RNNs for detecting depression in Huntington's Disease

Mariana Dias
mariana.g.dias@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

February 2020

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, which, unlike other symptoms of the disease, is not correlated to disease progression, but has been linked to greater functional damage and worse cognitive performance, while having 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. Longitudinal data of 9474 HD patients and 1481 controls from the *Enroll-HD* database was used, comprising information from annual clinical visits where several questionnaires are answered and exams are performed, regarding the evaluation of all clinical aspects of HD. Several Recurrent Neural Network architectures were tested and it was 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.

1. Introduction

HD is a rare neurodegenerative disease for which there exists no cure. The evolution of the disease is extremely heterogeneous, being characterized by a progressive course of a combination of motor, cognitive and behavioural impairments, which lead to an increasing dependency in daily life, resulting in patients requiring full-time care, and finally death [1], [2]. The onset of the symptoms usually occurs between the ages of 30 to 50 [1].

The most frequently occurring psychiatric sign is depression but no relation to disease progress has been evidenced [3], [4]. 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 [5]. Suicide is estimated to be the cause of 5-10% of the deaths in HD [6].

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 [7]. 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 [8]. Like HD, it is characterized by a heterogeneous symptomatology. Hence, the clinical treatment approach is usually a trial and error approach, which is extremely unadvantageous as antidepressants may have adverse secondary effects [9]. Moreover, depression has been linked to cognitive decline [10], [11].

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 [12], [13]. 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 [14].

While most studies regarding this issue focus on statistically associating specific phases of the disease and/or specific symptoms and signs to depression, in this work we aim to use Deep Learning (more concretely, Recurrent Neural Networks) for processing longitudinal clinical data to detect cases where depression (with or without a formally-stated diagnosis) has been a part of the medical history.

2. Background

2.1. Huntington's Disease and depression

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 (HTT) gene (gene IT-15) [1]. CAG is a trinucleotide that codes for the amino acid glutamine. Although the exact function of the HTT protein remains uncompletely understood, it appears to have an important role in the neurons functioning, being found in all neurons of the brain, as well as glial cells [15]. The presence of the expanded CAG segment leads to the production of an abnormally long version of the huntingtin protein [2], 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 [12].

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 [16], [17]. Common cognitive decline features in HD include deficits in attention, emotion recognition, visuomotor processing, decreased ability to learn and retrieve new information [17], [18], [19]. These cognitive impairments with simultaneous psychiatric problems result many times in lack of initiative, social disengagement, impulsivity and lack of awareness [20].

Psychiatric symptoms, particularly depression, can develop during the prodromal stage or when the disease is manifest [21]. Apathy is correlated to disease progress (cognitive deterioration and functional decline), whereas anxiety and depression are not [4].

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 [3], [22], it could be related to the disease associated alterations in the neurotransmitters in the brain that regulate mood [12] 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 [23].

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 [24].

2.2. Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a family of neural networks for processing sequential data [25], as they include feedback connections among hidden units, associated with a time delay [26]. The key point is that the recurrent connections allow a "memory" of previous inputs to persist in the network's internal state. RNNs, like other neural networks, are 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 input and the output from the previous timestep [27], as represented in figure 1.

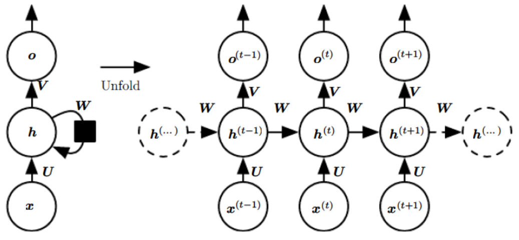


Figure 1: Schematic representation of an RNN [25].

The problem with RNNs is that the inputs cycle around the network's recurrent connections, leading to the problem of learning long-term dependencies: gradients propagated over many timesteps tend to vanish (or, more rarely, explode) [28].

The Long Short-Term Memory (LSTM) method was first introduced by Hochreiter and Schmidhuber in 1997 [29], with the aim of addressing the problem of long-term dependencies. 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 (represented in figure 2) which can maintain its state over time, and non-linear gating units which regulate the information flow into and out of the cell [30]: the input, output and forget gates.

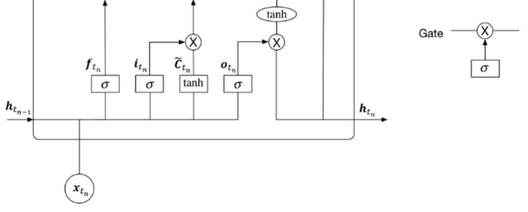


Figure 2: Schematic representation of an LSTM cell [31].

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 (from expression 1) 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} (from expression 4). The input gate (2) "decides" which values will be updated; the new values to add to the cell state are computed using expression 3. Finally, in the output gate (5) it is decided what part of the current cell state is going to be output (6).

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

$$i_t = \sigma(\ W_i \cdot [\ h_{t-1},\ x_t] +\ b_i) \tag{2}$$

$$\tilde{C_t} = tanh(\ W_C \cdot [\ h_{t-1},\ x_t] +\ b_C) \tag{3}$$

$$C_t =\ f_t *\ C_{t-1} +\ i_t *\ \tilde{C_t} \tag{4}$$

$$o_t = \sigma(\ W_o \cdot [\ h_{t-1},\ x_t] + b_o) \tag{5}$$

$$h_t = o_t * tanh(C_t) \tag{6}$$

Another type of RNN with a special gated architecture is the Gated Recurrent Unit (GRU) (illustrated in figure 3), which was developed in 2014 [32]. It can be thought of as a modification of the LSTM with a less complex architecture [33], once it uses two gates (the update and reset gates) instead of three and it doesn't use the cell state to transfer information, but rather the hidden state [34].

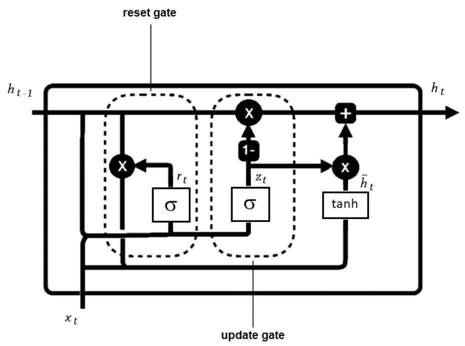


Figure 3: Schematic representation of a GRU (adapted from [35]).

The update gate, 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 7. The reset gate, r_t, allows the unit to forget the previous hidden states and it is computed as in expression 8. The candidate hidden state ($\tilde{h}_t$) is done similarly to that of the the LSTM (expression 9) [36]. 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 ($\tilde{h}_t$) [34].

$$z_t = \sigma(\ W_z \cdot [\ h_{t-1},\ x_t] + b_z) \tag{7}$$

$$r_t = \sigma(\ W_r \cdot [\ h_{t-1},\ x_t] + b_r) \tag{8}$$

$$\tilde{h}_t = tanh(\ W \cdot [\ h_{t-1} * r_t,\ x_t] + b_h) \tag{9}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \tag{10}$$

RNNs 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 [37], [38]), genomic analysis [39] and music generation [40].

GRUs have also been used for predicting Alzheimer's disease progression [41] and LSTMs have been used in the same task but in the context of Amyotrophic Lateral Sclerosis [42].

3. Methods

The present work was executed using Python language and the networks were developed using keras with Tensorflow backend.

3.1. 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 [43] and also *Adhoc* data . 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 [44]. Study procedures include annual assessments conducted during study visits and performed by trained clinical personnel [45].

Subjects from the Enroll-HD database 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. For the present study, from the whole dataset, those who didn't have any information regarding if depression had been part of their medical history were excluded, just like those only attended one visit (since this study is intended to be based on longitudinal data); also, the entries where the age at the moment of the visit wasn't specified were deleted. Table 1 describes the number of participants of the present study.

Two different types of data were used: sequential data (from the visits) and static data (the "profile" data).

	N	% female	#visits
HD	9474	46%	4.44 ± 2.57
controls	1481	40%	3.47 ± 1.48

Table 1: Final number of participants of each group and correspondent percentage of female subjects and mean number of visits per participant.

The visits' data comprise several items that correspond to the assessments made during each annual visit and include results from clinical examinations (such as cognitive, motor or psychiatric tests) and questions answered by the participant (regarding, for example, the activities of daily life, medical history, drug use, suicidal tendencies). These items come from a list of forms, which includes, for example, the *UHDRS* motor and functional sections, the *Cognitive Assessments* (result scores from 3 cognitive tests: Stroop Color and Word Reading Test [46], Symbol-Digit Modality Test [47] and Categorical Verbal Fluency Test [48]), the *Problem Behaviours Assessment - Short (PBA-s)* (frequency and severity of symptoms related to altered emotions, thought content and coping strategies [49]), the *Short Form Health Survey-12* (overall health status [50]) and the *Columbia Suicide Severity Rating Scale (C-SSRS)* (aimed to monitor suicidal events [51]). Some of these assessments are directly related to the diagnosis of depression and/or depressive behaviour (such as depressed mood scores, presence and frequency of suicidal thoughts): these items will be referred to as ”depression features” or ”DEP”.

The profile data contains general non-temporal 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, age at onset of HD. The binary variable ”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 from this data file.

3.2. Data Pre-Processing

As different features range between different values, it was necessary to perform standardization in order to set a common a scale [52]. Each column was separately standardized using the formula presented in expression 11 (z-normalization), 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 [53]. After applying this transformation, each feature has a mean of 0 and a standard deviation of 1.

$$z_i = \frac{x_i - \overline{x}}{\sigma} \tag{11}$$

Once the data described in the previous section includes both numerical and categorical features (although all were provided integer encoded, there often is no ordinal relationship between the values that the variable assumes), another standardization method for the categorical items was tested: one-hot encoding. This consists of converting each categorical feature in n binary features, with n being the number of possible categories [54]. Both methods were used and compared because the one-hot encoding approach implies increased computational complexity, as the feature dimensionality expands considerably.

Missing values represent 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”. 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”.

Finally, in order to use the visits' data 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.

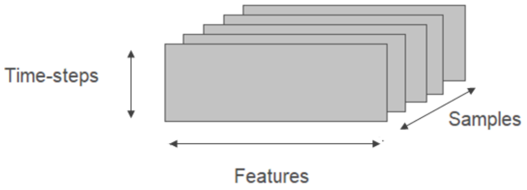


Figure 4: Representation of the visits' data as a 3D tensor.

Two different methods were used, in order to understand which worked better: (1) to correspond each sample to one participant; (2) to correspond each set of 3 time-steps to a sample (*i.e.*, from each participant, $n - 2$ samples are originated, with n representing the number of visits attended: for example, a participant who attended to 4 visits originates 2 samples: one corresponding to the first to the third visits and the other from the second to the fourth). As the number of visits attended per participant varied largely, in the first approach, each sample has 15 time-steps and is pre-padded with 0's. The second approach has the advantage of originating a larger number of samples and of not requiring long padding sequences.

3.3. Deep Learning Model

All developed models include an RNN with a "many-to-one" structure (as illustrated in figure 5), which receives as input the 3D matrix described in the previous section. The first layer is a masking layer for masking the time-steps filled with "0's" and the last layer is a Dense layer: a feed-forward layer, which outputs the computation of an activation function to its inputs.

The first architecture consists of a stack of LSTM layers, for processing the visit's data only.

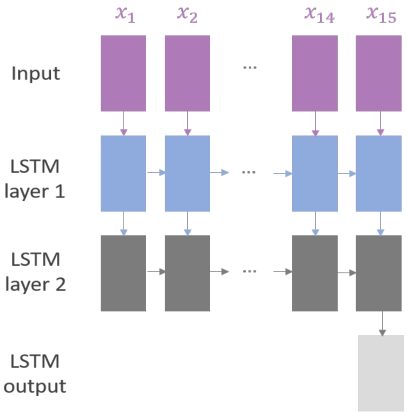


Figure 5: LSTM network "many-to-one" architecture.

In order to add the profile data, as this is a different type of data, a multi-input model was developed. For that, keras Functional API was used [55]. The single-input LSTM was developed using keras Sequential model [56]. Figure 6 illustrates the difference between the two architectures.

The purpose of the multi-input model is to process the different types of data accordingly. The sequential data is processed by an RNN and the profile data passes through a Dense layer with a linear activation, whose output will be concatenated with the RNN's output, as illustrated in figure 6(b).

GRUs and standard RNNs results were compared to those obtained with LSTMs.

3.4. Training and Testing the Deep Learning models

The data was randomly divided into training and testing sets, in a proportion of 80 to 20 percent, respectively. 20 % of the training data was 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.

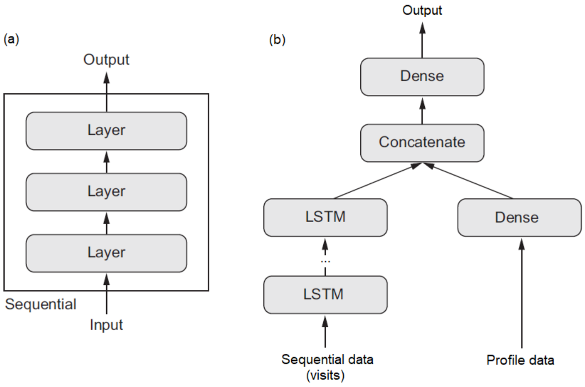


Figure 6: Schematic representations of the keras (a) sequential and (b) funcional API models (adapted from [33]).

In order 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 added 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, 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). For each number of layers, the model was trained with different sets of nodes and the one that lead to the lowest validation loss value would be selected.

Since we are predicting a binary outcome, the used loss function was the binary cross entropy (expression 13) and the activation function of the last layer of the neural network was the sigmoid function (as, from expression 12, 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) [33]. The used optimizer was the Adam optimizer [57].

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{12}$$

$$L(y,\hat{y}) = \\ -\frac{1}{N} * \sum_{i=1}^{N} y_i \cdot log(\hat{y}_i) + (1-y_i) \cdot log(1-\hat{y}_i) \tag{13}$$

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

The data being used to develop the classifier is imbalanced (see table 2), hence, it was necessary to deal with this problem in order to attenuate it. Accordingly, 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.

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

Table 2: Classes representativity in each group of participants.

Another regularization technique used was to add dropout to the network. Complex models, ie, 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 [58]. It is important to mention that the dropout was applied to the input connections (non-recurrent connections), as standard dropout does not work well on RNNs [59].

3.5. Performance Evaluation

In binary classification, a nomenclature we can use to distinguish the outputs is between *Positive* and *Negative*. In this study, the positives represent those with a medical history of depression (or ”1’s”).

The first metric used for evaluating the developed model was *Accuracy*, the most common metric for classifier evaluation, which assesses the overall effectiveness of the algorithm by computing the probability of the correct prediction [60], using the formula presented in expression 14 (where TP stands for True Positive, FP is False Positive, TN is True Negative and FN is False Negative) [61].

Since we are evaluating a binary classifier on an imbalanced dataset, *accuracy* alone can be misleading [60]. Hence, 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 15) whereas the TNR or specificity is the rate of participants belonging to the negative class who were correctly predicted as negative (16). Balanced accuracy is the mean value of these two metrics (17), 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.

$$\textbf{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

$$\textbf{TPR} = \frac{TP}{TP + FN} \tag{15}$$

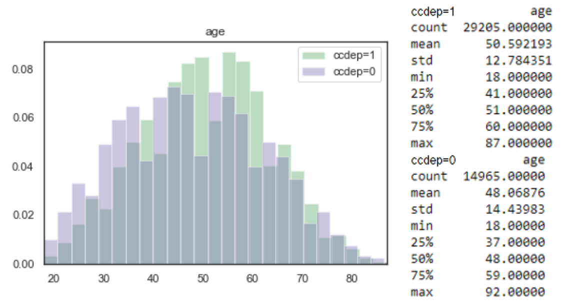
$$\textbf{TNR} = \frac{TN}{TN + FP} \tag{16}$$

$$\textbf{Balanced Accuracy} = \frac{TPR + TNR}{2} \tag{17}$$

The different types of the described RNN architectures were optimized using only the data regarding HD participants and the network architecture which gave the best results was, afterwards, also trained and tested using the whole dataset (HD + controls) and using only the control dataset. Finally, 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. In the following section, preceding the networks results, some of the features’ distributions are presented.

4. Results

4.1. Feature Analysis



```
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 7: Density distributions of the participants’ ages from the two classes.

HD is a progressive disease and, as such, the symptoms and impairments get worse with time. The distribution of the patients’ age from the 2 classes is shown in figure 7: the ages range between roughly the same values 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 (this is important for the stated reason that the age influences the deterioration state and some of the found differences in the other variables distributions could be due to this).

From the distributions of the depression scores (figure 8(a),(b)), where higher score values indicate higher severity of depressive symptoms, it is

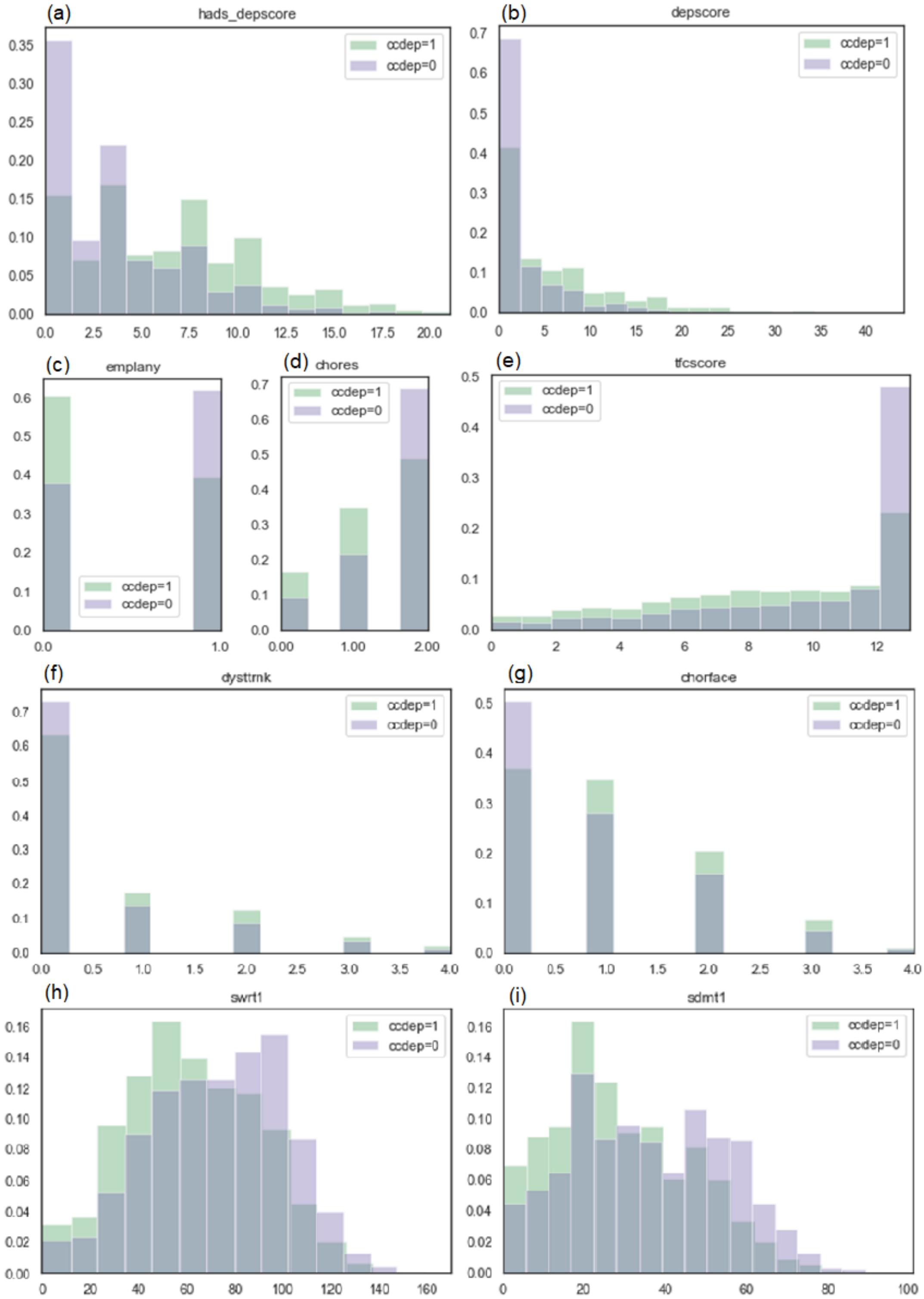


Figure 8: Density distributions of variables from the dataset, distinguishing the two classes *ccdep=1* (green) and *ccdep=0* (violet). (a) and (b) show the distributions of two depression scores, from the HADS-SIS and PBA-s forms, respectively. (c), (d) and (e) regard functional assessments: *emplany* is the binary variable wich answers the question ”Could subject engage in any kind of gainful employment?”, *chores* is the categorical variable regarding the capability to do the domestic chores (0- unable; 1- impaired; 2- normal) and *tfcscore* Total Functional Capacity Score (from the UHDRS). *Dysttrnk* (f) and *chorface* (g) are motor assessments regarding the trunk dystonia and facial choreatic movements, respectively (0- absent; 1- slight intermittent; 2- mild common or moderate intermittent; 3- moderate common; 4- marked prolonged). *swrt* (h) and *sdmt* (i) are the scores (total number of correct answers) obtained in the stroop word reading and symbol-digit modality tests.

possible to observe that there are many samples from *ccdep=1* with very low values of depression scores. First, it is important to retain that these values come from sequential data 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, explaining what was observed. On the other hand, there's also 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).

Figures 8 (c), (d) and (e) illustrate the distributions of three functional assessments, corroborating the idea that depression is associated to greater functional impairment, as there is a clear "shift" of the distributions of the positive class to worse functional performances, comparing to those of the negative class. Regarding the motor assessments, the inter-class variation in the distribution plots is not so evident, as shown in figures 8 (f) and (g). Nonetheless, there is a higher percentage of assessments where there's absent dystonia or facial choreatic movements in the negative class than in the other. Figures 8 (h) and (i) evidence that among those who have had depression there is a higher tendency to perform worse in cognition evaluation exams, such as the stroop word reading test and the symbol-digit modality test.

4.2. Deep Learning results using LSTM

Figure 9 shows the learning and validation loss curves from training an LSTM Sequential model with 3 layers, each curve from training it with a different set of nodes.

From the curves shown in figure 9, it is possible to see that the network with 512, 256, 128 nodes (curve 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.

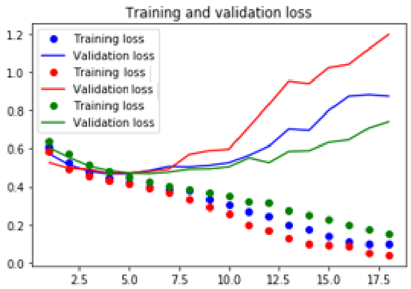


Figure 9: Comparison of the validation and training loss curves from training 3 LSTM networks differing in the number of nodes. Each is composed of 3 layers with the following number of nodes: red: 512, 256, 128; blue: 256, 128, 62; green: 128, 64, 32.

Table 4 summarizes the results obtained using the different approaches described in Methods. 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 each participant, the results are shown in the first two lines of table 4 and from these values, it is, undoubtedly, advantageous to use the longer samples. Hence, having more longitudinal information outperforms the advantage of having more samples, and so, more training data (the number of samples available using the two methods is indicated in table 3).

# ts	**# training samples**
3	14787
15	6063

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

Also, in table 4 it is evidenced that encoding the categorical features with a one-hot scheme lead to an improved performance of the model (all performance measures increased).

Adding dropout allowed to reduce considerably the overfitting phenomenon. The curves presented in figure 10 were obtained during training of the LSTM network, after encoding the categorical features using the one-hot scheme.

Finally, in table 4 we can also find the results obtained after adding the profile information about the patient to the model (with the adjacent necessary changes, explained in Methods). All of the performance measures improved, supporting the idea that this data (such as, the medical history of the parents, first symptoms noted, age at which they have been first noted) is informative for our pur-

Model	# ts	categorical features	Acc	TPR	TNR	BAcc
LSTM - seq	3	stand	0.738	0.840	0.482	0.661
LSTM - seq	15	stand	0.751	0.804	0.651	0.728
LSTM - seq	15	one-hot	0.772	0.821	0.672	0.747
LSTM - seq+stat	15	one-hot	**0.792**	**0.845**	**0.687**	**0.766**

Table 4: Performance comparison of the proposed models.

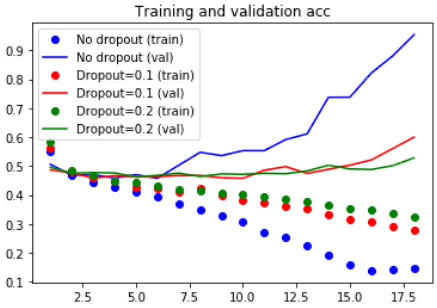


Figure 10: Comparison of the validation and training loss curves from training the same network without dropout (blue), with a dropout rate of 0.1 (red) and with a dropout rate of 0.2 (green).

pose.

4.3. GRU and standard RNN

Table 5 shows the results obtained with the different RNNs. The type of RNN that achieved the best performance was the GRU.

Model	Acc	TPR	TNR	BAcc
LSTM	0.792	0.845	0.687	0.766
Simple RNN	0.794	**0.856**	0.668	0.762
GRU	**0.796**	0.850	**0.690**	**0.770**

Table 5: Performance comparison of different RNN models.

Starting with the comparison between the LSTM model and the standard RNN (or ”SimpleRNN” as the Keras layer is called), the performance slightly worsened. Although the accuracy improved by 0.2%, 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 be achieved by simply having an increase in the number of samples being classified as belonging to the most common class. Nonetheless, the difference between performances was very small, indicating that the number of time-steps was low enough for not being notoriously affected by the vanishing gradient phenomenon.

GRUs are 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 better results.

4.4. HD vs controls

So far, the results that have been shown only regard the detection of depression in HD patients.

As it is possible to observe from table 6, the accuracy and the balanced accuracy 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 (1481 × 0.8 × 0.8 = 948), which is a limiting factor in Deep Learning [25]. Secondly, the used data comes from a database developed for purposes of studying the HD 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. There was an increase in the TNR and a decrease in the TPR, probably due to the classes representativity in the control group (indicated in table 2), showing that the method used during training to deal with the imbalance did not completely prevent the model’s tendency to classify a sample as belonging to the over-represented class.

Concerning the results from using the entire dataset (HD+control), although we had a larger training set and a diminished imbalance problem, the performance worsened, leading to the conclusion that the two groups differ in what is indicative of the class their participants belong to.

4.5. What is giving useful information to the network?

From table 7, we see that using the depression related features (DEP) along with the profile, the accuracy slightly increases (about 0.8%) comparing

Dataset	# training samples	Acc	TPR	TNR	BAcc
HD	6063	**0.796**	**0.850**	**0.690**	**0.770**
Controls	948	0.728	0.668	0.764	0.716
HD + Controls	7011	0.772	0.822	0.689	0.756

Table 6: Model performance comparison between groups.

Visits features	Profile	Acc	TPR	TNR	BAcc
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

Table 7: Performance metrics obtained using different sets of features.

with the use of all features; 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 underrepresented class. The presented values corroborate the idea that the DEP assessments benefit from being complemented with more objective data about the patient as a tool in detecting depression and should not be used alone.

5. Conclusions and future work

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. From 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 RNNs are able to use this data and extract useful outputs.

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. Nonetheless, the GRUs lead to the best results (although the performances were very similar).

One of the limitations of this work is the fact that only two classes were distinguished, using classes built on more restrict criteria would likely benefit the classification task. To improve the used method, it could be tested to add different data to the multi-input model (for example, imaging data).

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.

Acknowledgements

This document was written and made publically available as an institutional academic requirement and as a part of the evaluation of the MSc thesis in Biomedical Engineering of the author at Instituto Superior Técnico. The work described herein 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.

References

[1] Raymund A.C. Roos. Huntington's disease: A clinical review. *Orphanet Journal of Rare Diseases*, 5(1):2–9, 2010.

[2] Hoa Huu Phuc Nguyen and Patrick Weydt. Huntington disease. *Nature Reviews Disease Primers*, 1(1), 2015.

[3] James R. Slaughter, Matthew P. Martens, and Kathleen A. Slaughter. Depression and Huntington's Disease: Prevalence, Clinical Manifestations, Etiology, and Treatment. *CNS Spectrums*, 6(4):306–308,325–326, apr 2001.

[4] Paul Naarding, Joost G.E. Janzing, Paul Eling, Sieberen Van Der Werf, and Berry Kremer. Apathy is not depression in Huntington's disease. *Journal of Neuropsychiatry and Clinical Neurosciences*, 21(3):266–270, 2009.

[5] Aileen K. Ho, Abigail S. Gilbert, Sarah L. Mason, Anna O. Goodman, and Roger A. Barker. Health-related quality of life in Huntington's disease: Which factors matter most? *Movement Disorders*, 24(4):574–578, mar 2009.

[6] Eric A Epping and Jane S Paulsen. Depression in the early stages of Huntington disease. *Neurodegenerative Disease Management*, 1(5):407–414, oct 2011.

[7] Heath D Schmidt, Richard C Shelton, and Ronald S Duman. Functional Biomarkers of Depression: Diagnosis, Treatment, and Pathophysiology. *Neuropsychopharmacology*, 36:2375–2394, 2011.

[8] Shuang Gao, Vince D. Calhoun, and Jing Sui. Machine learning in major depression: From classification to treatment outcome prediction. *CNS Neuroscience and Therapeutics*, 24(11):1037–1052, 2018.

[9] André F Carvalho, Manu S Sharma, André R Brunoni, Eduard Vieta, and Giovanni A Fava. The Safety, Tolerability and Risks Associated with the Use of Newer Generation Antidepressant Drugs: A Critical Review of the Literature . *Psychother Psychosom*, 85:270–288, 2016.

[10] 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*, 26(5):975–989, sep 1996.

[11] Guy G. Potter and David C. Steffens. Contribution of depression to cognitive impairment and dementia in older adults. *Neurologist*, 13(3):105–117, 2007.

[12] Jane S Paulsen, Angèle Bénard Rsw, and Cyndy Moffat Forsyth. *Understanding Behaviour in Huntington Disease: A Guide for Professionals*. 3rd edition, 2016.

[13] Gordon J. Gilbert. Weight loss in Huntington disease increases with higher CAG repeat number. *Neurology*, 73(7):572, 2009.

[14] Jimmy Tan. Primed for Psychiatry: The role of artificial intelligence and machine learning in the optimization of depression treatment. Technical Report 1, 2019.

[15] Joost Schulte and J Troy Littleton. The biological function of the Huntingtin protein and its relevance to Huntington's Disease pathology. *Current trends in neurology*, 5:65–78, 2011.

[16] 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*, 79(8):874–880, 2008.

[17] Marie Noëlle W. Witjes-Ané, Maria Vegter-van der Vlis, Jeroen P.P. Van Vugt, Jan B.K. Lanser, Jo Hermans, Aeilko H. Zwinderman, Gert Jan B. Van Ommen, and Raymund A.C. Roos. Cognitive and motor functioning in gene carriers for Huntington's disease: A baseline study. *Journal of Neuropsychiatry and Clinical Neurosciences*, 15(1):7–16, 2003.

[18] Julie C. Stout, Jane S. Paulsen, Sarah Queller, Andrea C. Solomon, Kathryn B. Whitlock, J. Colin Campbell, Noelle Carlozzi, Kevin Duff, Leigh J. Beglinger, Douglas R. Langbehn, Shannon A. Johnson, Kevin M. Biglan, and Elizabeth H. Aylward. Neurocognitive Signs in Prodromal Huntington Disease. *Neuropsychology*, 25(1):1–14, 2011.

[19] Marina Papoutsi, Izelle Labuschagne, Sarah J. Tabrizi, and Julie C. Stout. The cognitive burden in Huntington's disease: Pathology, phenotype, and mechanisms of compensation, apr 2014.

[20] Kevin Duff, Jane S Paulsen, Leigh J Beglinger, Douglas R Langbehn, Chiachi Wang, MS C Julie Stout, Christopher A Ross, Elizabeth Aylward, Noelle E Carlozzi, and Sarah 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*, 22(2):196–207, 2010.

[21] Kevin Duff, Jane S. Paulsen, Leigh J. Beglinger, Douglas R. Langbehn, and Julie C. Stout. Psychiatric Symptoms in Huntington's Disease before Diagnosis: The Predict-HD Study. *Biological Psychiatry*, 62(12):1341–1346, dec 2007.

[22] Carol Efron Peyser and Susan E Folstein. Huntington's Disease as a Model for Mood Disorders Clues from Neuropathology and Neurochemistry. Technical report, 1990.

[23] 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*, 19(4):441–448, 2007.

[24] Huntington Study Group. Unified Huntington's Disease Rating Scale: Reliability and Consistency. *Movement Disorders*, 11(2):136–142, 1996.

[25] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[26] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. Understanding the exploding gradient problem. Technical report, Université de Montréal, 2012.

[27] Alex Graves. *Supervised Sequence Labelling with Recurrent Neural Networks*. 1 edition, 2012.

[28] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning Long-Term Dependencies with Gradient Descent is difficult. pages 157–166, 1994.

[29] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8, November 15, 1997):1735–1780, 1997.

[30] Klaus Greff, Rupesh K. Srivastava, Jan Koutnik, Bas R. Steunebrink, and Jurgen Schmidhuber. LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10):2222–2232, 2017.

[31] Tingyan Wang, Robin G. Qiu, and Ming Yu. Predictive Modeling of the Progression of Alzheimer's Disease with Recurrent Neural Networks. *Scientific Reports*, 8(1):1–12, 2018.

[32] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua 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)*, pages 1724–1734. Association for Computational Linguistics, 2014.

[33] Nikhil Ketkar. *Deep Learning with Python*. 2017.

[34] Junyoung Chung, Caglar Gulcehre, and Kyunghyun Cho. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[35] Michael Nguyen. Illustrated Guide to LSTM's and GRU's: A step by step explanation, 2018.

[36] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. Gated Feedback Recurrent Neural Networks. In *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, 2015.

[37] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

[38] Alex Graves. Generating Sequences With Recurrent Neural Networks. 2013.

[39] Rui Xu, Donald C. Wunsch, and Ronald L. Frank. Inference of genetic regulatory networks with recurrent neural network models using particle swarm optimization. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(4):681–692, 2007.

[40] Douglas Eck and Jürgen Schmidhuber. A First Look at Music Composition using LSTM Recurrent Neural Networks. Technical report, IDSIA, Manno, Switzerland, 2002.

[41] Garam Lee, Kwangsik Nho, Byungkon Kang, Kyung Ah Sohn, and Dokyoon Kim. Predicting Alzheimer's disease progression using multi-modal deep learning approach. *Scientific Reports*, 9(1):1–12, 2019.

[42] Aviv Nahon and Boaz Lerner. Temporal modeling of ALS using longitudinal data and long-short term memory-based algorithm. (April):25–27, 2018.

[43] REGISTRY Study Protocol Version 3.0 Replacing Version 2.0 REGISTRY-an observational study of the European Huntington-Disease Network (EHDN). Technical report, 2009.

[44] Data dictionary of enroll-hd - periodic dataset. https://www.enroll-hd.org/enrollhd$_{d}ocuments/2018-10-R1/Enroll-HD-DataDictionary-2018-10-R1.pdf$.

[45] Enroll-hd: A prospective registry study in a global huntington's disease cohort. clinical study protocol version 1. https://www.enroll-hd.org/enrollhd$_{d}ocuments/Enroll-HD-Protocol-1.0.pdf$.

[46] Michael Noll-Hussong, Francesca Burgio, Yoshifumi Ikeda, Federica Scarpina, and Sofia Tagini. The Stroop Color and Word Test. *The Stroop Color and Word Test Front. Psychol*, 8:557, 2017.

[47] Laura K. Sheridan, Hiram E. Fitzgerald, Kenneth M. Adams, Joel T. Nigg, Michelle M. Martel, Leon I. Puttler, Maria M. Wong, and Robert A. Zucker. Normative Symbol Digit Modalities Test performance in a community-based sample. *Archives of Clinical Neuropsychology*, 21:23–28, 2006.

[48] Stan Smith, Nelson Butters, Roberta White, Lauren Lyon, and Eric Granholm. Priming semantic relations in patients with Huntington's Disease. *Brain and Language*, 33(1):27–40, 1988.

[49] Jenny Callaghan, Cheryl Stopford, Natalie Arran, Marie-Francoise Boisse, Allison Coleman, Rachelle Dar Santos, Eve M Dumas, Ellen P Hart, Damian Justo, Gail Owen, Joy Read, Miranda J Say, Alexandra Durr, Blair R Leavitt, Raymund A C Roos, Sarah J Tabrizi, Anne-Catherine Bachoud-Levi, Catherine Bourdet, Erik van Duijn, and David 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*, 27(1):59–64, 2015.

[50] John E. Ware, Mark Kosinsky, and Susan D. Keller. A 12-Item Short-Form Health Survey: construction of scales and preliminary tests of reliability and validity. *Medical Care*, 34(3):220–233, 1996.

[51] Kelly Posner, Gregory K Brown, Barbara Stanley, David A Brent, Kseniya V Yershova, Maria A Oquendo, Glenn W Currier, Glenn A Melvin, Laurence Greenhill, Sa Shen, and J John Mann. The Columbia-Suicide Severity Rating Scale: Initial Validity and Internal Consistency Findings From Three Multisite Studies With Adolescents and Adults. *Am J Psychiatry*, 168(12):1266–1277, 2011.

[52] Yann LeCun, Leon Bottou, Genevieve B. Orr, and Klaus-Robert Muller. Efficient BackProp. Technical report, 1998.

[53] Zyad Kasasbeh Basel Shalabi, Luai Al Shaaban. Data Mining : A Preprocessing Engine. *Journal of Computer Science*, 2006.

[54] Yang Liu. Encoding Categorical Features - Towards Data Science, 2018.

[55] Model (functional API) - Keras Documentation.

[56] Sequential - Keras Documentation.

[57] Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic Optimization. In *International Conference on Learning Representations*, pages 1–13, 2015.

[58] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, and Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[59] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent Neural Network Regularization. 2014.

[60] Taklit Akrouf Alitouche Mohamed Bekkar, Hassiba Kheliouane Djemaa. Evaluation Measures for Models Assessment over Imbalanced Data Sets. *Journal of Information Engineering and Applications*, 2013.

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