18 to 87 in the "positive" class and from 18 to 92 in the "negatives") and their mean values are of 50.6 and 48.1. There's, nonetheless, a slight deviation to the younger ages in the negative class - or, more precisely, while in the positive class the distribution has a sharper shape (50% of the ages are between 41 and 60), in the negative class there are less people around the mean age and slightly more between 18 and around 40. This is important for the stated reason that the age influences the deterioration state and some of the found differences in the variables distributions could be due to this.

Notice that figures from 5.1 to 5.8 are histograms that represent the distribution of data variables by forming bins along the ranges of the data and the bars show the relative frequency of observations which fall in each bin.



ccdep=1	age
count	29205.000000
mean	50.592193
std	12.784351
min	18.000000
25%	41.000000
50%	51.000000
75%	60.000000
max	87.000000

ccdep=0	age
count	14965.00000
mean	48.06876
std	14.43983
min	18.00000
25%	37.00000
50%	48.00000
75%	59.00000
max	92.00000

Figure 5.1: Density distributions of the participants' ages from the two classes.

Two other plots worth displaying are the distributions of the depression scores of the PBA-s and HAD-SIS forms, where higher score values indicate higher severity of depressive symptoms. First, it is important to retain that these values come from sequential data (see in figure 5.1 that the count is the number of visits and not the number of participants) and that we are distinguishing people that have never had depression from those that did and the data regarding those that did is not only from the period of time while they were experiencing it - in other words, in the data from "ccdep=1" there is information about moments prior and/or posterior to depression. This explains why, for example, there are so many samples from "ccdep=1" with very low values of depression scores. On the other hand, there's a considerable quantity of high depression scores from participants with "ccdep=0" which might

indicate that some people were not aware that they might have had this disorder, or that these scores are not completely accurate or even the discussed hypothesis that these parameters are subjective and may be interpreted in different ways by different people (making it important to have additional information and to use objective automatic methods and not only human interpretation).

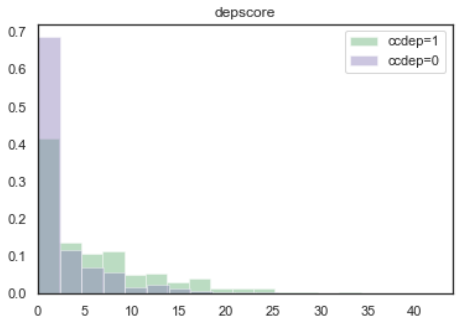


Figure 5.2: Density distribution of the *depscore* feature.

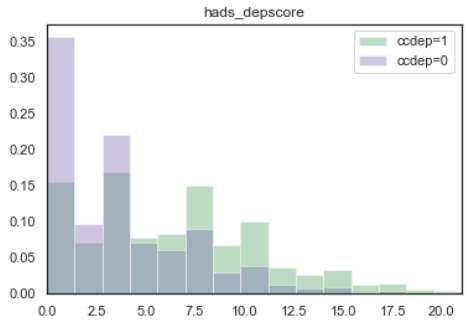


Figure 5.3: Density distribution of the *hads_depscore* feature.

As discussed in chapter 2, there's great evidence that experiencing depression in HD is positively correlated with greater functional impairments, due to a number of factors. In figure 5.4 some of the assessments that measure these impairments are represented and the shown distributions corroborate this idea.

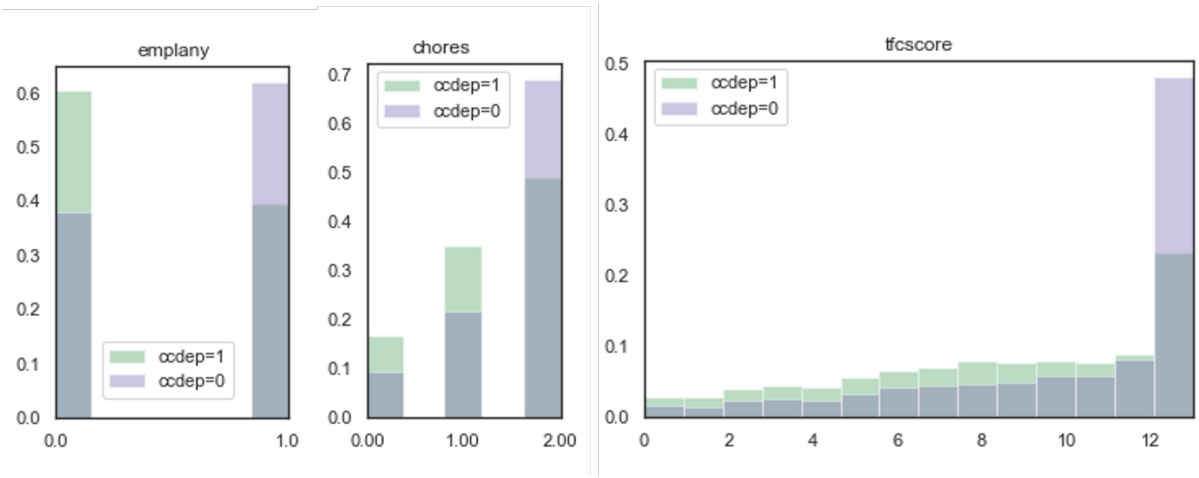


Figure 5.4: Functional assessments density distributions: (a) *emplany* is the binary variable wich answers the question "Could subject engage in any kind of gainful employment?", (b) *chores* is the categorical variable regarding the capability to do the domestic chores (0- unable; 1- impaired; 2- normal) and (c) *tfcscore* Total Functional Capacity Score (from the UHDRS, see section 4.2.3.A).

Regarding the motor assessments, the inter-class variation in the distribution plots is not so evident, as shown in figures 5.5 and 5.6. Nonetheless, there is a higher percentage of assessments where there's absent dystonia in the negative class and, for example, while in the positive class there are roughly as many assessments made where there's absent facial choreatic as where there's slight intermittent movements, in the negative class from the first two the second category, the incidence reduces by around 20%.

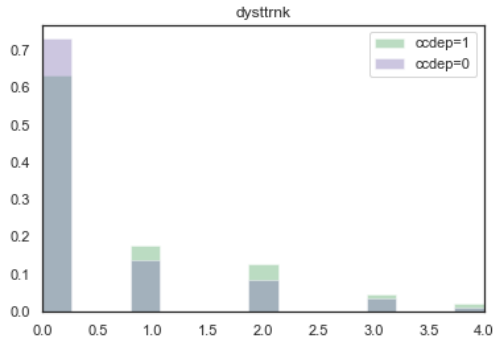


Figure 5.5: Density distribution of the *dysttrnk* feature, a motor assessment regarding the trunk dystonia (0- absent; 1- slight intermittent; 2- mild common or moderate intermittent; 3- moderate common; 4- marked prolonged).

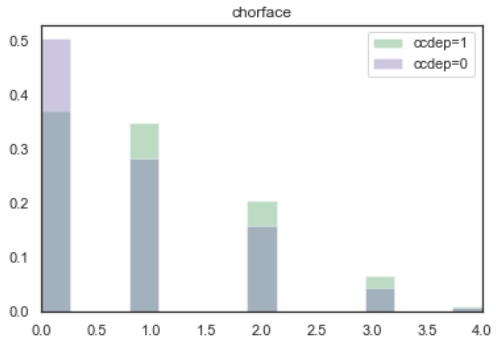


Figure 5.6: Density distribution of the *chorface* feature, a motor assessment regarding facial choreatic movements (0- absent; 1- slight intermittent; 2- mild common or moderate intermittent; 3- moderate common; 4- marked prolonged).

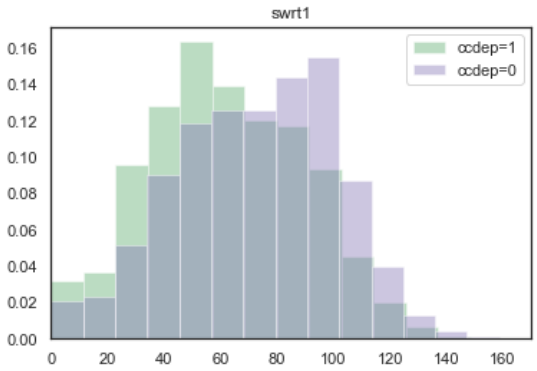


Figure 5.7: Density distribution of the *swrt* (stroop word reading test) scores.

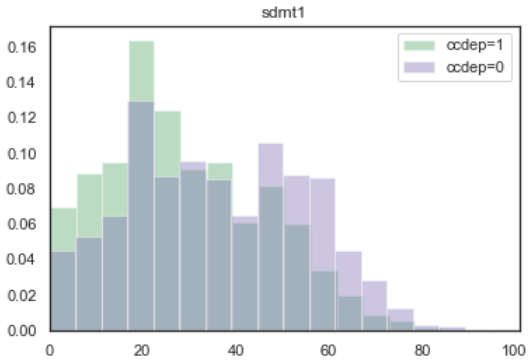


Figure 5.8: Density distribution of the *sdmt* (symbol-digit modality test) scores.

Both in HD and in the general population, depression has been correlated to cognitive dysfunction, as mentioned in chapter 2. Figures 5.7 and 5.8 show that those from the positive class have higher tendency to perform worse in cognition evaluation exams, such as the *stroop word reading test* and the *symbol-digit modality test* (see section 4.2.3.A). The values in the x axis of the graphics from figures 5.7 and 5.8 are the total number of correct answers given during those exams and lower scores are indicative of lower cognitive performance.

Although it is interesting to observe these distribution plots, no conclusions can be drawn and no criteria to distinguish one class from the other could be extracted from this information alone.

5.2 Deep Learning results

5.2.1 LSTM models

As explained in 4.4.2.B, in order to optimize the model's architecture, several hyperparameters were tuned. Figure illustrates the training and validation curves from training the LSTM Sequential model (from 4.4.1.A) with 3 layers, each curve from training it with a different set of nodes.

From the curves shown in figure 5.9, it is possible to see that the network with 512, 256, 128 nodes

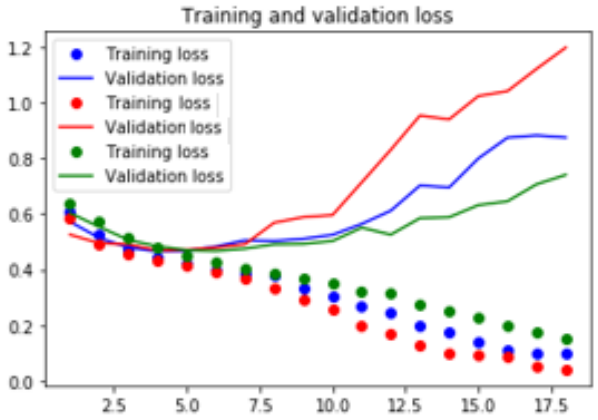


Figure 5.9: Comparison of the validation and training loss curves from training 3 LSTM networks only differing in the number of nodes. Each is composed of 3 LSTM layers with the following number of nodes: red: 512, 256, 128; blue: 256, 128, 62; green: 128, 64, 32.

(curves shown in red) overfits to the training data not only early in the training (the validation loss starts increasing from the 4th epoch) but also very rapidly (by 18th epoch, the validation loss is the double of that after the 1st epoch). Regarding the smallest network from these 3 (represented by the green curves), although it is the one with the least overfitting tendency (from the 6th epoch forward, it provides the lowest validation loss of the three curves), it never reaches a loss value in the validation set as low as the blue line in the 4th epoch (of 0.465), which indicates lower representational power.

Regarding the matter of whether it was more advantageous to use samples of 3 time-steps or longer sequences, each representing the total longitudinal information of one participant (issue posed in section 4.3.3), the results are shown in table 5.1. From these values, it is, undoubtedly, advantageous to use the longer samples. These results show that having more longitudinal information outperforms the advantage of having more samples, and so, more training data (table 5.2 indicates the difference in the number of samples available using the two methods).

Table 5.1: Comparison between using samples of 3 timesteps each or of 15 timesteps each.

	#timesteps	Accuracy	TPR	TNR	Balanced Accuracy
2 layers	3	0.737	0.841	0.476	0.658
	15	0.745	**0.858**	0.528	0.693
3 layers	3	0.738	0.840	0.482	0.661
	15	**0.751**	0.804	**0.651**	**0.728**
4 layers	3	0.732	0.831	0.483	0.657
	15	0.749	0.825	0.600	0.713
5 layers	3	0.732	0.856	0.454	0.654
	15	0.752	0.805	0.646	0.726

Table 5.3 shows the results obtained using the same LSTM Sequential model as in 5.1, but giving as input the data with a different pre-processing approach, with the categorical features encoded using a one-hot scheme (using the samples of 15 time-steps, like all the results presented subsequently).

Table 5.2: Number of data samples available for training, validation and testing, when using samples of 3 and 15 timesteps.

# ts	# training samples	# validation samples	# testing samples
3	14787	3697	4620
15	6063	1516	1895

Table 5.3: Model Performance after encoding the categorical features using a one-hot scheme.

	Accuracy	**TPR**	**TNR**	**Balanced Accuracy**
2 layers	0.766	**0.841**	0.619	0.730
3 layers	0.766	0.824	0.648	0.736
4 layers	**0.772**	0.821	**0.672**	**0.747**
5 layers	**0.772**	0,831	0.655	0.743

Overall, the performance improved. Nonetheless, it is interesting to notice that while previously the best results were achieved with 3 LSTM layers, after performing this encoding it was necessary to add an extra layer to obtain higher values of our metrics. This can easily be explained by the great expansion of dimensionality of the feature space that performing one hot encoding implies - each categorical feature was "transformed" into as many features as the number of categories it includes (as explained in 4.3.1) - going, this way, from having 284 to 680 features. More complex datasets require more complex networks.

Figure 5.10 illustrates the effects of adding dropout during training.

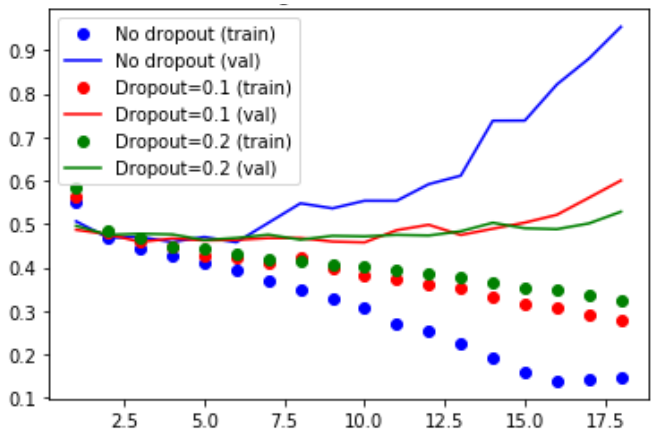


Figure 5.10: Comparison of the validation and training loss curves from training the same network without dropout (blue), with a dropout rate of 0.1 (red) and with a dropout rate of 0.2 (green).

Adding dropout allowed to reduce considerably the overfitting phenomenon. These curves were obtained during training of the 4-layer LSTM network, after encoding the categorical features using the one-hot scheme. The tendency to overfit tends to worsen with the size of the model, as it is possible to confirm when comparing the curves of figure 5.9 with the blue one of figure 5.10. Using a dropout rate (fraction of units to "drop") of 0.1 allowed to dissipate this problem very considerably, as by the 10th epoch the validation loss had still not rised and by the 18th epoch it was about 3 decimals lower than the

validation loss obtained without using dropout. A dropout rate of 0.2 would not improve the performance as the best achieved values of validation loss were worse than when using a rate of 0.1.

Reflecting over the values obtained for each metric with the data regarding the patient's annual visits, there are some things worth noticing. First, with this information, it is possible to distinguish between the two classes with an accuracy of 77%. Secondly, as expected once we are dealing with an imbalanced dataset (with 66.6% of the participants belonging to the "positive" class), there is a considerable discrepancy between the True Positive Rate and the True Negative Rate: while (in the best case) the network can distinguish correctly 82% of the positives as belonging to that class, from the negative class, it only identifies them as such in 67% of the cases - the model has a greater tendency to classify a sample as being positive. This shows that giving higher importance (through the enhancement of the loss function value) to a misclassification of a sample from the minority class during training did not make the TPR and TNR have similar values.

Next, it will be assessed whether complementing the model with additional, non-sequential data brings improvements.

5.2.2 Combining sequential and profile data - the functional model

Table 5.4 shows the results obtained after adding the profile information about the patient to the model, with the adjacent changes (explained in 4.4.1.B).

Table 5.4: Model Performance after adding the "profile" information.

Data	Accuracy	TPR	TNR	Balanced Accuracy
visits	0.772	0.821	0.672	0.747
visits + profile	**0.792**	**0.845**	**0.687**	**0.766**

All of the performance measures improved, supporting the idea that these data are informative for our purpose. Information like the medical history of the parents, first symptoms noted, age at which they have been first noted help in the task of detecting depression.

5.2.3 Comparison with other RNNs: "SimpleRNN" and GRU

As proposed, the LSTM model performance was compared to a standard RNN and GRU.

Starting with the comparison between the LSTM model and the standard RNN (or "SimpleRNN" as the Keras layer is called), although by looking at the values from 5.5 it is not immediately obvious (once the accuracy increased by 0.2%), the performance slightly worsened. Although the accuracy improved, it did at the cost of a worse specificity: as previously discussed, in an imbalanced dataset accuracy alone can be misleading as an improvement could simply be due to an increase in the number of samples being classified as belonging to the most common class. Another aspect to notice is the standard deviation of the result values - for the simple RNN, the metrics values had much larger variations than when using LSTMs, which indicates lower consistency. Nonetheless, the difference between performances was very

Table 5.5: Performance comparison of different RNN models (mean and standard deviation of the metrics obtained when using different train and test sets).

Model	Accuracy (sd)	TPR (sd)	TNR (sd)	Balanced Accuracy (sd)
LSTM	0.792 (0.007)	0.845 (0.028)	0.687 (0.039)	0.766 (0.008)
Simple RNN	0.794 (0.009)	**0.856** (0.046)	0.668 (0.082)	0.762 (0.021)
GRU	**0.796** (0.008)	0.850 (0.025)	**0.690** (0.043)	**0.770** (0.011)

Table 5.6: Sizes of the models.

Model	# layers	# units	#parameters
LSTM	4	480	$\approx 1,36 \times 10^6$
Simple RNN	3	448	$\approx 3,4 \times 10^5$
GRU	4	480	$\approx 1 \times 10^6$

small, indicating that the number of time-steps was low enough for not being affected by the vanishing gradient phenomenon.

The GRUs were the type of RNN that lead to better results. As discussed in chapter 3, GRUs are very similar to LSTMs, as they are both gated structured RNNs developed with the purpose of solving the vanishing gradient problem. The main difference relies on the complexity of the architectures - the GRU can be thought of as a simpler version of the LSTM. As it was observed, the standard RNN's (the simplest of the networks, a simple tanh) performance did not differ much from the LSTM, giving the idea that learning this data for the present task does not require a very complex algorithm, which may be the reason for the GRU model to achieve slightly better results.

All the results shown in the next sections were obtained using the functional model with the GRU network for processing the sequential data.

5.2.4 HD vs controls

So far, the results that have been shown only regard the detection of depression in HD patients. In this section we will compare the predictability of the presence of a medical history of depression in patients with HD only, in controls only and in all participants.

Table 5.7: Model performance comparison between groups.

Dataset	Accuracy	TPR	TNR	Balanced Accuracy
HD	0.796	0.850	0.690	0.770
Controls	0.728	0.668	0.764	0.716
HD + Controls	0.772	0.822	0.689	0.756

Comparing the first two lines of table 5.7, we see that both the accuracy and the balanced accuracy

Table 5.8: Number of data samples available for training, validation and testing, when using each of the datasets.

Dataset	# training samples	# validation samples	# testing samples
HD	6063	1516	1895
control	948	237	296
HD + control	7011	1753	2191

of the model worsens when using the controls data, which can be due to a number of factors. First, the amount of training data is very small (as indicated in table 5.8, when using the HD dataset 6063 samples are used for training while with the controls dataset, only 948), which is a limiting factor in Deep Learning. Secondly, the used data comes from a database developed for purposes of studying the Huntington's Disease and the information gathered is, therefore, probably not ideal for the objective of distinguishing healthy people from having or not a medical history of depression, despite including assessments regarding this matter. It is worth noticing the increase in the TNR (and the "swap" between the specificity and the sensitivity of the model), which is probably a consequence of the fact that in the control group the representativity of the classes is of 35.4% and 64.6% for class "positive" and "negative", respectively, an almost reverse context comparing to the HD group. This shows that the method used during training to balance classes (of weighting the loss function according to the classes representativity) did not completely prevent the model's lower tendency to classify a sample as belonging to the underrepresented class.

Concerning the results from using the entire dataset (HD+control), it can be concluded that what is indicative of the class the participant belongs to in one group is not the same as in the other. Using the entire dataset, there were two reasons that could have contributed to accomplishing better results: a larger training set (4.2) and a diminished imbalance between classes (while 66.6% of the HD samples belong to the "positive" class, in the whole dataset this class represents 62.3% of the totality of samples).

One possible reason for the higher rate of wrong predictions is that features that are informative for one group may not give any useful information to the model when making a prediction for the other. Or, even if the same features are informative, it could be in different ways (for example, anxiety or apathy scores could be both informative for healthy and for HD participants regarding depression but a score value that for a healthy participant could be an evidence of depression for an HD patient could indicate the opposite). Finally, there could also be different temporal and non-temporal patterns behind the classification of the two groups.

5.2.5 What is giving useful information to the network?

As mentioned in chapter 4, there are items in the data that are used in clinical practice to make the diagnosis of depression. In this section, it is aimed to analyze the impact of various sets of features on the performance of the network.

Observing the results from table 5.9, we see that using only the depression related features (DEP) along with the profile, the accuracy slightly increases (about 0.8%) comparing with the use of all features

Table 5.9: Performance metrics obtained using different sets of features.

Visits features	Profile	Accuracy	TPR	TNR	Balanced Accuracy
All	Yes	0.796	0.850	**0.690**	**0.770**
All	No	0.772	0.822	0.672	0.747
All\DEP	Yes	0.770	0.831	0.648	0.740
All\DEP	No	0.740	0.820	0.580	0.700
DEP	Yes	**0.804**	**0.875**	0.665	**0.770**
DEP	No	0.774	0.831	0.659	0.745
-	Yes	0.772	0.861	0.592	0.727

(DEP+profile+others); nonetheless, the specificity of the model worsens (by about 2.5%); the balanced accuracy remains because the TPR increased by the same percentage as the TNR decreased. This means that the model became less capable of detecting the samples from the less represented class when it didn't have access to all the other assessments made during the visits. Regarding the results when the only given information is the DEP set of features, all the metrics values worsen: comparing with the values obtained using all features, the accuracy decreases 2.2%, the TPR 1.9% and the TNR 3.1%. This corroborates the idea that these scores and assessments alone are not an as effective tool in detecting depression and should be complemented with more objective data about the patient.

Nevertheless, these features are informative and should be included - this is proven by the results obtained when they were not used, which were worse than all that were obtained when using them. The other visits' assessments alone are the less informative set of features, originating an accuracy of 74% and a specificity of only 58%.

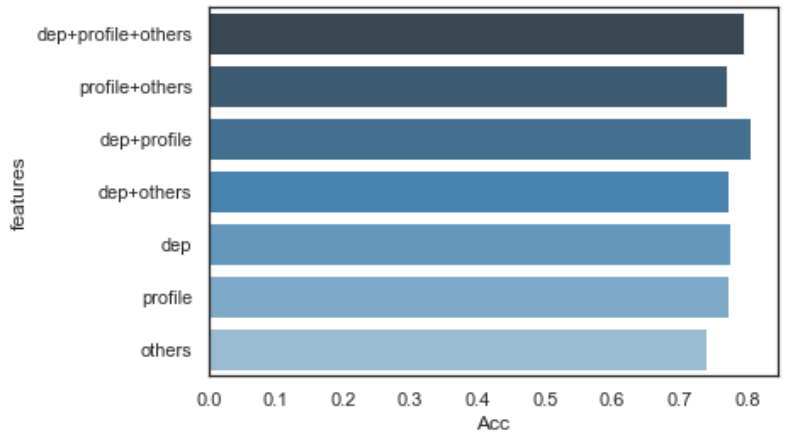


Figure 5.11: Comparison of the accuracy obtained with the different sets of features.

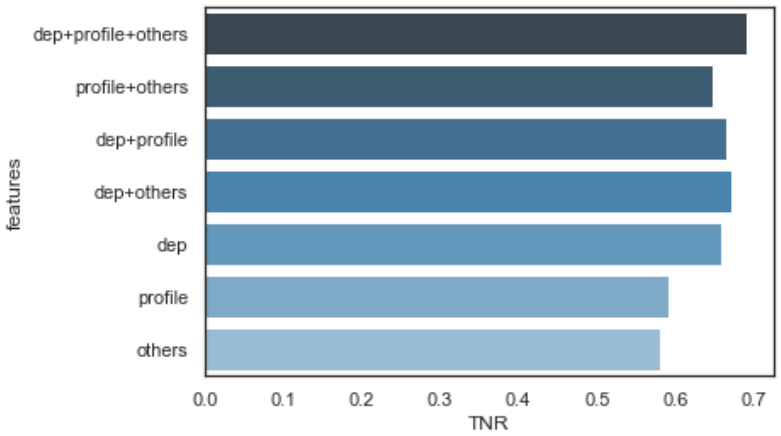


Figure 5.12: Comparison of the specificity obtained with the different sets of features.

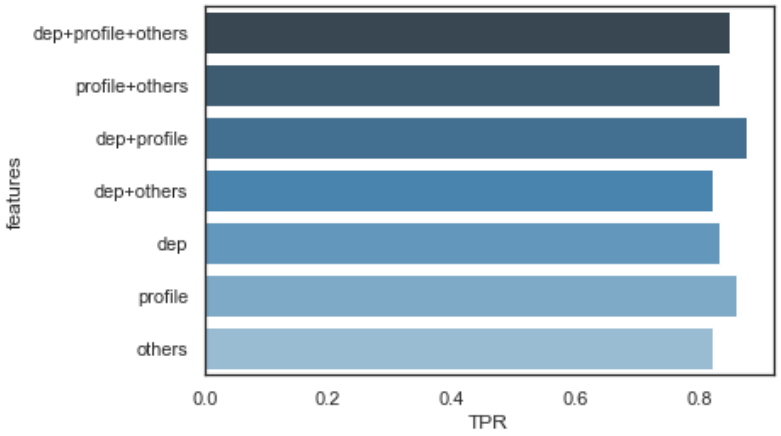


Figure 5.13: Comparison of the sensitivity obtained with the different sets of features.

6

Conclusions and Future Work

Contents

6.1 Conclusions

In HD, psychiatric symptoms, particularly depression, can develop during the prodromal stage or when the disease is manifest. Unlike many other psychiatric and non-psychiatric symptoms, depression is not correlated to disease progress (a person is not more likely to develop this disorder or experience enhanced depressive symptoms with the aggravation of the disease) and it is actually less frequent in late stages. With this in mind, from the obtained distributions of the different assessments of HD participants from the two distinguished classes, it is plausible to conclude that there is a higher tendency of those who suffered from depression to experience exacerbated dysfunctions. No causality relation can be inferred but the question of whether the physiological mechanisms behind depression could lead to an exacerbation of the impairments experienced in HD remains. Obviously, this relation could be indirect, as if, for example, it was the administered antidepressants that were on the origin of the aggravated symptoms.

The main objective of the dissertation was to build a model able to detect, from clinical longitudinal data, cases where depression had been a part of the medical history (includes treatment with antidepressants with or without a medical diagnosis). Regarding the results obtained with it, overall, we can say that the applied method fits the proposed task and that, with further improvements, it is very plausible that it could be used in clinical practice (as we are dealing with a disease that leads to dementia, the patients may reach a point where they may not be able to tell if they have had or not been through depression and in that case this could be useful). Furthermore, the obtained results show that the approach of using clinical data (not only from neuropsychiatric tests but also regarding cognitive, motor, functional aspects and more general personal data) is informative for the purpose of detecting depression and that Recurrent Neural Networks are able to use this data and extract useful outputs.

The first thing that was tested was if it would be more gainful to use the whole clinical picture (the complete sequences of data) of each participant or to divide it into smaller sequences of 3 time-steps, with the advantage of having an increased number of available samples (which is very important for training Deep Learning networks). Using the entire sequences lead to better results, indicating that having more longitudinal data was more important for the present task than having more training samples.

Another experience that was performed was to use additional non-sequential and general data about each patient and to give it to our model to see if its performance would benefit from it. For that, the Keras Functional API was used and a regular ANN layer was added and its output concatenated to the already built RNN's output. This was done based on the hypothesis that demographic data or family history, and more general disease-related information would be informative, which was confirmed.

The original idea was to use LSTM networks in this dissertation, as it is the state of the art Recurrent Neural Network with greatest representational power and it has been used in the longitudinal study of other diseases. Nonetheless, two different types of RNN were tested: the standard RNN (with the tanh activation) and the GRU. With the methodologies used, the GRU lead to the best results (although the performance measures were very similar, especially between the LSTM and GRU models). Some reasons could be behind this, for example the small number of time-steps that composed the sequences.

Finally, it was analysed whether the different types of features that were used were giving useful information to the network. For that, the network was tested using different sets of features. Although when isolated from the other features the most informative set was the "DEP" set of features (assessments directly related to depression), the other clinical assessments were of great use in the task of detecting depression.

To conclude, regardless of not having reached perfect values of accuracy, specificity nor sensitivity, we can say that this Deep Learning method applied to this area of research can be extremely useful and that this work is indicative that in future work even better results are very likely to be achieved.

6.2 Future work

One of the limitations of this work is the fact that only two classes were distinguished. This is probably very limiting as the distinguished "positive" class may simultaneously include individuals who have been through a Major Depressive Disorder and others that only experienced some depressive symptoms. Using classes built on more restrict criteria would likely benefit the classification task, as we believe there may be participants with a similar history of depression who in this study belong to different classes (for example, there are records of high depression scores in assessments from the "negative" class). To improve the used method, it could also be tested to use different information about the patient (for example, imaging data).

It would also be interesting to understand which features are not adding useful inputs to the model and could be deleted (the least the number of features, the better in terms of data acquisition and also the lower model complexity).

One of the big questions that remain (which was not the purpose of the developed work) is what is behind this strict relation between HD and depression. Having a better understanding of the temporal patterns, possibly detected by the RNN, could bring great insights regarding this issue: is depression prior to a specific pattern of clinical evolution? Is it a consequence of it? For that, mining algorithms built with the purpose of finding sequential patterns would be interesting to use in our context.

Finally, a very similar approach could be used for predicting if the person will have depression in the future or to predict the stage of the disease or of a specific aspect of it in a future visit.

Bibliography

[1] J. R. Slaughter, M. P. Martens, and K. A. Slaughter, “Depression and Huntington’s Disease: Prevalence, Clinical Manifestations, Etiology, and Treatment,” *CNS Spectrums*, vol. 6, no. 4, pp. 306–308,325–326, apr 2001.

[2] P. Naarding, J. G. Janzing, P. Eling, S. Van Der Werf, and B. Kremer, “Apathy is not depression in Huntington’s disease,” *Journal of Neuropsychiatry and Clinical Neurosciences*, vol. 21, no. 3, pp. 266–270, 2009.

[3] E. A. Epping and J. S. Paulsen, “Depression in the early stages of Huntington disease,” *Neurodegenerative Disease Management*, vol. 1, no. 5, pp. 407–414, oct 2011.

[4] H. D. Schmidt, R. C. Shelton, and R. S. Duman, “Functional Biomarkers of Depression: Diagnosis, Treatment, and Pathophysiology,” *Neuropsychopharmacology*, vol. 36, pp. 2375–2394, 2011.

[5] A. F. Carvalho, M. S. Sharma, A. R. Brunoni, E. Vieta, and G. A. Fava, “The Safety, Tolerability and Risks Associated with the Use of Newer Generation Antidepressant Drugs: A Critical Review of the Literature ,” *Psychother Psychosom*, vol. 85, pp. 270–288, 2016. [Online]. Available: www.karger.com/pps

[6] R. Elliott, B. J. Sahakian, A. P. McKay, J. J. Herrod, T. W. Robbins, and E. S. Paykel, “Neuropsychological impairments in unipolar depression: the influence of perceived failure on subsequent performance,” *Psychological Medicine*, vol. 26, no. 5, pp. 975–989, sep 1996.

[7] G. G. Potter and D. C. Steffens, “Contribution of depression to cognitive impairment and dementia in older adults,” *Neurologist*, vol. 13, no. 3, pp. 105–117, 2007.

[8] J. S. Paulsen, A. B. Rsw, and C. M. Forsyth, *Understanding Behaviour in Huntington Disease: A Guide for Professionals*, 3rd ed., A. Bénard, C. M. Forsyth, and J. Papke, Eds., 2016. [Online]. Available: https://www.huntingtonsociety.ca/wp-content/uploads/2013/10/HSC{_}UnderstandingBehaviour{_}3rdEdition.pdf

[9] J. Tan, “Primed for Psychiatry: The role of artificial intelligence and machine learning in the optimization of depression treatment,” Tech. Rep. 1, 2019.

[10] R. A. Roos, “Huntington’s disease: A clinical review,” *Orphanet Journal of Rare Diseases*, vol. 5, no. 1, pp. 2–9, 2010.

[11] H. H. P. Nguyen and P. Weydt, “Huntington disease,” *Nature Reviews Disease Primers*, vol. 1, no. 1, 2015. [Online]. Available: https://doi.org/10.1038/nrdp.2015.52

[12] F. O. Walker, “Huntington’s disease,” pp. 218–228, jan 2007.

[13] T. Pringsheim, K. Wiltshire, L. Day, J. Dykeman, T. Steeves, and N. Jette, “The incidence and prevalence of Huntington’s disease: A systematic review and meta-analysis,” *Movement Disorders*, vol. 27, no. 9, pp. 1083–1091, aug 2012.

[14] F. Squitieri, S. E. Andrew, Y. P. Goldberg, B. Kremer, N. Spence, J. Zelsler, K. Nichol, J. Theilmann, J. Greenberg, J. Goto, I. Kanazawa, J. Vesa, L. Peltonen, E. Almqvist, M. Anvret, H. Telenius, B. Lin, G. Napolitano, K. Morgan, and M. R. Hayden, “DNA haplotype analysis of huntington disease reveals clues to the origins and mechanisms of CAG expansion and reasons for geographic variations of prevalence,” *Human Molecular Genetics*, vol. 3, no. 12, pp. 2103–2114, dec 1994.

[15] J. S. Paulsen, D. R. Langbehn, J. C. Stout, E. Aylward, C. A. Ross, M. Nance, M. Guttman, S. Johnson, M. MacDonald, L. J. Beglinger, K. Duff, E. Kayson, K. Biglan, I. Shoulson, D. Oakes, and M. Hayden, “Detection of Huntington’s disease decades before diagnosis: The Predict-HD study,” *Journal of Neurology, Neurosurgery and Psychiatry*, vol. 79, no. 8, pp. 874–880, 2008.

[16] O. Quarrell, K. L. O’Donovan, O. Bandmann, and M. Strong, “The Prevalence of Juvenile Huntington’s Disease: A Review of the Literature and Meta-Analysis,” *PLoS Currents*, vol. 4, p. e4f8606b742ef3, jul 2012.

[17] A.-W. Heemskerk and R. A. C. Roos, “E04 Causes of death in Huntington’s disease,” *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 81, sep 2010.

[18] D. R. Langbehn, M. Hayden, and J. S. Paulsen, “CAG-Repeat Length and the Age of Onset in Huntington Disease (HD): A Review and Validation Study of Statistical Approaches.”

[19] O. W. J. Quarrell, M. A. Nance, P. Nopoulos, J. S. Paulsen, J. A. Smith, and F. Squitieri, “Managing juvenile Huntington’s disease,” *Neurodegener Dis Manag*, vol. 3, no. 3, 2013.

[20] E. J. Wild and S. J. Tabrizi, “Huntington’s disease phenocopy syndromes,” *Current Opinion in Neurology*, vol. 20, no. 6, pp. 681–687, dec 2007.

[21] “HTT gene - Genetics Home Reference - NIH,” 2008. [Online]. Available: https://ghr.nlm.nih.gov/gene/HTT{#}conditions

[22] J. Schulte and J. T. Littleton, “The biological function of the Huntingtin protein and its relevance to Huntington’s Disease pathology,” *Current trends in neurology*, vol. 5, pp. 65–78, 2011.

[23] M. N. W. Witjes-Ané, M. Vegter-van der Vlis, J. P. Van Vugt, J. B. Lanser, J. Hermans, A. H. Zwinderman, G. J. B. Van Ommen, and R. A. Roos, “Cognitive and motor functioning in gene carriers for Huntington’s disease: A baseline study,” *Journal of Neuropsychiatry and Clinical Neurosciences*, vol. 15, no. 1, pp. 7–16, 2003.

[24] Â. Miranda, R. Lavrador, F. Júlio, C. Januário, M. Castelo-Branco, and G. Caetano, “Classification of Huntington’s disease stage with support vector machines: A study on oculomotor performance,” *Behavior Research Methods*, vol. 48, no. 4, pp. 1667–1677, 2016. [Online]. Available: http://dx.doi.org/10.3758/s13428-015-0683-z

[25] R. D. Kent, G. Weismer, J. F. Kent, and J. C. Rosenbek, “Toward phonetic intelligibility testing in dysarthria,” *Journal of Speech and Hearing Disorders*, vol. 54, no. 4, pp. 482–499, 1989.

[26] A. A. Guccione, R. Wong, and D. Avers, *Geriatric Physical Therapy*, 3rd ed. Mosby, 2011.

[27] J. B. Penney, A. B. Young, I. Shoulson, S. Starosta-Rubenstein, S. R. Snodgrass, J. Sanchez-Ramos, M. Ramos-Arroyo, F. Gomez, G. Penchaszadeh, J. Alvir, J. Esteves, I. DeQuiroz, N. Marsol, H. Moreno, P. M. Conneally, E. Bonilla, and N. S. Wexler, “Huntington’s disease in venezuela: 7 years of follow-up on symptomatic and asymptomatic individuals,” *Movement Disorders*, vol. 5, no. 2, pp. 93–99, 1990.

[28] J. C. Stout, J. S. Paulsen, S. Queller, A. C. Solomon, K. B. Whitlock, J. C. Campbell, N. Carlozzi, K. Duff, L. J. Beglinger, D. R. Langbehn, S. A. Johnson, K. M. Biglan, and E. H. Aylward, “Neurocognitive Signs in Prodromal Huntington Disease,” *Neuropsychology*, vol. 25, no. 1, pp. 1–14, 2011.

[29] E. Aretouli and J. Brandt, “Episodic Memory in Dementia: Characteristics of New Learning that Differentiate Alzheimer’s, Huntington’s, and Parkinson’s Diseases.”

[30] M. Papoutsi, I. Labuschagne, S. J. Tabrizi, and J. C. Stout, “The cognitive burden in Huntington’s disease: Pathology, phenotype, and mechanisms of compensation,” pp. 673–683, apr 2014.

[31] K. Duff, J. S. Paulsen, L. J. Beglinger, D. R. Langbehn, C. Wang, M. C. Julie Stout, C. A. Ross, E. Aylward, N. E. Carlozzi, and S. Queller, “"Frontal" Behaviors Before the Diagnosis of Huntington’s Disease and Their Relationship to Markers of Disease Progression: Evidence of Early Lack of Awareness,” *Journal of Neuropsychiatry and Clinical Neurosciences*, vol. 22, no. 2, pp. 196–207, 2010.

[32] V. L. Wheelock, T. Tempkin, K. Marder, M. Nance, and R. H. Myers, “Predictors of nursing home placement in Huntington disease,” *Neurology*, vol. 60, no. 6, pp. 998–1001, 2003.

[33] M. H. Abbott, G. A. Chase, B. A. Jensen, and M. F. Folstein, “The Association of Affective Disorder with Huntington’s Disease in a Case Series and in Families,” *Psychological Medicine*, vol. 13, no. 3, pp. 537–542, 1983.

[34] K. Duff, J. S. Paulsen, L. J. Beglinger, D. R. Langbehn, and J. C. Stout, “Psychiatric Symptoms in Huntington’s Disease before Diagnosis: The Predict-HD Study,” *Biological Psychiatry*, vol. 62, no. 12, pp. 1341–1346, dec 2007.

[35] G. J. Gilbert, “Weight loss in Huntington disease increases with higher CAG repeat number,” *Neurology*, vol. 73, no. 7, p. 572, 2009.

[36] J. S. Paulsen, R. E. Ready, J. M. Hamilton, and J. L. Cummings, “Neuropsychiatric aspects of Huntington’s disease,” *J Neurol Neurosurg Psychiatry*, vol. 71, pp. 310–314, 2001.

[37] G. Huntington, “On chorea. George Huntington, M.D.” pp. 109–112, 2003.

[38] J. S. Paulsen, C. Nehl, B. Karin Ferneyhough Hoth, M. E. Jason Kanz, M. Michelle Benjamin, B. Rachel Conybeare, B. Bradley McDowell, and B. Turner, “Depression and Stages of Huntington’s Disease,” *The Journal of Neuropsychiatry and Clinical Neurosciences*, vol. 17, pp. 496–502, 2005.

[39] N. P. Rocha, B. Mwangi, C. A. Candano, C. Sampaio, E. F. Stimming, and A. L. Teixeira, “The clinical picture of psychosis in manifest Huntington’s disease: A comprehensive analysis of the enroll-HD Database,” *Frontiers in Neurology*, vol. 9, no. NOV, pp. 1–11, 2018.

[40] E. Van Duijn, E. M. Kingma, and R. C. Van Der Mast, “Psychopathology in verified Huntington’s disease gene carriers,” *Journal of Neuropsychiatry and Clinical Neurosciences*, vol. 19, no. 4, pp. 441–448, 2007.

[41] Huntington Study Group, “Unified Huntington’s Disease Rating Scale: Reliability and Consistency,” *Movement Disorders*, vol. 11, no. 2, pp. 136–142, 1996.

[42] S. Siesling, J. P. P. van Vugt, K. A. H. Zwinderman, K. Kieburtz, and R. A. C. Roos, “Unified Huntington’s disease rating scale: A follow up,” *Movement Disorders*, vol. 13, no. 6, pp. 915–919, nov 1998. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/mds.870130609

[43] World Health Organization, “Depression,” 2019. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/depression

[44] M. C. Angermeyer, A. Holzinger, H. Matschinger, and K. Stengler-Wenzke, “Depression and quality of life: Results of a follow-up study,” *International Journal of Social Psychiatry*, vol. 48, no. 3, pp. 189–199, 2002.

[45] S. Bachmann, “Epidemiology of suicide and the psychiatric perspective,” *International Journal of Environmental Research and Public Health*, vol. 15, no. 7, jul 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6068947/pdf/ijerph-15-01425.pdf

[46] K. S. Dobson and D. J. Dozois, “Chapter 1 - introduction: Assessing risk and resilience factors in models of depression,” in *Risk Factors in Depression*, K. S. Dobson and D. J. Dozois, Eds. San Diego: Elsevier, 2008, pp. 1 – 16. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780080450780000010

[47] M. J. Patel, A. Khalaf, and H. J. Aizenstein, “Studying depression using imaging and machine learning methods,” *NeuroImage: Clinical*, vol. 10, pp. 115–123, 2016. [Online]. Available: http://dx.doi.org/10.1016/j.nicl.2015.11.003

[48] W. H. Farah, M. Alsawas, M. Mainou, F. Alahdab, M. H. Farah, A. T. Ahmed, E. A. Mohamed, J. Almasri, M. R. Gionfriddo, A. Castaneda-Guarderas, K. Mohammed, Z. Wang, N. Asi, C. N. Sawchuk, M. D. Williams, L. J. Prokop, M. H. Murad, and A. Leblanc, "Non-pharmacological treatment of depression: A systematic review and evidence map," pp. 214–221, dec 2016.

[49] J. LeMoult and I. H. Gotlib, "Depression: A cognitive perspective," *Clinical Psychology Review*, vol. 69, pp. 51–66, apr 2019.

[50] M. Pandya, M. Altinay, D. A. Malone, and A. Anand, "Where in the brain is depression?" *Current Psychiatry Reports*, vol. 14, no. 6, pp. 634–642, 2012.

[51] V. Lorenzetti, N. B. Allen, A. Fornito, and M. Yücel, "Structural brain abnormalities in major depressive disorder: A selective review of recent MRI studies," pp. 1–17, sep 2009.

[52] T. A. Kimbrell, T. A. Ketter, M. S. George, J. T. Little, B. E. Benson, M. W. Willis, P. Herscovitch, and R. M. Post, "Regional cerebral glucose utilization in patients with a range of severities of unipolar depression," *Biological Psychiatry*, vol. 51, no. 3, pp. 237–252, feb 2002.

[53] K. Helm, K. Viol, T. M. Weiger, P. A. Tass, C. Grefkes, D. Del Monte, and G. Schiepek, "Neuropsychiatric Disease and Treatment Dovepress Neuronal connectivity in major depressive disorder: a systematic review," *Neuropsychiatric Disease and Treatment*, 2018. [Online]. Available: http://dx.doi.org/10.2147/NDT.S170989

[54] L. Chen, Y. Wang, C. Niu, S. Zhong, H. Hu, P. Chen, S. Zhang, G. Chen, F. Deng, S. Lai, J. Wang, L. Huang, and R. Huang, "Common and distinct abnormal frontal-limbic system structural and functional patterns in patients with major depression and bipolar disorder," 2018. [Online]. Available: https://doi.org/10.1016/j.nicl.2018.07.002

[55] M. P. Luber, J. P. Hollenberg, P. Williams-Russo, T. N. Didomenico, B. S. Meyers, G. S. Alexopoulos, and M. E. Charlson, "Diagnosis, treatment, comorbidity, and resource utilization of depressed patients in a general medical practice," *International Journal of Psychiatry in Medicine*, vol. 30, no. 1, pp. 1–13, 2000.

[56] A. K. Ho, A. S. Gilbert, S. L. Mason, A. O. Goodman, and R. A. Barker, "Health-related quality of life in Huntington's disease: Which factors matter most?" *Movement Disorders*, vol. 24, no. 4, pp. 574–578, mar 2009.

[57] D. Craufurd, J. C. Thompson, and J. S. Snowden, "Behavioral Changes in Huntington Disease," *Neuropsychiatry, Neuropsychology, and Behavioral Neurology*, vol. 14, no. 4, pp. 219–226, 2001. [Online]. Available: https://doi.org/10.1176/jnp.14.1.37

[58] C. Nehl, R. E. Ready, J. Hamilton, and J. S. Paulsen, "Effects of depression on working memory in presymptomatic Huntington's disease," *Journal of Neuropsychiatry and Clinical Neurosciences*, vol. 13, no. 3, pp. 342–346, 2001.

[59] C. E. Peyser and S. E. Folstein, “Huntington’s Disease as a Model for Mood Disorders Clues from Neuropathology and Neurochemistry,” Tech. Rep., 1990.

[60] N. Ketkar, *Deep Learning with Python*, 2017.

[61] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[62] T. G. Dietterich, *Machine learning*, 1997.

[63] J. Hopfield, “Artificial neural networks,” *IEEE Circuits and Devices Magazine*, vol. 4, no. 5, pp. 3–10, 1988.

[64] L. Fu, “Knowledge Discovery Based on Neural Networks,” *Communications of the ACM*, vol. 42, no. 11, pp. 47–50, 1999.

[65] S. Shanmuganathan, “Artificial Neural Network Modelling: An Introduction,” in *Artificial Neural Network Modelling*. Springer, Cham, 2016, no. July, pp. 1–14.

[66] S. Ruder, “An overview of gradient descent optimization algorithms,” Insight Centre for Data Analytics, NUI Galway, Dublin, Tech. Rep., 2017.

[67] N. Qian, “On the Momentum Term in Gradient Descent Learning Algorithms,” *Neural Networks*, pp. 145–151, 1999.

[68] D. P. Kingma and J. Lei Ba, “Adam: A method for stochastic Optimization,” in *International Conference on Learning Representations*, 2015, pp. 1–13.

[69] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, 1986. [Online]. Available: https://doi.org/10.1038/323533a0

[70] I. Flood, “A Gaussian-Based Feedforward Network Architecture and Complementary Training Algorithm.”

[71] R. Pascanu, T. Mikolov, and Y. Bengio, “Understanding the exploding gradient problem,” Université de Montréal, Tech. Rep., 2012.

[72] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*, 1st ed., 2012.

[73] R. J. Williams and D. Zipser, “Gradient-Based Learning Algorithms for Recurrent Networks and Their Computational Complexity,” pp. 433–486.

[74] Y. Bengio, P. Simard, and P. Frasconi, “Learning Long-Term Dependencies with Gradient Descent is difficult,” pp. 157–166, 1994.

[75] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Computation*, vol. 9, no. 8, November 15, 1997, pp. 1735–1780, 1997. [Online]. Available: http://www.bioinf.jku.at/publications/older/2604.pdf

[76] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, “LSTM: A Search Space Odyssey,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 10, pp. 2222–2232, 2017.

[77] T. Wang, R. G. Qiu, and M. Yu, “Predictive Modeling of the Progression of Alzheimer’s Disease with Recurrent Neural Networks,” *Scientific Reports*, vol. 8, no. 1, pp. 1–12, 2018. [Online]. Available: http://dx.doi.org/10.1038/s41598-018-27337-w

[78] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using RNN encoder-decoder for statistical machine translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014, pp. 1724–1734. [Online]. Available: https://www.aclweb.org/anthology/D14-1179

[79] J. Chung, C. Gulcehre, and K. Cho, “Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling,” in *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014. [Online]. Available: http://arxiv.org/abs/1412.3555

[80] M. Nguyen, “Illustrated Guide to LSTM’s and GRU’s: A step by step explanation,” 2018. [Online]. Available: https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

[81] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Gated Feedback Recurrent Neural Networks,” in *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, 2015. [Online]. Available: https://arxiv.org/abs/1502.02367

[82] A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional LSTM and other neural network architectures,” *Neural Networks*, vol. 18, no. 5-6, pp. 602–610, 2005. [Online]. Available: https://doi.org/10.1016/j.neunet.2005.06.042

[83] A. Graves, “Generating Sequences With Recurrent Neural Networks,” 2013. [Online]. Available: http://arxiv.org/abs/1308.0850

[84] R. Xu, D. C. Wunsch, and R. L. Frank, “Inference of genetic regulatory networks with recurrent neural network models using particle swarm optimization,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 4, no. 4, pp. 681–692, 2007.

[85] D. Eck and J. Schmidhuber, “A First Look at Music Composition using LSTM Recurrent Neural Networks,” IDSIA, Manno, Switzerland, Tech. Rep., 2002.

[86] Z. C. Lipton, D. C. Kale, C. Elkan, R. P. Wetzel Laura, and L. K. Whittier Virtual PICU, “LEARNING TO DIAGNOSE WITH LSTM RECURRENT NEURAL NETWORKS,” in *iclr 2016*, 2016.

[87] G. Lee, K. Nho, B. Kang, K. A. Sohn, and D. Kim, “Predicting Alzheimer’s disease progression using multi-modal deep learning approach,” *Scientific Reports*, vol. 9, no. 1, pp. 1–12, 2019.

[88] A. Nahon and B. Lerner, “Temporal modeling of ALS using longitudinal data and long-short term memory-based algorithm,” no. April, pp. 25–27, 2018.

[89] “Python,” https://www.python.org.

[90] “Jupyter,” https://www.jupyter.org.

[91] “Matplotlib,” https://www.matplotlib.org.

[92] “Seaborn,” https://www.seaborn.pydata.org.

[93] “Data dictionary of enroll-hd - periodic dataset,” https://www.enroll-hd.org/enrollhd_documents/2018-10-R1/Enroll-HD-DataDictionary-2018-10-R1.pdf.

[94] “REGISTRY Study Protocol Version 3.0 Replacing Version 2.0 REGISTRY-an observational study of the European Huntington-Disease Network (EHDN),” Tech. Rep., 2009. [Online]. Available: https://www.enroll-hd.org/enrollhd{_}documents/2016-10-R1/registry-protocol-3.0.pdf

[95] “Enroll-hd: A prospective registry study in a global huntington’s disease cohort. clinical study protocol version 1,” https://www.enroll-hd.org/enrollhd_documents/Enroll-HD-Protocol-1.0.pdf.

[96] S. Smith, N. Butters, R. White, L. Lyon, and E. Granholm, “Priming semantic relations in patients with Huntington’s Disease,” *Brain and Language*, vol. 33, no. 1, pp. 27–40, 1988.

[97] L. K. Sheridan, H. E. Fitzgerald, K. M. Adams, J. T. Nigg, M. M. Martel, L. I. Puttler, M. M. Wong, and R. A. Zucker, “Normative Symbol Digit Modalities Test performance in a community-based sample,” *Archives of Clinical Neuropsychology*, vol. 21, pp. 23–28, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0887617705001174

[98] M. Noll-Hussong, F. Burgio, Y. Ikeda, F. Scarpina, and S. Tagini, “The Stroop Color and Word Test,” *The Stroop Color and Word Test Front. Psychol*, vol. 8, p. 557, 2017.

[99] I. Santana, D. Duro, R. Lemos, V. Costa, M. Pereira, M. R. Simões, and S. Freitas, “Mini-mental state examination: Avaliação dos novos dados normativos no rastreio e diagnóstico do défice cognitivo,” *Acta Medica Portuguesa*, vol. 29, no. 4, pp. 240–248, 2016.

[100] J. Callaghan, C. Stopford, N. Arran, M.-F. Boisse, A. Coleman, R. Dar Santos, E. M. Dumas, E. P. Hart, D. Justo, G. Owen, J. Read, M. J. Say, A. Durr, B. R. Leavitt, R. A. C Roos, S. J. Tabrizi, A.-C. Bachoud-Levi, C. Bourdet, E. van Duijn, and D. Craufurd, “Reliability and Factor Structure of the Short Problem Behaviors Assessment for Huntington’s Disease (PBA-s) in the TRACK-HD and REGISTRY studies,” *Journal of Neuropsychiatry and Clinical Neurosciences*, vol. 27, no. 1, pp. 59–64, 2015.

[101] J. E. Ware, M. Kosinsky, and S. D. Keller, “A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity,” *Medical Care*, vol. 34, no. 3, pp. 220–233, 1996.

[102] K. Posner, G. K. Brown, B. Stanley, D. A. Brent, K. V. Yershova, M. A. Oquendo, G. W. Currier, G. A. Melvin, L. Greenhill, S. Shen, and J. J. Mann, "The Columbia-Suicide Severity Rating Scale: Initial Validity and Internal Consistency Findings From Three Multisite Studies With Adolescents and Adults," *Am J Psychiatry*, vol. 168, no. 12, pp. 1266–1277, 2011. [Online]. Available: www.cssrs.columbia.edu

[103] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Muller, "Efficient BackProp," Tech. Rep., 1998.

[104] Z. K. B. Shalabi, Luai Al Shaaban, "Data Mining : A Preprocessing Engine," *Journal of Computer Science*, 2006.

[105] Y. Liu, "Encoding Categorical Features - Towards Data Science," 2018. [Online]. Available: https://towardsdatascience.com/encoding-categorical-features-21a2651a065c

[106] D. M. Reddy and S. Reddy, "Effects of padding on LSTMs and CNNs," Manipal Institute of Technology, Tech. Rep., 2019.

[107] "Core Layers - Keras Documentation." [Online]. Available: https://keras.io/layers/core/

[108] "Sequential - Keras Documentation." [Online]. Available: https://keras.io/models/sequential/

[109] "Guide to the Sequential model - Keras Documentation." [Online]. Available: https://keras.io/getting-started/sequential-model-guide/

[110] "Model (functional API) - Keras Documentation." [Online]. Available: https://keras.io/models/model/

[111] "Guide to the Functional API - Keras Documentation." [Online]. Available: https://keras.io/getting-started/functional-api-guide/

[112] N. Srivastava, G. Hinton, A. Krizhevsky, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, 2014.

[113] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent Neural Network Regularization," 2014. [Online]. Available: http://arxiv.org/abs/1409.2329

[114] T. A. A. Mohamed Bekkar, Hassiba Kheliouane Djemaa, "Evaluation Measures for Models Assessment over Imbalanced Data Sets," *Journal of Information Engineering and Applications*, 2013.

[115] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets," *PLOS ONE*, 2015.