PREDICTING CLINICAL SYMPTOMS OF DEPRESSION FROM ACOUSTIC SPEECH SIGNALS USING NEURAL NETWORKS

by

Sebastián Arturo Rodríguez Ordóñez

Submitted in partial fulfillment of the requirements
for the degree of Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
December 2021

My family gives me the strength and motivation to be a better version of myself. Those who were, are and will be, I dedicate this to all of them.

Table of Contents

List of Tables

List of Figures

Abstract

Approximately 280 million people suffer from depression, a disabling illness. Early diagnosis and effective monitoring are known to reduce adverse effects. Still, they require extensive clinical resources, thus motivating considerable work in automatic *detection* of depression, including from acoustic speech signals, with some recent success using deep learning. Much less work has been done for automated *assessment*. We make progress towards automated assessment by presenting the first approach to use acoustic features of speech to predict responses for individual items on validated clinical assessment tools and demonstrate results better than a majority-based baseline on many of the items. We achieve this using CNN, and LSTM architectures whose inputs are speech signals' acoustic features and outputs are distributions over individual item responses corresponding roughly to presence/absence of each such symptom. This approach provides valuable explanatory power as it inherently predicts which symptoms might lead to the overall assessment score.

Abbreviations

AASS Autobiographical Adult Speech Samples. iii, v, vi, 2, 3, 14–17, 19, 28, 29, 35, 36, 38–43

ANN Artificial Neural Networks. iii, 2, 8, 11

CDRIN Canadian Depression Research and Intervention Network. 14

CNN Convolutional Neural Network. iii, vi, vii, 9, 12, 30–34, 39, 40, 43, 46, 47

DAIC Distress Analysis Interview Corpus. 15

DAIC-WOZ Distress Analysis Interview Corpus - Wizard of OZ. iii, v, vi, 2, 3, 12, 15, 16, 18, 19, 28, 29, 38–40, 43–45

dB decibels. 30

DD Depression Detection. iii, 6, 7, 12

DSE Depression Severity Estimation. iii, 6, 12

eGeMAPS extended Geneva Minimalistic Acoustic Parameter Set. iv, vii, 2, 12, 22, 25, 31, 33, 34, 39, 40, 46, 47

FFT fast Fourier Transformation. vi, 20, 21

FORBOW Families Overcoming Risks and Building Opportunities for Well-being. 14

GeMAPS Geneva Minimalistic Acoustic Parameter Set. 22, 24, 25

Hz hertz. 28

IIDD Individual Item Depression Detection. iv, 17, 28, 29, 36, 38

LSTM Long Short-Term Memory. iii, vi, 2, 10, 12, 30, 32–34, 39, 40

MADRS Montgomery-Åsberg Depression Rating Scale. iii, vi, vii, 2, 4–6, 15–17, 39, 43, 46, 47

ms milliseconds. 30, 32

PHQ-8 Patient Health Questionnaire-8. iii, vi, 2, 5, 6, 15–18, 39, 40, 43

ReLU Rectified Linear Unit. 11

STFT Short-Time Fourier Transformation. 20

VMP Vocal Mind Project. 14

Acknowledgements

First, I would like to thank my supervisor, Dr. Sageev Oore, whose experience and support have massively shaped and guided this work. I am incredibly grateful for having had the chance to learn from him.

I would also like to thank Dalhousie University and the Vector Institute[1] for giving me the opportunity of being part of such renowned institutions in the pursue of this research.

In addition, I would like to thank Dr. Rudolf Uher, Sheri Rempel, and the whole Depression-Speech team for their advice, expertise, and essential role in creating crucial resources for this project.

To my colleagues at Dr. Oore's lab, especially to Sri Harsha Dumpala, I thank you for the ideas, conversations, and plentiful knowledge shared with me.

Last but not least, I would like to thank my parents, my sister, and the rest of my family for their guidance and constant support. Also, none of this would have been possible without my spouse, Ana Silva, who helps me navigate through the ups and downs with the most selfless love and support I have ever seen.

[1] www.vectorinstitute.ai

Chapter 1

Introduction

Depression, or major depressive disorder, is a mental disorder that affects approximately 280 million people worldwide [20]. By 1990, depressive disorders were amongst the 25 most disabling diseases [6]. The current outlook is not better, with depressive disorders ranked 13th in the same ranking and still the highest-ranked mental disorder [6].

The early assessment of depression in individuals is a fundamental step in offsetting the burden brought by this mental disorder. Multiple instruments have been created to measure depression severity; some of these are self-reported [25, 4, 36] and some others require a trained clinician [15, 30, 36]. These instruments are usually questionnaires/scales with various items that have to be answered. When aggregating the answers of all items, it results in a severity estimate.

Due to advances in technology and the importance of early depression assessment, many alternatives have been proposed to detect and estimate depression in an automated way from the estimates obtained using popular depression assessment instruments. These approaches have evolved through the years, starting with the usage of traditional machine learning models [49, 7, 29], all the way to artificial neural networks [19, 44, 10, 53, 9, 1, 50]. Moreover, different types of input have been used for automated depression detection or severity estimation. Existing approaches have used text [44], speech [10, 53, 9, 1, 7, 29] and video [50, 7, 29].

So far, existing automated methods have focused on tasks where the total score obtained from a depression assessment instrument determines the output to be predicted. However, none of the proposed methods utilize the individual items contained in the assessment tools. Considering these could be helpful both because it could suggest what symptoms/aspects of depression can be identified by an automated system. Also, it could provide potential explanatory power, e.g., “the overall severity score might be high because the patient might have a high score on items (a,b,c)”.

In this work, we define a binary classification task for detecting the presence of depressive symptoms in individual items of clinical depression assessment instruments. To train models for this task, we use two datasets: the Autobiographical Adult Speech Samples (AASS) and the Distress Analysis Interview Corpus - Wizard of OZ (DAIC-WOZ) [14], which contain recordings from hundreds of interviews with multiple participants, along with transcripts and depression severity estimates. The AASS dataset contains estimates from the clinician-rated Montgomery-Åsberg Depression Rating Scale (MADRS) [30] and the DAIC-WOZ dataset includes the self-rated Patient Health Questionnaire-8 (PHQ-8).

To detect depressive symptoms in individual items, we extract acoustic features from the speech recordings contained in both datasets, namely mel spectrograms and the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [12]. Together with these acoustic features, we propose architectures of Artificial Neural Networks (ANN), containing convolutional layers [27] and Long Short-Term Memory (LSTM) [17] layers. We train these models to detect depressive symptoms in each item in the depression assessment instruments.

We evaluate our proposed approaches using the F1 score, and we compare them to determine which features and models obtain better results in the different items, datasets, and depression instruments. Our results show that the proposed models perform better than an empirical majority-based baseline and show that no single proposed model is better than the rest at predicting all the items. Moreover, our results indicate that the scores vary across items, showing that some items of the depression assessment instruments are more easily detectable than others from the acoustic features of speech alone. For example, presence of suicidal thoughts are easier to detect than reduced appetite.

1.1 Contributions

In this work, we make the following contributions:

1. We introduce the problem of automatically predicting scores on individual items of depression assessment instruments from acoustic features of speech.

2. We cast this problem as a set of binary classification tasks, roughly corresponding

to detecting the presence of the individual clinical depressive symptoms.

3. We implement acoustic-based deep learning architectures to solve the proposed binary tasks.

4. We evaluate our models and provide a detailed analysis of the results on two different depression datasets scored with two severity estimation instruments: a self-reported and a clinician-rated.

1.2 Outline

The structure of the rest of the document is the following.

Chapter 2 We explain some background concepts for our proposed approaches and results, as well as existing related work. In particular, we provide background on depression, measurement, and model components. The related work explores research efforts in the field of depression detection and severity estimation.

Chapter 3 We discuss the datasets used across our experiments, AASS, and DAIC-WOZ, and we give information about the train, validation, and test splits used. Additionally, we present the task and show how this definition impacts the distribution of the datasets.

Chapter 4 We describe the feature extraction process and the steps taken when extracting the acoustic features from speech recordings.

Chapter 5 We describe some challenges we observed when working on depression-related tasks. We explain our approaches, including the architectures and their interaction with the previously defined acoustic features and challenges. We also briefly highlight some other approaches throughout the project and the intuition behind them regarding the challenges.

Chapter 6 We explain some details of the training procedure. We also evaluate and analyze the results from our approaches.

Chapter 7 Finally, we present concluding remarks, considerations, and future work.

Chapter 2

Background

2.1 Measuring Depression

Measuring depression is challenging. Over the last century, multiple depression assessment instruments/scales have been proposed. These scales are primarily divided into two types: self-rated [25, 4, 36] and clinician-rated [15, 30, 36]. We used datasets with depression severity estimates obtained with some of these scales/questionnaires during our experiments. This section contains an explanation of these questionnaires.

2.1.1 Montgomery-Åsberg Depression Rating Scale (MADRS)

Montgomery and Åsberg designed this 10-item scale in 1979 for measuring the depression severity of patients [30]. To estimate a patient's depression level, a trained clinician rates each item from MADRS on a scale from 0 to 6. In this scale, the rater decides if the score is defined in the main scale steps (0, 2, 4, 6) or not (1, 3, 5). The items to grade are the following:

1. Apparent sadness
2. Reported sadness
3. Inner tension
4. Reduced sleep
5. Reduced appetite
6. Concentration difficulties
7. Lassitude
8. Inability to feel

9. Pessimistic thoughts
10. Suicidal thoughts

The scores of the individual items are added to obtain a total score. This total score indicates the depression severity estimated by the MADRS questionnaire.

2.1.2 Patient Health Questionnaire-8 (PHQ-8)

The PHQ-8 [25] is a self-rated questionnaire designed for estimating depression severity. Since the questions are self-rated, people using this questionnaire answer it based on their own experience and get a depression estimation.
In this questionnaire, participants are asked: *"Over the past two weeks, how often have you been bothered by any of the following problems?"*. Where the problems are:

1. Little interest or pleasure in doing things
2. Feeling down, depressed, or hopeless
3. Trouble falling or staying asleep, or sleeping too much
4. Feeling tired or having little energy
5. Poor appetite or overeating
6. Feeling bad about yourself - or that you are a failure or have let yourself or your family down
7. Trouble concentrating on things, such as reading the newspaper or watching television
8. Moving or speaking so slowly that other people could have noticed. Or the opposite - being so fidgety or restless that you have been moving around a lot more than usual

There are four possible responses to the questions, each of them representing a score from 0 to 3: Not at all (0), several days (1), more than half the days (2), nearly every day (3). After answering the PHQ-8, patients will find themselves with an estimated score from 0 to 24 after adding the scores obtained in each item. The higher the score, the higher the estimated depression severity according to the PHQ-8 questionnaire.

2.2 Automated Depression Assessment

Automated assessment of depression can be done in multiple ways. These alternatives use the results from depression scales for determining the label. So far, existing approaches have focused on two tasks that are explained in this section.

2.2.1 Depression Severity Estimation (DSE)

The main objective of the task of Estimating Depression Severity is to predict the total score obtained when adding all the items present in the depression scale used to assess a participant.

2.2.2 Depression Detection (DD)

Depression Detection (DD) is a binary classification task. The objective is to accurately predict the presence or absence of depression in a person by considering their characteristics (e.g., speech) that might contain information about the depressive state. Since depression assessment is done using scales, defining a scale-dependent threshold over the total score is necessary to distinguish between depressed and healthy samples. For MADRS (Section 2.1.1), a cutoff of 10 and above has been found to indicate the presence of depression [55, 16]. For PHQ-8 (Section 2.1.2), its authors also recommend that any score of 10 or higher be classified as current depression [25].

2.3 Metrics

Classification models are usually measured with the accuracy metric, i.e., accuracy = $\frac{\text{\# correctly predicted samples}}{\text{\# samples}}$. However, when there is an interest in the performance of each class, a different set of metrics is used. This set of metrics is also used when there is the presence of class imbalance. In this section, we will use the confusion matrix on table 2.1to explain how to calculate the metrics. Table 2.1 contains an example of the usage of a confusion matrix on a binary depression detection task. We refrain from using vocabulary designed for the binary classification scenario since metric names may vary from positive to negative cases (e.g., sensitivity and specificity).

Table 2.1: Example of a confusion matrix. Columns indicate the actual label and rows correspond to the predicted label for a two class depression detection task (Depressed/Not Depressed).

		Actual label: Depressed	Actual label: Not Depressed	
Predicted label	Depressed	A	B	Predicted Depressed = $A+B$
	Not Depressed	C	D	Predicted Not Depressed = $C+D$
		Actual Depressed = $A+C$	Actual Not Depressed = $B+D$	

2.3.1 Precision

The way to compute this metric for one specific model is to find all the correctly predicted samples to have a class and divide that by the all samples predicted to have that class (correctly and incorrectly predicted). That is, precision measures the proportion of positive predictions that were indeed correct. Using the confusion matrix for the DD task in table 2.1 a more straightforward way to view it is.

$$\text{Precision(Depressed)} = \frac{A}{\text{Predicted Depressed}} = \frac{A}{A+B}$$

$$\text{Precision(Not Depressed)} = \frac{D}{\text{Predicted Not Depressed}} = \frac{D}{C+D}$$

2.3.2 Recall

Recall measures the proportion of the actual samples in a class that were correctly predicted. The way to calculate it is to find all those samples that were correctly predicted to have a class and then divide it by all those samples that truly belong to that class. Using the confusion matrix in table 2.1 for a simpler explanation on the DD task:

$$\text{Recall(Depressed)} = \frac{A}{\text{Actual Depressed}} = \frac{A}{A+C}$$

$$\text{Recall(Not Depressed)} = \frac{D}{\text{Actual Not Depressed}} = \frac{D}{B+D}$$

2.3.3 F1 Score

By definition, the F1 Score is the harmonic mean of the recall and precision metrics. It is a useful metric that takes into consideration the information obtained from both of the metrics. It is defined as:

$$\text{F1-score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad .$$

Usually, there is a trade-off between precision and recall. The higher the precision the lower the recall and the other way around. That is why the F1-Score is commonly used, given that it is a metric that considers the joint values of these two metrics.

2.4 Artificial Neural Networks (ANN)

Researchers have studied ANNs for decades. Initially, they were motivated by how the neurons in the brain work [28]. In recent years, their popularity has risen since they have been proven successful in various learning tasks. ANNs can be represented as a directed graph, where a set of nodes (called neurons) are joined by directed edges, where the edges represent weights that are learned during the training process. The activation (value) of a neuron depends on the activations of those neurons pointing towards it and the weights of the edges between them, as well as a non-linear activation function. To adapt an ANN to a task it is trained by adjusting those weights to minimize a loss function [38]. There are many different types of components in an ANN, those used in our proposed approaches will be explained in this section.

2.4.1 Linear Layers

Linear components in ANNs represent a set of neurons (layer) that is fully connected to another set of neurons (layer). Each neuron of the first layer has an edge pointing towards each neuron in the following layer. These edges contain a weight [38]. An example of an ANN made up of one linear hidden layer (not an input or an output) can be observed in image 2.1. ANNs with only linear layers are commonly called Feedforward Neural Networks. In order to calculate the value of a specific neuron i, $v_{n,i}$, in the layer V_n a weighted sum is applied across the neurons from layer V_{n-1}, where the weightings

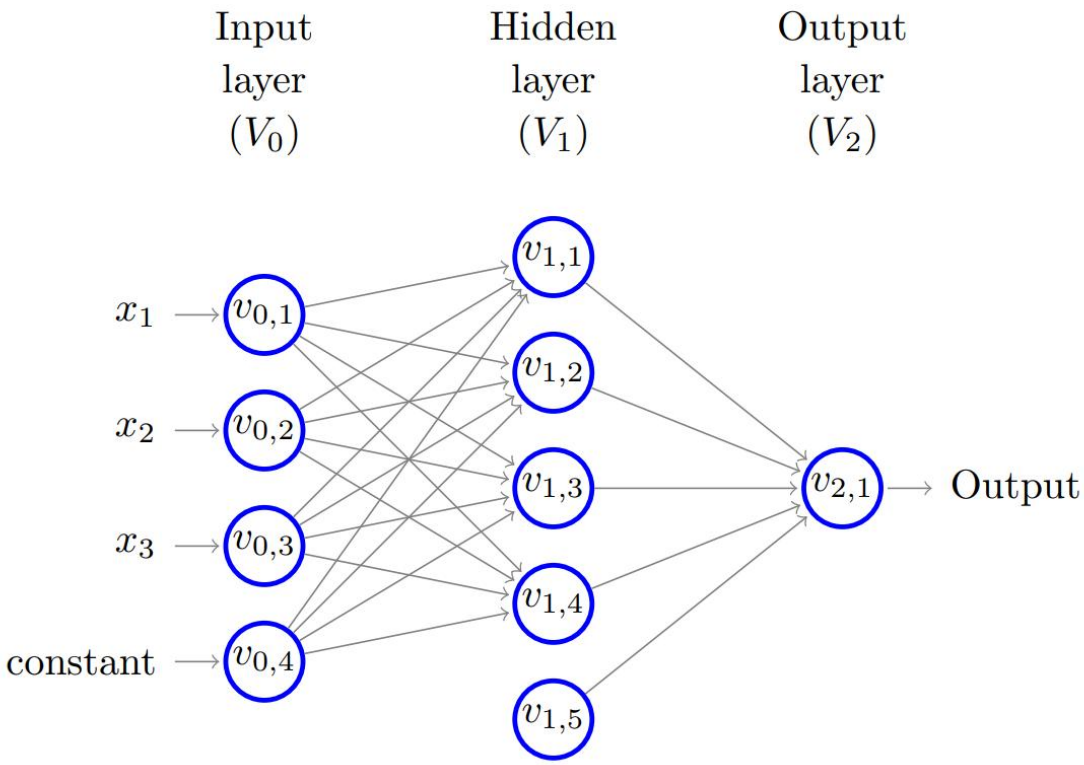


Figure 2.1: Example of a Feedforward Neural Network.

This neural network has one hidden layer. Extracted from [38]

used are the weights contained in the edges pointing towards neuron $v_{n,i}$.

$$v_{n,i} = \sum_{j=1}^{J} v_{n-1,j} w_{v_{n-1,j},v_{n,i}}$$

J is the number of neurons in layer V_{n-1} , and $w_{v_{n-1,j},v_{n,i}}$ is the weight connecting $v_{n-1,j}$ to $v_{n,i}$. It is important to note that usually an activation function is applied to each neuron to obtain the activation value used for the next layer, in this explanation it has been omitted, however, we briefly explain activation functions and their relevance in section 2.4.4.

2.4.2 Convolutional Neural Network (CNN)

CNNs date back to the late 1980s and early 1990s, having zipcode identification from handwritten digits [27] as one of the earliest applications. The biggest change brought by CNNs is the use of convolutional layers. Convolutional layers aim to learn kernels that are applied to the input. [13]. Convolutional layers are commonly applied across 1

and 2 dimensions. An example of the application of a 2-dimensional kernel to an input can be observed in figure 2.2.

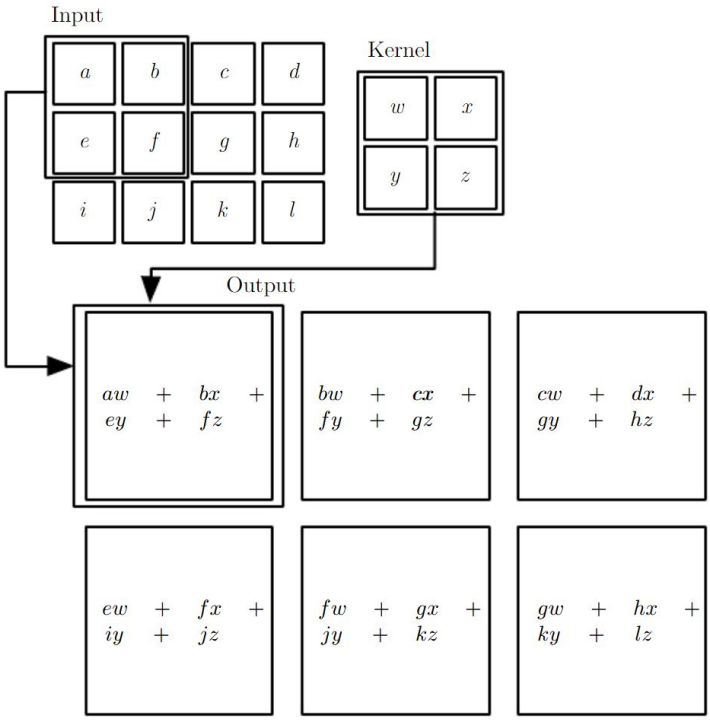


Figure 2.2: Example of the calculation of the output of a convolutional layer

The image shows a $2x2$ kernel being applied to a $3x4$ input, and the output that this process generates. Extracted from [13].

2.4.3 Long Short-Term Memory (LSTM)

The LSTM layer was proposed in 1997 [17]. Researchers designed this component to work with sequence data. The objective of this kind of model is to create an intermediate representation on one step x_t of the data, and use it as part of the inputs needed when creating the representation for the next step x_{t+1} in the sequence of data. This pattern is common across Recurrent Neural Networks [39], however, LSTM introduces a group of gates in order to make it easier for the gradient to spread across longer sequences [13]. Figure 2.3 shows the components inside of an LSTM layer. Namely, three different gates are introduced as part of this component a forget gate, an input gate, and an output gate. Using these three gates, the input, and the information shared from the previous sequence step, the information for the next element in the sequence is obtained [13].

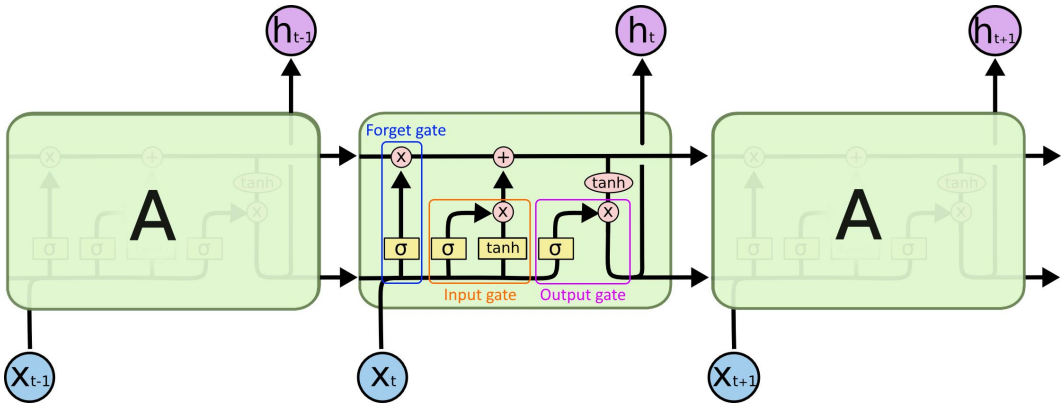


Figure 2.3: LSTM usage and components.

The diagram shows how information is shared between time steps of sequence data. σ (Sigmoid) and tanh represent the activation functions (section 2.4.4). Image extracted online from [5], we added annotations for clarity.

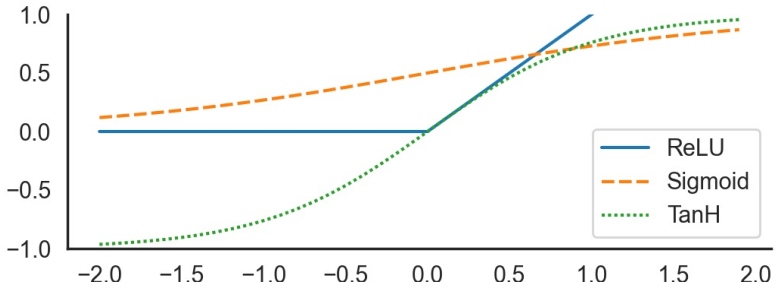


Figure 2.4: Activation functions visualized.

2.4.4 Activation Functions

Activation functions are a fundamental part of modern ANNs, without them, ANNs would be linear models regardless of the number of layers. This component can be applied to any neuron across a network to introduce non-linear relations between neurons. These non-linear relations help the models learn weights that generalize over the existing non-linearities prevalent in data. Some of the most common activation functions are the Sigmoid, the Hyperbolic Tangent (tanh), and the Rectified Linear Unit (ReLU) activation functions. They are defined as follows:

$$\text{ReLU}(x) = \begin{cases} x, & \text{if } x > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$\text{Sigmoid}(x) = \frac{1}{1 + e^{-}x}$$

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

Figure 2.4 shows a visual example of the output values of these activation functions given an input value x.

2.5 Related Work

Due to its importance, automated depression assessment has gained popularity in the last decades, fueled by existing workshops with depression-related challenges, such as the Audio/Visual Emotion Challenge 2016 [47] and 2019 [35], and by the release of an open depression dataset containing text, audio and video samples, DAIC-WOZ [14], depression tasks have had an increase in popularity. Solutions have been proposed using transcriptions (text) [44], speech, and video recordings [50].

Speech-based approaches have used a variety of features. From standard acoustic feature sets (such as INTERSPEECH 2009 [37], AVEC 2013 [46], COVAREP [8], eGeMAPS [12]) extracted with freely available toolkits (such as COVAREP [8], OpenSMILE [11] and the WORLD Vocoder [31]), Spectrograms [2], and even proposing innovative features [49, 19].

As alternatives for the existing depression-related tasks, DD and DSE, and, in conjunction with these previously mentioned features, machine learning models have been proposed, including Support Vector Regressors [46] and Gaussian Mixture Models [49]. Moreover, many neural architectures have been recently proposed. Some of these architectures have taken advantage of the progress done in the computer vision field by using CNNs [9, 54]. Others have used speech features as a sequence of data, using LSTMs [1, 53] for depression-related tasks. Recently, attention mechanisms have been also proposed as part of proposed solutions [54, 53]. Other authors have even explored existing approaches from the speech recognition task to enhance the performance of their depression models [10, 53].

Individual items from depression assessment scales are fundamental when determining the depression level present in a sample. These individual items have been used in the past as a core element in depression bio-marker studies [45, 18, 3]. To the best

of our knowledge, no automatic individual item assessment systems/models have been proposed in previous studies.

Chapter 3

Datasets & Tasks

In this chapter, we describe the datasets used during our experiments, along with the depression scales used for rating the depression severity for the participants. Additionally, we explain the task we focus on during our work and its differences from the tasks described in the section 2.2.2.

3.1 Datasets

In total, we use two datasets for our experiments which follow different collection processes and contain depression estimates from different depression assessment instruments. Using multiple datasets might give a better idea of how effective our approaches can be with different depression assessment scales. In this section, we explain these datasets and give a brief overview of their contents.

3.1.1 Autobiographical Adult Speech Samples (AASS)

This private dataset contains audio samples recorded by the Families Overcoming Risks and Building Opportunities for Well-being (FORBOW) team as part of projects FORBOW, Canadian Depression Research and Intervention Network (CDRIN), and Vocal Mind Project (VMP). The recording process follows a predefined standard operating procedure that establishes the recording device, settings, location in the room, interviewer prerequisites, and more. Each of the recordings corresponds to a person speaking about their last few weeks. The interviewer asks each participant these three prompts:

1. (Neutral) Tell me how you have been feeling and what you have been up to lately.
2. (Positive) Think about when you had a positive experience or when something good may have happened to you.
3. (Negative) Think about when you had a negative experience or when something bad may have happened to you.

Each participant is given a brief introduction to be ready for the interview and is expected to talk for 3 minutes in each prompt. Afterward, trained clinicians score the recordings using the MADRS scale. Some of the participants undertake multiple interviews on different days to assess their depression levels across time. The recordings are also split into different segments where raters mark the sentiment and emotion levels of the speaker throughout the recording. We consider a total of 131 recordings containing MADRS item scores and transcripts as part of our experiments. We separate these recordings across different splits; train split includes 94 samples, validation 16 samples, and test 21 samples.

3.1.2 Distress Analysis Interview Corpus - Wizard of OZ (DAIC-WOZ)

The DAIC-WOZ dataset is available upon request and is part of the bigger Distress Analysis Interview Corpus (DAIC) [14], which comprises a group of clinical interviews created to help with diagnosing depression. This dataset has been used as part of the Audio/Visual Emotion Challenges [47, 35].

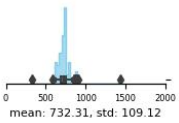
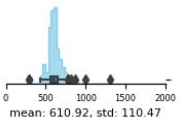
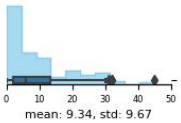
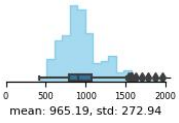
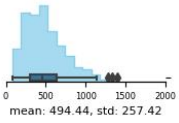
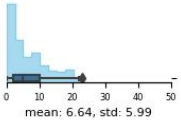
This dataset contains audio, video, and transcripts for multiple recording sessions. It was collected in a "Wizard of OZ" or "OZ Paradigm" setup [22] where Ellie, a virtual interviewer controlled by a person in a different room, interviews each participant. The transcripts for the recordings are also provided, along with the depression severity scores that participants reported using the PHQ-8 questionnaire.

The authors of the dataset provide suggested train, validation, and test splits. However, the testing split does not contain detailed information about the PHQ-8 item scores. For this reason, we decided to use the proposed validation set as our test set, and we split the proposed train set into train and validation splits. In total, we consider 219 samples, with 56 on the test set, 138 on the train test, and 25 on the validation set.

3.1.3 Exploration

Even though both datasets (AASS and DAIC-WOZ) have been collected to estimate depression severity, the content in each of them is pretty different. A quick summary of the primary differences, including recording length, samples, and depression scores, is located in table 3.1. From observing the table, it is clear that the datasets are very different in nature, for example, there is a clear difference in the length of the audio clips,

Table 3.1: Datasets' descriptive statistics.

Dataset	Study Status	# Samples	Length (seconds)	Voiced Length (seconds)	Total Depression Score
AASS (MADRS)	Ongoing	131			
DAIC-WOZ (PHQ-8)	Done	219			

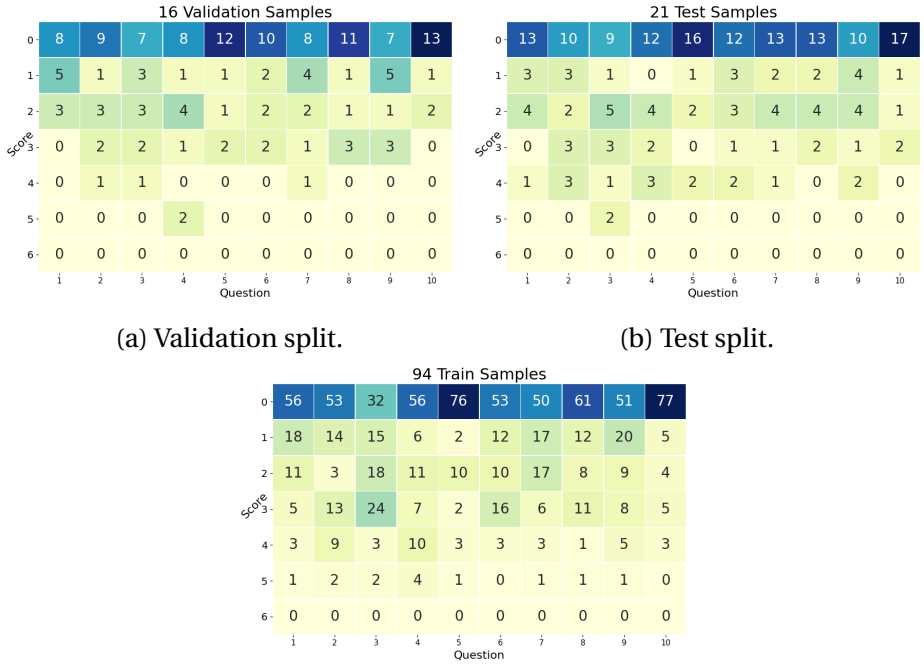
which is carried into the voiced (interviewer) version. Moreover, the total depression score distribution ranges vary due to the change in scale across the datasets, however, a high concentration can be seen in the lower depression sections in both datasets.

It is also fundamental to understand how depression item scores are distributed across datasets. Figures 3.1 and 3.2 show the item scores for the datasets used. It is important to note that the depression scales used for the datasets are different and, thus, the scores will show a different structure. From these two illustrations, it can be confirmed that there is a concentration in lower depression scores, noticeable in the total scores from table 3.1, and evident in figures 3.1 and 3.2.

Another consideration that can be derived from figures 3.1 and 3.2 is that extreme depressive symptoms are not frequent. In the case of the AASS dataset, none of the MADRS items achieves the maximum score of 6, defined by the scale. In the DAIC-WOZ dataset, scores of 3 are reported; however, it is the least common score across the eight questions in the questionnaire.

3.2 Task

We gave some background on the existing depression-related tasks in section 2.2 as a context of the automated approaches related to depression. However, our work does not focus on any of those tasks. We decided to propose an alternative task that takes



(a) Validation split.

(b) Test split.

(c) Train split.

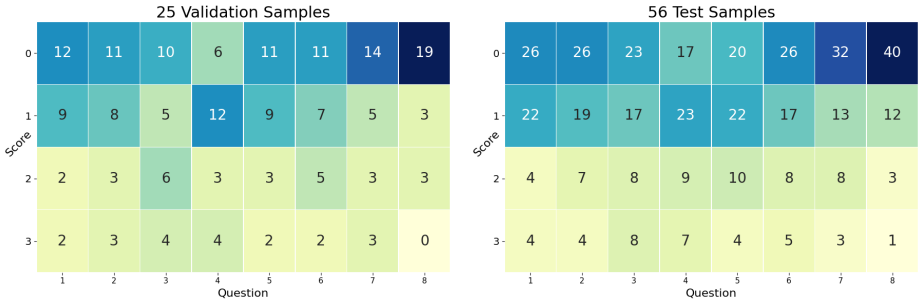
Figure 3.1: AASS dataset item/question depression assessment score distribution according to the MADRS scale.

into consideration the individual items of depression assessment instruments. We will explain the task in this section.

3.2.1 Individual Item Depression Detection (IIDD)

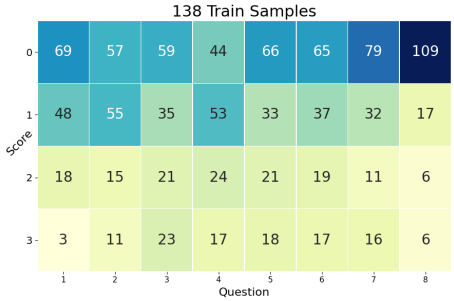
We propose the IIDD task as the task of predicting the presence of depressive symptoms corresponding to above-zero scores on individual items in either depression scale, MADRS or PHQ-8. This task has not been previously used as the target of any depression-related system. The main objective of this task is to find out which items from the depression severity scales can be coarsely but effectively detected using automated systems if any.

After applying the task definition, we are left with the distribution of samples for the test set shown in table 3.2. It is important to note that the number of recordings in this table does not reflect the total number of segments used during the training stages. Interestingly, the application of these thresholds reduces the class imbalance observed in section 3.1.3. Some of the depression assessment instruments items show a value



(a) Validation split.

(b) Test split.



(c) Train split.

Figure 3.2: DAIC-WOZ dataset item/question depression assessment score distribution according to the PHQ-8 questionnaire.

distribution close to uniform across the classes in both datasets.

Table 3.2: Samples on test splits after applying the depression thresholds defined in sections 2.2.2 and 3.2.1. Top section contains items for the AASS dataset, bottom section for the DAIC-WOZ. #(ND) Indicates the number of samples considered as "Not Depressed"/"No signs of Depression" and #(D) indicates those "Depressed"/"Presence of Depression" samples according to the thresholds

	Train set		**Validation set**		**Test set**	
Item / Question	#(ND)	#(D)	#(ND)	#(D)	#(ND)	#(D)
AASS (MADRS)						
(1) Apparent sadness	56	38	8	8	13	8
(2) Reported sadness	53	41	9	7	10	11
(3) Inner tension	32	62	7	9	9	12
(4) Reduced sleep	56	38	8	8	12	9
(5) Reduced appetite	76	18	12	4	16	5
(6) Conc. difficulties	53	41	10	6	12	9
(7) Lassitude	50	44	8	8	13	8
(8) Inability to feel	61	33	11	5	13	8
(9) Pessimistic thoughts	51	43	7	9	10	11
(10) Suicidal thoughts	77	17	13	3	17	4
DAICWOZ (PHQ-8)						
(1) Little interest	69	69	12	13	26	30
(2) Feeling down	57	81	11	14	26	30
(3) Trouble sleeping	59	79	10	15	23	33
(4) Feeling tired	44	94	6	19	17	39
(5) Poor apetite	66	72	11	14	20	36
(6) Self-dissapointment	65	73	11	14	26	30
(7) Conc. difficulties	79	59	14	11	32	24
(8) Restlessness	109	29	19	6	40	16

Chapter 4

Feature extraction

The feature extraction process is an essential step in our work since this allows us to work with larger audio windows while maintaining relatively small models, given the relatively small datasets available. In this chapter, we explain the acoustic features and dataset scaling techniques applied to the audio recordings on the datasets. We use this features throughout our approaches. Chapter 5 describes how we do this.

4.1 Features

4.1.1 Spectrograms

The Spectrogram is a visual representation of the frequency, time, and intensity of an audio signal's spectrum [24]. To create this representation an FFT is applied across short-time windows (Short-Time Fourier Transformation (STFT)) [32, 2]. Figure 4.1 shows this process, where short-time windows are selected from an audio signal. Afterward, a low-pass filter is applied to the signal, and then the FFT takes place. In the end, the result contains information about the intensity and frequency of the input window. The results for subsequent windows are appended one after another to obtain a spectrogram. Figure 4.2 contains an example of a spectrogram.

4.1.2 Mel-Spectrograms

Mel-Spectrograms build upon the spectrograms described in section 4.1.1. Moreover, to obtain mel spectrograms, a transformation has to be done to the spectrogram, following the mel scale. The mel scale [43] is a non-linear subjective scale, designed according to how humans perceive sound, where the distance in the scale represents pitches that sound equally distant for a human listener. The mel spectrogram is a version of the spectrogram where the frequency dimension is replaced by the mel scale, with evenly

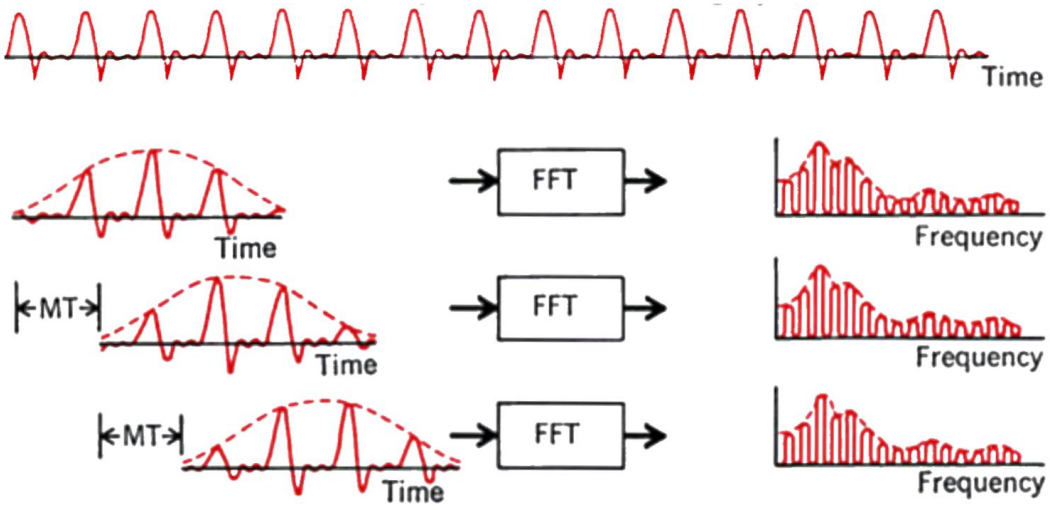


Figure 4.1: Computation of running spectrum using FFT.

Subsequent time windows are extracted from the original audio signal (top-left of the image), go through a low-pass filter (dotted line on the graphs on the left), have an FFT applied to them, and result in a frequency & intensity representation (right side of the image). Extracted from [32].

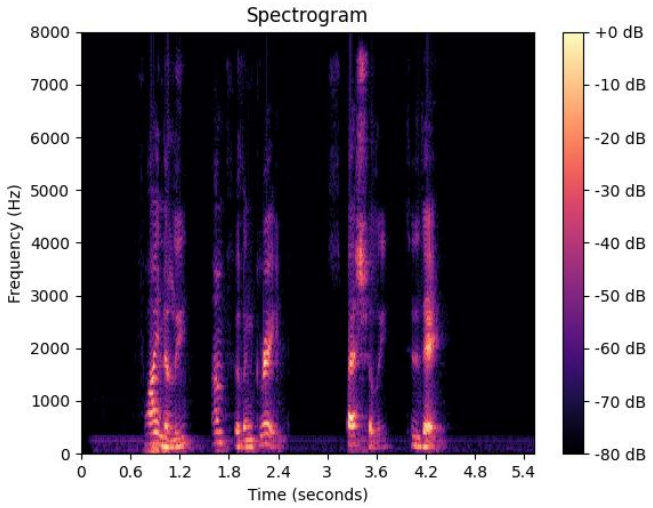


Figure 4.2: Example of a spectrogram.

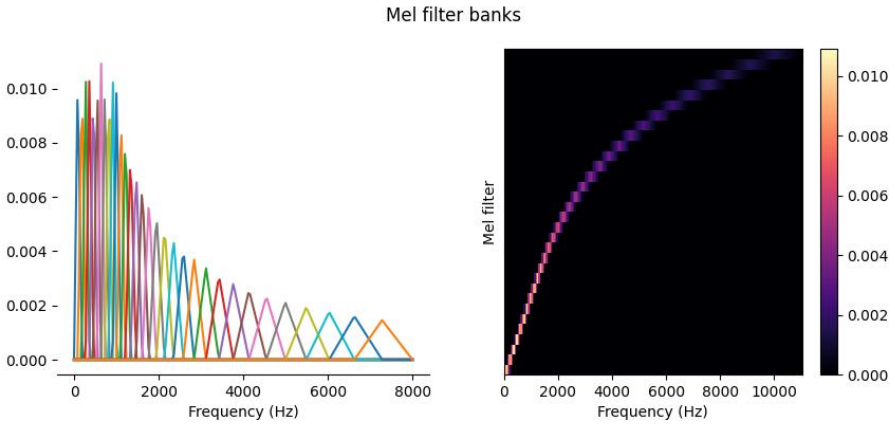


Figure 4.3: Example of 32 Mel Filter banks.

This figure contains two alternative visualizations of 32 Mel filter banks.

distributed mel bins. This representation can be achieved by using filter banks that define the linear transformation from frequency to the mel scale [40]. We have created two visualizations of 32 mel filter banks, which we added as figure 4.3. Both images point out that higher filter banks in the frequency spectrum take information from a more extensive range of frequencies. This filter bank transformation is applied to a spectrogram equally across time, resulting in a mel spectrogram. Figure 4.4 shows the change of a spectrogram into a mel spectrogram. Some differences are easily noticeable. In general, Mel Spectrograms have been preferred over Spectrograms for deep learning applications since they are a closer representation of how humans perceive sound. Mel Spectrograms have been widely used in many audio-based applications, including speaker verification [48], speech emotion recognition [34], and depression estimation [53].

4.1.3 extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS)

The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) feature set was proposed in 2016 as a minimalist feature set to work with speech, and emotion applications, based on multiple acoustic analyses [12]. eGeMAPS was proposed along with the GeMAPS acoustic feature set and is an extended version of it.

Geneva Minimalistic Acoustic Parameter Set (GeMAPS)

This acoustic feature set contains 18 different parameters or low-level descriptors, listed as follows with the number of parameters per item in parenthesis:

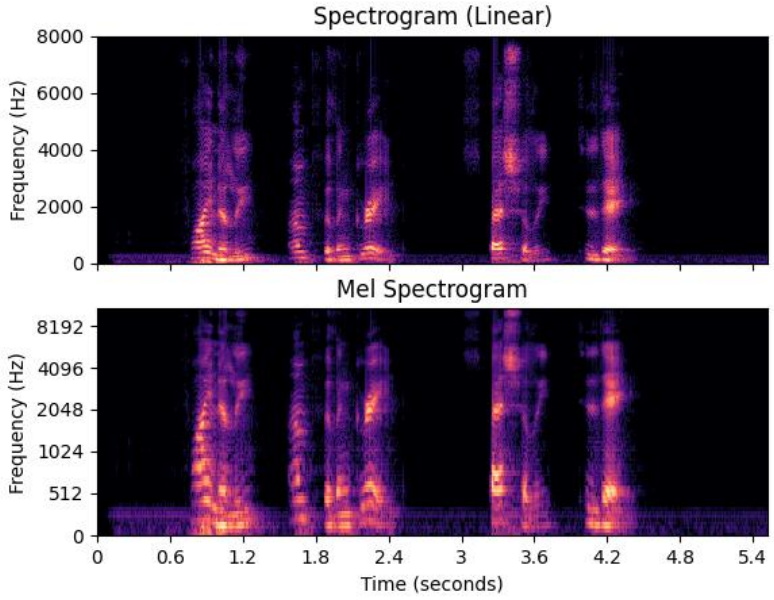


Figure 4.4: Spectrogram and Mel Spectrogram side-by-side.

1. (1) Pitch.
2. (1) Jitter.
3. (3) Center frequency of Formants 1, 2, and 3.
4. (1) Bandwidth of the Formant 1.
5. (1) Shimmer, the difference of the peak amplitudes of consecutive F0 periods.
6. (1) Loudness.
7. (1) Harmonics-to-noise ratio (HNR).
8. (1) Alpha Ratio.
9. (1) Hammarberg Index.
10. (2) Spectral slope 0-500 Hz and 500-1500 Hz of the logarithmic power spectrum.
11. (3) Formant 1, 2, and 3 relative energy.
12. (1) Harmonic difference first F0 harmonic (H1) to the energy of the second F0 harmonic from the third formant (H2).

13. (1) Harmonic difference of the first F0 harmonic(H1) to the energy of the highest harmonic from the third formant (A3).

The original publication [12] contains a better detail of the features. These features are also smoothed across time using a symmetric moving average with a length of 3 frames [12].

In addition to this set of features, some statistics, functions, and temporal features are calculated at the recording level, which summarizes the mentioned features across time, giving a single element [12]. The statistics applied, with the number of resulting features in parenthesis, are:

1. (36) Arithmetic mean and coefficient of variation of all the 18 features.
2. (16) 20th percentile, 50th percentile, 80th percentile, range 20th to 80th percentile, mean of the rising signal sections, standard deviation of the increasing signal sections, mean of the falling signal sections, and standard deviation of the falling signal sections of the features loudness and pitch.
3. (4) Arithmetic mean of the features Alpha Ratio, Hammarberg Index, Spectral slope 0-500 Hz, and Spectral slope 500-1500 Hz.
4. (1) Rate of loudness peaks.
5. (2) Mean length and the standard deviation of continuously voiced regions.
6. (2) Mean length and the standard deviation of unvoiced regions.
7. (1) Number of continuous voiced regions per second.

In total, the GeMAPS feature set contains 62 parameters when considering all the functions, statistics, and time features.

Extension

The extended version contains some additional features, which are:

1. (4) Mel-Frequency Cepstral Coefficients 1-4.
2. (1) Spectral flux.

3. (2) Bandwidth of the Formants 2 and 3.

Similar to the process done in GeMAPS, the next step is to calculate additional statistics and functionals of these features across time.

1. (10) Arithmetic mean and coefficient of variation of features Mel-Frequency Cepstral Coefficients 1-4, and Spectral flux.

2. (4) Arithmetic mean and coefficient of variation of features Formant 2, 3 bandwidth for voiced sections only.

3. (1) Arithmetic mean of features Spectral flux for unvoiced sections only.

4. (10) Arithmetic mean and coefficient of variation to features MFCC 1-4, Spectral flux for voiced sections only.

5. (1) Equivalent sound level.

In total, the number of additional statistics and functionals contained in eGeMAPS is 26. The authors of the eGeMAPS feature set [12] explain this set of feature in a more detailed manner. When adding these 26 parameters and the 62 parameters contained in GeMAPS, the grand total count of the number of parameters in eGeMAPS is 88.

4.2 Data Scaling

An essential step in the feature creation process is scaling the features used by the models. This step is beneficial since it is common to have features in different scales, which can negatively impact the training process of neural networks. We apply data scaling to all the dataset splits (train, validation, and test). Below, we explain the data scaling techniques used across our approaches.

4.2.1 Range Scaling

This scaling technique can be used when the features are bounded by a specific range that does not vary across samples and datasets. After defining an upper limit and a lower limit, the range normalized value of a sample s, at time t, for the feature f is:

$$\tilde{x}_{s,t,f} = \frac{x_{s,t,f} - \text{lower bound}}{\text{upper bound} - \text{lower bound}}$$

4.2.2 Min-Max Scaling

A maximum and a minimum value are obtained from the training dataset for each feature across time and samples. These maximum and minimum values are then used as the upper and lower bounds of the range scaling technique, respectively. Given the minimum value for the feature f, $\min f$, and the maximum value, $\max f$, the final value for a sample s, at time t, for the feature f is:

$$\tilde{x}_{s,t,f} = \frac{x_{s,t,f} - \min f}{\max f - \min f}$$

Chapter 5

Automated Assessment

In this chapter, we explore the automated assessment of the task described in section 3.2.1. First, we describe questions and challenges that arose during this work. These challenges led us to try many different approaches. We describe in detail the approaches we found to be most successful. Finally, we briefly summarize some of those experiments we tried that we did not include in the final set of results in section 6

5.1 Challenges

Throughout our work with depression-related tasks, we have noticed some important challenges to keep in mind. In this section, we explain them. In our approaches (section 5.2) we mention how we expect them to solve one or more challenges.

5.1.1 Length of Samples

An important aspect that has to be considered is the length of the recordings. By observing table 3.1, it is clear that the datasets have a variety of lengths with average lengths of more than 8 minutes. Having samples with these durations brings another level of complexity, since using a whole recording as the input of a model, given the proposed features, becomes a difficult task due to the length of the sequence.

5.1.2 Recording predictions

Moreover, if we split recordings into multiple segments, all the segments from a single recording will have the same depression label and the model will make predictions at the segment label. Since our interest lies in predicting the depression level for a complete recording, which represents the current depression state of a person, a strategy to aggregate the segment level predictions into recording/track level predictions has to be used.

5.1.3 Unbalanced Data

From section 3.1.3, its tables, and figures we know that the explored datasets contain unbalanced classes. To be precise, there is a high concentration of labels on the lower end of depression presence, while there are very few recordings from extremely depressed individuals. Since labels do not resemble a uniform distribution, it is necessary to consider solutions that consider this issue.

5.1.4 Limited Samples

An essential aspect of training a model for a specific task is to have enough samples to generalize well on unseen data. In this case, we have ~22 hours and ~30 hours in the AASS and DAIC-WOZ, respectively. However, the recordings have an average duration of over 8 minutes. Therefore, each dataset has 'effectively' less than 300 samples. For this reason, a significant challenge is to explore existing alternatives to deal with a limited amount of samples.

5.2 Approaches

This section outlines the different approaches used throughout our experiments. It is important to note that the outputs of the models will be logistic units predicting binary variables (for the IIDD task). The features and samples can vary across each of the approaches and, therefore, we will describe them for each approach.

There are some processes shared by all our approaches, spanning from pre-processing to getting track-level predictions. Those processes are the following:

- **Audio re-sampling.** The recordings in both of the datasets are re-sampled to have a sampling frequency of 16,000 hertz (Hz). This step is necessary since we found some recordings to have higher sampling rates than the rest. We applied a down-sampling process using the Sound eXchange (SoX) utility [41] for these recordings.

- **Only participant speech.** An important pre-processing step shared by all our approaches is that we remove all the speech that does not belong to the speaker prior to the creation of any features. We do this by using the recording transcripts

from the datasets. The processed recordings only contain audio from those sections spoken by the participant/interviewee.

- **Logistic output (Binary classification).** All of our approaches contain models trained on the task IIDD (section 3.2.1) for all the items in the depression severity estimates of the AASS and DAIC-WOZ datasets. Therefore, the output could be interpreted as representing the probability of presence/absence of each item (i.e., depressive symptom). As we showed in section 3.2.1, having this task as the objective helps reduce the prevalence of the challenge of unbalanced data (section 5.1.3), in some of the depression items data is close to being balanced.

- **Majority Vote for recording-level predictions.** Due to the long samples (see section 5.1.1), we split all recordings into segments of varying length. This introduces the secondary challenge of obtaining recording-level predictions from segment-level predictions (see section 5.1.2). To obtain the prediction for a recording, we use *majority voting*. This voting scheme is a frequently used strategy in automated depression detection models [47]: a majority vote is carried across the binary predictions for each of the segments. An example of the majority vote can

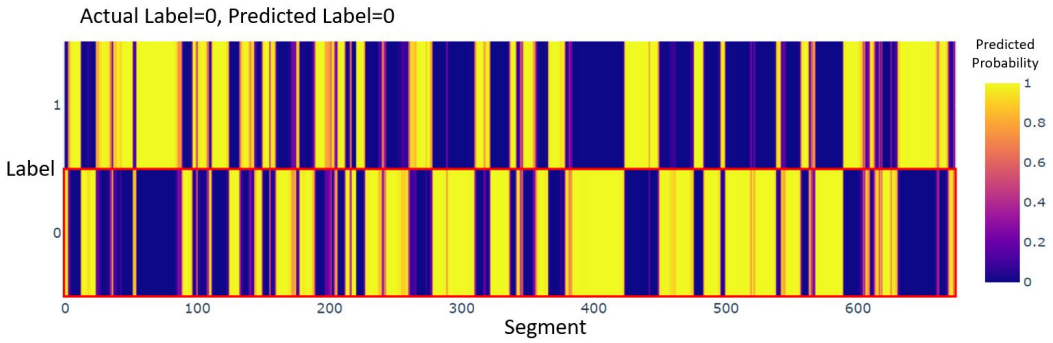


Figure 5.1: An Example of the majority vote.

This example corresponds to the heat map of the predictions of a recording that has been split into 660 segments. Top row indicates the probability of a segment of being 'Depressed' (Label=1) and the bottom row correspond to that of being 'Not Depressed' (Label=0). Each column corresponds to those predictions for a segment, when adding both probabilities (top and bottom rows) the results is always 1 for any segment. It is clear that the predictions vary across segments, some are clearly predicted as 'Not Depressed' (Label=0), some as 'Depressed' (Label=1), and some have similar probabilities for both classes. At the end of the majority vote process, the prediction is 0 (Not Depressed). This prediction is correct since we know that the actual label, marked by the red box, is also 0 (Not Depressed).

by observed in image 5.1. This process follows a defined set of steps. First, a prediction is obtained in each of the segments by selecting the class with the higher predicted probability. Afterward, a majority vote is carried across the segments; therefore, the most commonly predicted class will correspond to the final prediction for the recording.

5.2.1 Models

Spectrogram CNN

Features & Samples This approach uses mel spectrograms, explained in section 4.1.2and shown in figure 4.4, as the features. We create the mel spectrograms using a 20 milliseconds (ms) window, 20 ms steps, and 64 mel filter banks. The scaling technique used is range scaling since the values of the mel spectrogram have a range from -80 decibels (dB) to 0 dB. We use there as the lower and upper bounds, respectively. As a solution to the length of samples challenge (section 5.1.1), recordings are split into segments. One sample/segment is defined to be the mel spectrogram extracted from 4 seconds of a recording. There is a step of 1 second from sample to sample, meaning there is an overlap between contiguous samples.

Model Architecture In figure 5.2 we describe the main component of this network, it consists of 3 parallel strided 2-dimensional convolutional layers, with different kernel sizes, dropout [42] and batch normalization [21]. The outputs of the layers are flattened and appended into a single 156-dimensional embedding. This embedding is sent through a linear layer that transforms it into the desired output, two units, sent through the final activation function (softmax). Recently, similar strided convolutional architectures with parallel layers have been proposed for depression-based models with success [10, 19]. The explained component follows a similar pattern.

Spectrogram CNN-LSTM

Features & Samples Similar to the Spectrogram CNN, this model uses mel spectrograms extracted from the recordings using a window of 20 ms and 20 ms steps, with

64 mel filter banks applied. Even though we generate the features using the same parameters, the samples we use to train the models are different. In this case, one sample/segment corresponds to the spectrograms from 10 4-second chunks with steps of 1 second at a time, meaning that in total, one sample contains overlapping spectrograms corresponding to 13 seconds of the recording. The is a step of 1 second between one sample/segment and another.

Model Architecture This model builds on the Spectrogram CNN. We use the main component from figure 5.2 across multiple spectrograms. We use the resulting embeddings as the input of an LSTM with a hidden state size of 64 with dropout, as we show in figure 5.3. The output of the LSTM is sent through a fully connected layer that produces the logits received by the last activation function (softmax).

eGeMAPS CNN

Features & Samples The features used in this approach are explained in section 4.1.3. The eGeMAPS feature set suggests a list of low-level acoustic descriptors and a list of functions/statistics to apply across time which removes the time dimension. The features used in our approach correspond to the list of low-level descriptors extracted by

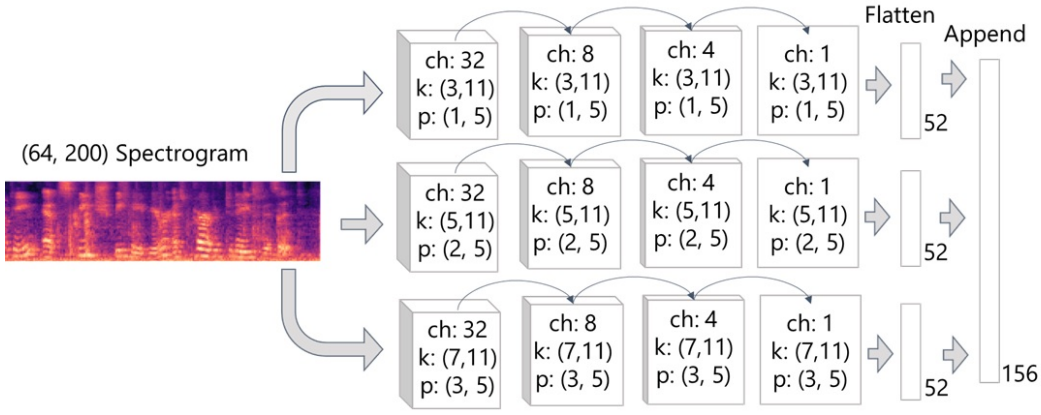


Figure 5.2: Spectrogram CNN component.

Each box represents a 2-d convolution followed by a ReLU activation function. All the convolutions had a stride of (2, 2). (ch, k, p) represent the number of channels, kernel size, and padding, respectively. Each convolutional layer has dropout and batch normalization. Strided convolutional architectures with parallel layers have been previously used for depression tasks with success [19, 10]. The proposed architecture follows a similar pattern.

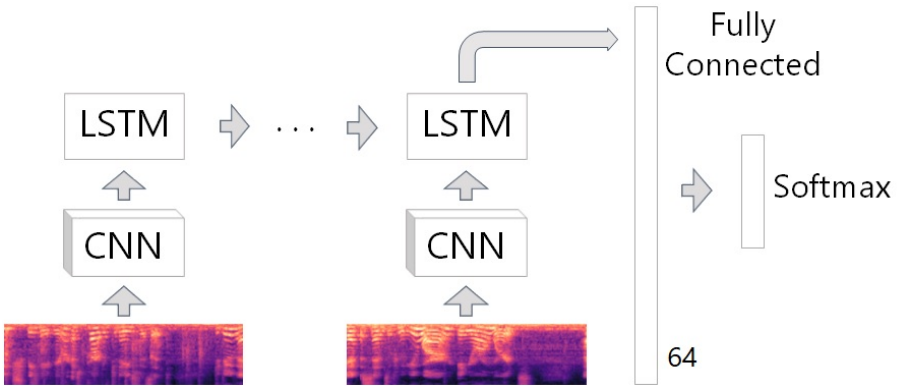


Figure 5.3: Complete Spectrogram CNN-LSTM architecture.

Complete architecture for the model Spectrogram CNN-LSTM. The CNN corresponds to the convolutional component explained in figure 5.2. The LSTM layer has a hidden state size of 64.

the OpenSMILE toolkit [11] from the eGeMAPS feature set before applying the functionals across time that the authors suggest. The configuration defined in OpenSMILE generates features with a step of 100 ms and a window of 600 ms. The features extracted with the OpenSMILE toolkit are 23 in total (number of features in brackets): (1) Pitch, (1) Jitter, (3) Formant 1, 2, and 3 frequency, (1) Formant 1 bandwidth, (1) Shimmer, (1) Loudness, (1) Harmonics-to-noise ratio (HNR), (1) Alpha Ratio, (1) Hammarberg Index, (2) Spectral Slope 0-500 Hz and 500-1500 Hz, (3) Formant 1, 2, and 3 relative energy, (1) Harmonic difference H1-H2, (1) Harmonic difference H1-A3, (4) MFCC 1-4, and (1) Spectral flux. After the extraction process, we have a sequence of frames for each of the samples that we use in the model architecture explained in the following section. The extracted features are then scaled using the min-max scaling technique. One sample/segment on this proposed approach corresponds to the eGeMAPS low-level descriptors extracted from 15 seconds of audio. Samples have a step of 1 second between each other.

Model Architecture This architecture is described in figure 5.4, and is exclusively used with the eGeMAPS features. It contains 3 parallel layers of strided 1-D convolutional layers. After applying the convolutional layers, we append their outputs and are send them into two fully connected layers with ReLU activation functions. The convolutional layers and the two linear layers have dropout [42] and a batch normalization [21] components. These components have varying parameters defined during the hyperparameter tuning

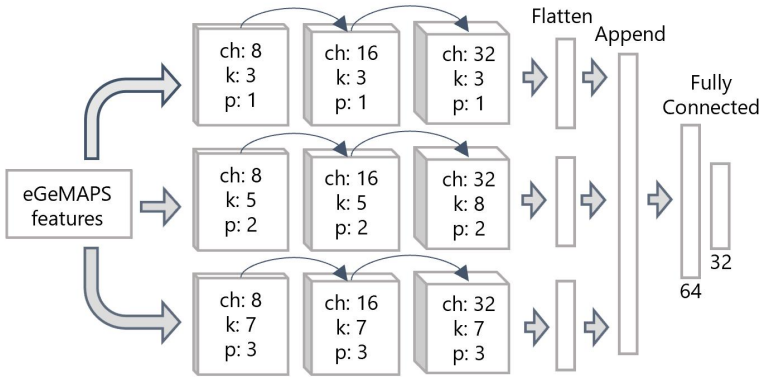


Figure 5.4: eGeMAPS CNN architecture.

Architecture of the eGeMAPS CNN. Each box represents a 1-D convolution across time, with a stride of 2. Fully connected layers have ReLU activation functions. (ch, k, p) represent the number of channels, kernel size, and padding, respectively. Every convolutional and linear layer has batch normalization and dropout.

stage. We add one last linear layer at the end of the network, which creates the logits that we use in the last activation function (softmax).

eGeMAPS CNN-LSTM

Features & Samples This architecture uses the eGeMAPS features as the input, as explained in the previous model. We define a sample as the eGeMAPS features extracted from 24 seconds of audio for this model. We split these features into 10 chunks, where each chunk contains the features corresponding to 15 seconds of audio. There is a step of 1 second between chunks.

Model Architecture The architecture proposed for this model is an extension of the eGeMAPS CNN. We include the main component from the eGeMAPS CNN, figure 5.4, as part of this model. Then, we use the output embeddings from that component as a sequence of data that is fed into an LSTM. This LSTM has dropout and a hidden state of size 64 (see figure 5.5 for more details). Finally, a fully connected layer receives the output from the last step of the LSTM and produces the logits processed by the softmax activation function.

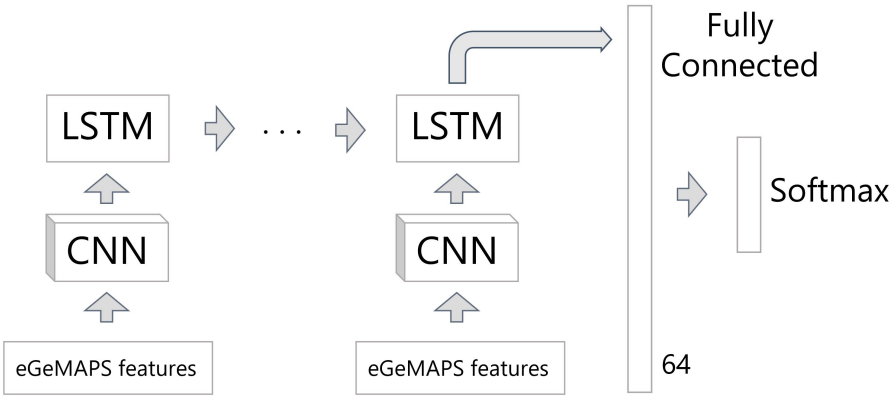


Figure 5.5: Complete eGeMAPS CNN-LSTM architecture.

Complete architecture for the model eGeMAPS CNN-LSTM. The CNN corresponds to the convolutional component explained in figure 5.4. The LSTM layer has a hidden state size of 64.

5.2.2 Voting Alternative

On section 5.2 we explained how we use majority voting for obtaining a recording level prediction. Here, we propose one alternative to obtain the track-level prediction as an additional solution to the recording-level prediction challenges from section 5.1.2. We will also refer to majority voting as hard voting, borrowing the naming convention from ensemble learning.

Soft Voting

As an alternative to majority voting or hard voting, we decided to use soft voting. The main difference is that, in soft voting, the probabilities of all the segments are averaged, and then a prediction is made. This naming convention (soft and hard voting) has been borrowed from ensemble learning where it is common when considering the predictions from multiple models to obtain a final prediction.

This is a brief example showing how the prediction from these two strategies might differ. In this example we consider a recording that has been split into 3 segments and we need to find out the prediction at the recording level, therefore we need to vote. We know the probability predictions for each of the segments, which are:

Segment 1: $P(\text{Depressed}) = 0.60$, $P(\text{Not Depressed}) = 0.40$

Segment 2: $P(\text{Depressed}) = 0.55$, $P(\text{Not Depressed}) = 0.45$

Segment 3: $P(\text{Depressed}) = 0.10$, $P(\text{Not Depressed}) = 0.90$

If we do a majority vote over the predictions of each segment, we first observe that Segment 1 is predicted as Depressed, Segment 2 is predicted as Depressed, and Segment 3 is predicted as Not Depressed. Thus, looking at the majority vote across these three segments, the track level prediction is Depressed. On the other hand, if we use the soft voting approach we find that the average of the predicted probabilities across the segments is $P(\text{Depressed}) = 0.417$, $P(\text{Not Depressed}) = 0.583$, by looking at this result we find ourselves with a prediction of Not Depressed at the track level. Thus, we are in a scenario where the prediction made by soft voting differs from that of majority voting (hard voting).

5.3 Other Approaches

This section briefly mentions other approaches that were considered and tested for tackling the challenges explained in the section 5.1. These approaches are not part of our final results since we decided not to pursue them during our initial stages.

Data Augmentation To mitigate the possible impact of data scarcity 5.1.4 we applied data augmentation. For those approaches using spectrograms, we implemented and tested those augmentations presented in the SpecAugment paper [33] due to its success in a variety of audio applications. Namely, those augmentations tested were frequency masking, time masking, and time warping. Additionally, we implemented the MixUp augmentation [51] for all our approaches. We tested these techniques on an earlier version of the AASS dataset. By then, the improvements observed were little, and, thus, we decided not to pursue this approach any further.

Noise Robust loss functions We hypothesized that having the same label for all the segments in a recording might cause the problem of noisy labels. The reasoning behind this is that even if a person is Depressed, they might not show strong and consistent acoustic signals of depression across the recording. Thus, some of the segments should ideally be labeled differently. The loss function tested for this was Generalized Crossentropy [52]. Similar to data augmentation, we tested this loss function on a previous

version of the AASS dataset. Preliminary results did not show any major performance improvement.

Class weights To address the data imbalance problem left for some of the items of the depression assessment scales, after applying the definition of the IIDD task, we tested the usage of class weights in the loss functions used for training the models. Once again, we tested class weights on an earlier version of the AASS dataset, and the preliminary results did not show significant improvements.

Variable depression & voting thresholds This was another alternative to majority voting that we tested to solve the challenge of obtaining recording-level predictions 5.1.2. The idea behind this strategy is to select the best combination for the depression and voting thresholds. The depression threshold defines the probability required for a segment to be labeled as 'Depressed' instead of selecting the highest probability. The voting threshold determines the proportion of the segments required to vote 'Depressed' to predict 'Depressed' at the recording level. In this scenario, we can represent the majority vote as having a depression threshold of 0.5 (i.e., When the predicted probability of 'Depressed' of a segment is 0.5 or more the segment is predicted to be 'Depressed') and a voting threshold of 0.5 (i.e., The prediction for the recording is 'Depressed' if 50% or more of the total segments are predicted as 'Depressed'). An example of the F-Score surface created by varying both thresholds is available in figure 5.6. We observed from this approach that even though it was possible to perform better than the majority vote in some scenarios, on average, the majority vote has a more consistent behavior. In contrast, it was common to choose a voting threshold and a depression threshold that caused more harm than benefit due to overfitting.

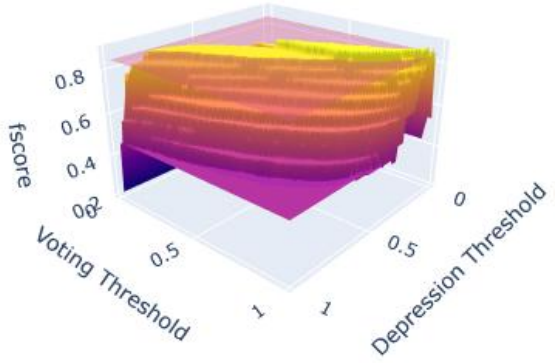


Figure 5.6: F1-Score when varying depression and voting threshold.

The plane at F1-Score equal to 0.8 corresponds to the F1-Score obtained by the majority voting technique. It can be observed how the F1-Score varies with changes on the thresholds. In this example, The figure also shows that the majority vote technique does not achieve the highest F1 score.

Chapter 6

Evaluation

6.1 Training details

As part of our experiments, we trained the approaches presented in section 5.2 on the task of IIDD for each of the items of the depression assessment instruments for datasets AASS and DAIC-WOZ. All models were trained using the Adam optimizer [23] for 25 epochs, with parameters β_1:0.9, β_2: 0.999, α: 0.0005. We also used *L2* regularization [26] in all our experiments. Additionally, we applied a random search of hyperparameters for deciding on the exact parameters of the batch normalization [21] and dropout [42] components in the layers of the proposed models. We used the random seed 1234 across all our experiments for replicability purposes.

6.2 Results

After training the models, we obtain track-level predictions using the majority vote approach explained at the start of section 5.2. We will explore these results in each of the datasets separately in the upcoming sections.

6.2.1 Models

Autobiographical Adult Speech Samples (AASS)

The results of our approaches, when trained on the AASS dataset, can be observed in table 6.1. At first glance, we can see that the performance across the different depression items (IIDD task) is varied. Moreover, no single 'optimal' approach/model outperforms all the other approaches when looking at the F1-score.

When looking at individual items, those where our acoustic-based approaches can more effectively detect depressive symptoms are (8) Inability to feel (F1-score of 0.8), (9) Pessimistic thoughts (F1-score of 0.81), and (10) Suicidal thoughts (F1-score of 0.9),

Table 6.1: Track level F1 Scores of the models trained on each of the individual items of the AASS dataset. Each cell contains metrics in the format: Weighted F1 Score / F1 Score (ND) / F1 Score (D). Last column corresponds to a dummy predictor that always predicts the most common class. Numbers in bold indicate the highest weighted F1-scores obtained for each of the depression items.

	AASS				
MADRS Item / Question	Spectrogram CNN	Spectrogram CNN LSTM	eGeMAPS CNN	eGeMAPS CNN LSTM	Dummy Predictor
(1) Apparent sadness	0.36/0.32/0.43	0.61/0.71/0.43	0.66/0.74/0.53	**0.76**/0.81/0.67	0.47/0.76/0.00
(2) Reported sadness	0.62/0.60/0.64	**0.71**/0.73/0.70	0.61/0.67/0.56	**0.71**/0.75/0.67	0.36/0.00/0.69
(3) Inner tension	0.67/0.63/0.70	0.65/0.53/0.74	**0.69**/0.57/0.79	**0.69**/0.57/0.79	0.42/0.00/0.73
(4) Reduced sleep	0.67/0.67/0.67	**0.76**/0.76/0.76	0.74/0.83/0.62	0.75/0.81/0.67	0.42/0.73/0.00
(5) Reduced appetite	0.49/0.65/0.00	0.65/0.79/0.22	0.62/0.75/0.20	0.55/0.73/0.00	**0.66**/0.86/0.00
(6) Conc. difficulties	0.41/0.27/0.59	0.66/0.72/0.59	0.63/0.76/0.46	**0.71**/0.75/0.67	0.42/0.73/0.00
(7) Lassitude	0.33/0.46/0.13	0.42/0.40/0.45	0.57/0.67/0.40	**0.67**/0.72/0.59	0.47/0.76/0.00
(8) Inability to feel	0.67/0.72/0.59	0.52/0.62/0.38	**0.80**/0.86/0.71	0.65/0.76/0.46	0.47/0.76/0.00
(9) Pessimistic thoughts	0.76/0.78/0.74	0.62/0.60/0.64	**0.81**/0.82/0.80	**0.81**/0.82/0.80	0.36/0.00/0.69
(10) Suicidal thoughts	0.70/0.74/0.53	0.75/0.86/0.29	**0.90**/0.94/0.75	0.62/0.69/0.31	0.72/0.89/0.00

scores by eGeMAPS CNN and (9) also achieved by eGeMAPS CNN LSTM. In general, there is a clear improvement for most of the items when comparing our approaches against the dummy predictor, except for item (5) Reduced appetite, where none of the models achieve an increase in the weighted F1-score.

Distress Analysis Interview Corpus - Wizard of OZ (DAIC-WOZ)

The results for the DAICWOZ were also obtained and are available in table 6.2. Overall, we can observe that the results vary across the depression items. It is also noticeable that the samples on this dataset represent a bigger challenge when compared to the previous dataset, having, on average, lower performances according to the F1-score. However, when compared to the dummy predictor, our models increase the performance throughout all the items of the PHQ-8 questionnaire.

6.2.2 Voting Alternatives

Soft Voting

After applying soft voting on the results obtained when training the model, as explained in section 5.2.2, we obtain the results that we show in this section.

Table 6.2: Track level F1-Scores of the models trained on each of the individual items of the DAIC-WOZ dataset. Each cell contains metrics in the format: Weighted F1-Score / F1-Score (ND) / F1-Score (D). Last column corresponds a dummy predictor that always predicts the most common class.

	DAICWOZ				
PHQ-8 Item / Question	Spectrogram CNN	Spectrogram CNN LSTM	eGeMAPS CNN	eGeMAPS CNN LSTM	Dummy Predictor
(1) Little interest	0.51/0.54/0.49	0.44/0.47/0.41	0.54/0.60/0.49	**0.59**/0.60/0.58	0.37/0.00/0.70
(2) Feeling down	0.45/0.35/0.55	0.47/0.38/0.55	**0.61**/0.51/0.70	0.43/0.38/0.47	0.37/0.00/0.70
(3) Trouble sleeping	0.49/0.27/0.64	**0.55**/0.47/0.62	0.45/0.22/0.62	0.51/0.32/0.65	0.44/0.00/0.74
(4) Feeling tired	**0.64**/0.34/0.77	0.59/0.22/0.75	0.56/0.09/0.76	0.53/0.14/0.70	0.57/0.00/0.82
(5) Poor appetite	**0.55**/0.26/0.72	0.46/0.39/0.49	0.48/0.29/0.59	0.54/0.43/0.61	0.50/0.00/0.78
(6) Self-disappointment	0.44/0.18/0.66	**0.54**/0.44/0.63	0.46/0.40/0.52	0.42/0.13/0.67	0.37/0.00/0.70
(7) Conc. difficulties	0.52/0.57/0.45	0.52/0.51/0.53	**0.60**/0.69/0.48	0.57/0.68/0.44	0.42/0.73/0.00
(8) Slow/Agitated	0.61/0.75/0.28	**0.68**/0.80/0.37	0.66/0.79/0.36	0.61/0.77/0.23	0.60/0.83/0.00

Autobiographical Adult Speech Samples (AASS): Table 6.3 contains the results for the AASS when calculating the predictions using the soft voting alternative. There are some changes when compared with the results obtained previously with the majority/hard vote. Figures 6.1 and 6.2 contain visualizations that make the comparison between voting schemes and approaches easier. We can not claim that soft voting is a better alternative since sometimes it improves the F1 score, and sometimes it does not. We want to highlight the performance of this voting alternative in one item where there were positive changes on the F1 score. Item (9) Pessimistic thoughts obtains the highest weighted F1-score using the eGeMAPS CNN approach with a value of 0.86, an improvement over the previous score of 0.81 obtained with hard voting. Additionally, there are some examples where there is an improvement within the results of a model. An example is the behavior of question (6) Concentration difficulties, where a weighted F1-score of 0.69 is obtained with soft voting using the eGeMAPS CNN approach, in contrast to the score obtained with hard voting (0.63).

Distress Analysis Interview Corpus - Wizard of OZ (DAIC-WOZ): Table 6.4 shows the results of this alternative. Like the patterns evidenced on the AASS dataset, soft voting obtains a better performance in some occasions. There are no clear improvements in the F1 score across the questions of this questionnaire. Importantly, soft voting manages to obtain the highest F1 scores across multiple items, such as (1) Little interest (eGeMAPS CNN LSTM), (3) Trouble sleeping (Spectrogram CNN and Spectrogram CNN LSTM), and

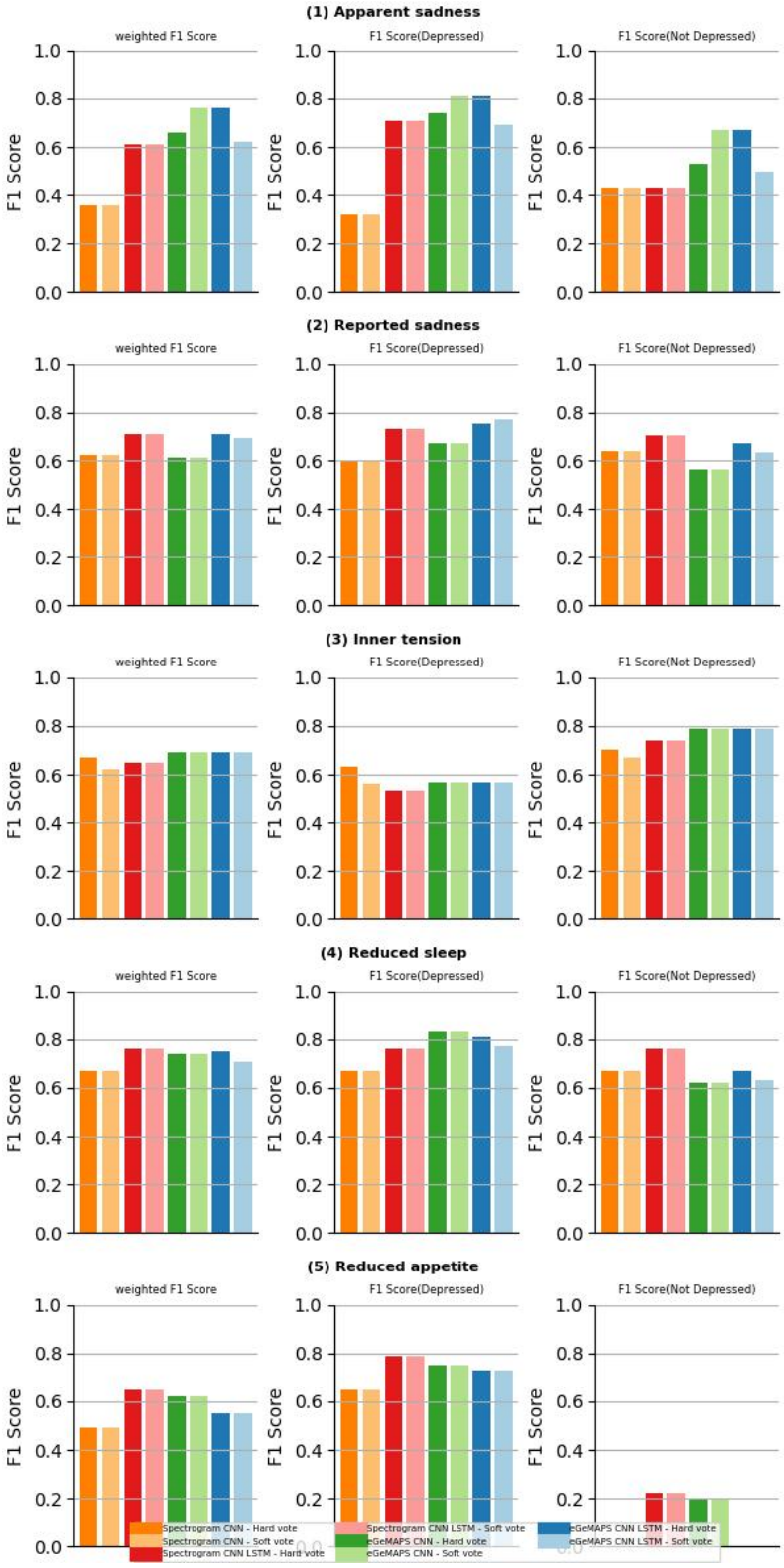


Figure 6.1: AASS F1 Score comparison for hard and soft voting, questions 1 to 5.

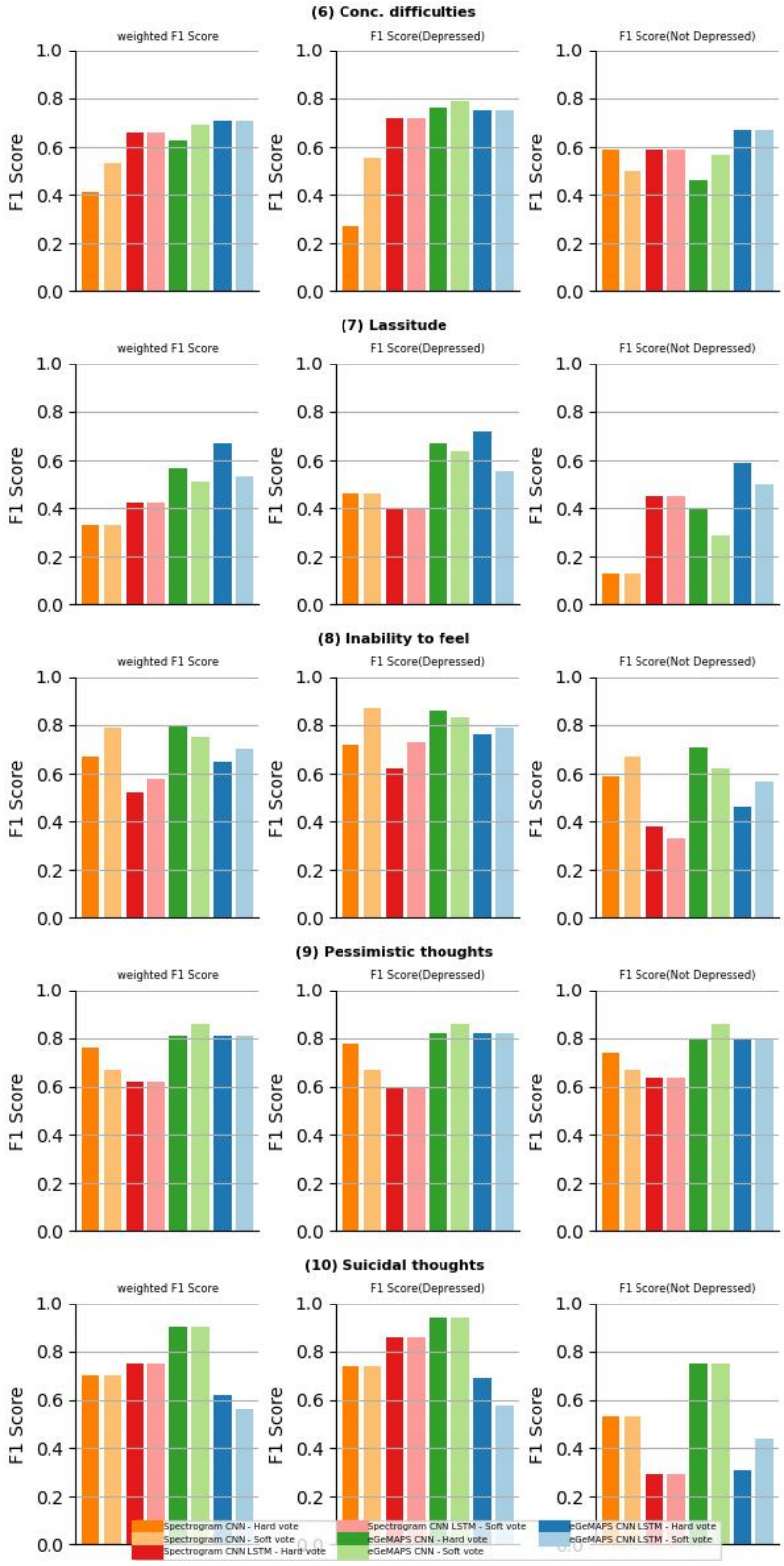


Figure 6.2: AASS F1 Score comparison for hard and soft voting, questions 6 to 10.

Table 6.3: Track level F1-Scores of the models trained on each of the individual items of the AASS dataset when using soft voting. Each cell contains metrics in the format: Weighted F1-Score / F1-Score (ND) / F1-Score (D).

AASS				
MADRS Item / Question	Spectrogram CNN	Spectrogram CNN LSTM	eGeMAPS CNN	eGeMAPS CNN LSTM
(1) Apparent sadness	0.36/0.32/0.43	0.61/0.71/0.43	0.76/0.81/0.67	0.62/0.69/0.50
(2) Reported sadness	0.62/0.60/0.64	0.71/0.73/0.70	0.61/0.67/0.56	0.69/0.77/0.63
(3) Inner tension	0.62/0.56/0.67	0.65/0.53/0.74	0.69/0.57/0.79	0.69/0.57/0.79
(4) Reduced sleep	0.67/0.67/0.67	0.76/0.76/0.76	0.74/0.83/0.62	0.71/0.77/0.63
(5) Reduced appetite	0.49/0.65/0.00	0.65/0.79/0.22	0.62/0.75/0.20	0.55/0.73/0.00
(6) Conc. difficulties	0.53/0.55/0.50	0.66/0.72/0.59	0.69/0.79/0.57	0.71/0.75/0.67
(7) Lassitude	0.33/0.46/0.13	0.42/0.40/0.45	0.51/0.64/0.29	0.53/0.55/0.50
(8) Inability to feel	0.79/0.87/0.67	0.58/0.73/0.33	0.75/0.83/0.62	0.70/0.79/0.57
(9) Pessimistic thoughts	0.67/0.67/0.67	0.62/0.60/0.64	0.86/0.86/0.86	0.81/0.82/0.80
(10) Suicidal thoughts	0.70/0.74/0.53	0.75/0.86/0.29	0.90/0.94/0.75	0.56/0.58/0.44

(4) feeling tired (Spectrogram CNN). Once again, there are changes in the performance across multiple items indicating that the tested voting schemes can obtain different results.

Table 6.4: Track level F1-Scores of the models trained on each of the individual items of the DAIC-WOZ dataset when using soft voting. Each cell contains metrics in the format: Weighted F1-Score / F1-Score (ND) / F1-Score (D).

DAICWOZ				
PHQ-8 Item / Question	Spectrogram CNN	Spectrogram CNN LSTM	eGeMAPS CNN	eGeMAPS CNN LSTM
(1) Little interest	0.53/0.55/0.52	0.55/0.38/0.69	0.53/0.63/0.44	0.59/0.61/0.57
(2) Feeling down	0.49/0.55/0.44	0.47/0.38/0.55	0.49/0.34/0.62	0.48/0.47/0.49
(3) Trouble sleeping	0.55/0.44/0.63	0.55/0.47/0.62	0.51/0.37/0.61	0.50/0.31/0.63
(4) Feeling tired	0.64/0.34/0.77	0.59/0.22/0.75	0.56/0.09/0.76	0.58/0.10/0.79
(5) Poor apetite	0.51/0.14/0.71	0.46/0.39/0.49	0.46/0.25/0.58	0.47/0.42/0.50
(6) Self-disappointment	0.44/0.18/0.66	0.50/0.44/0.55	0.46/0.40/0.52	0.42/0.13/0.67
(7) Conc. difficulties	0.59/0.61/0.57	0.52/0.51/0.53	0.55/0.60/0.49	0.55/0.67/0.40
(8) Slow/Agitated	0.60/0.77/0.17	0.68/0.80/0.37	0.64/0.78/0.30	0.60/0.75/0.22

Understanding the improvements: To understand how can the change from majority/hard vote to soft vote brings an improvement, we have found an example where

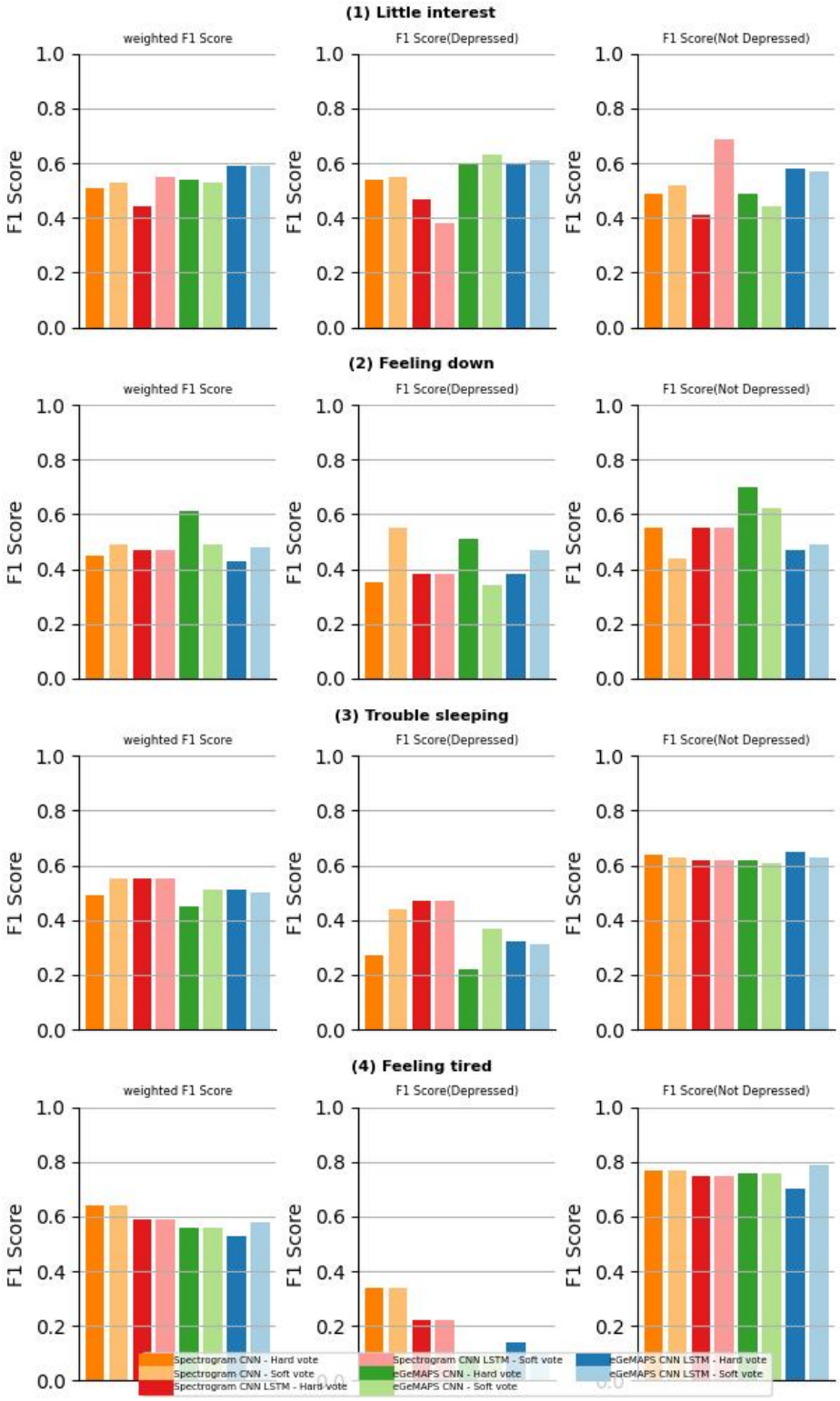


Figure 6.3: DAIC-WOZ F1 Score comparison for hard and soft voting, questions 1 to 4.

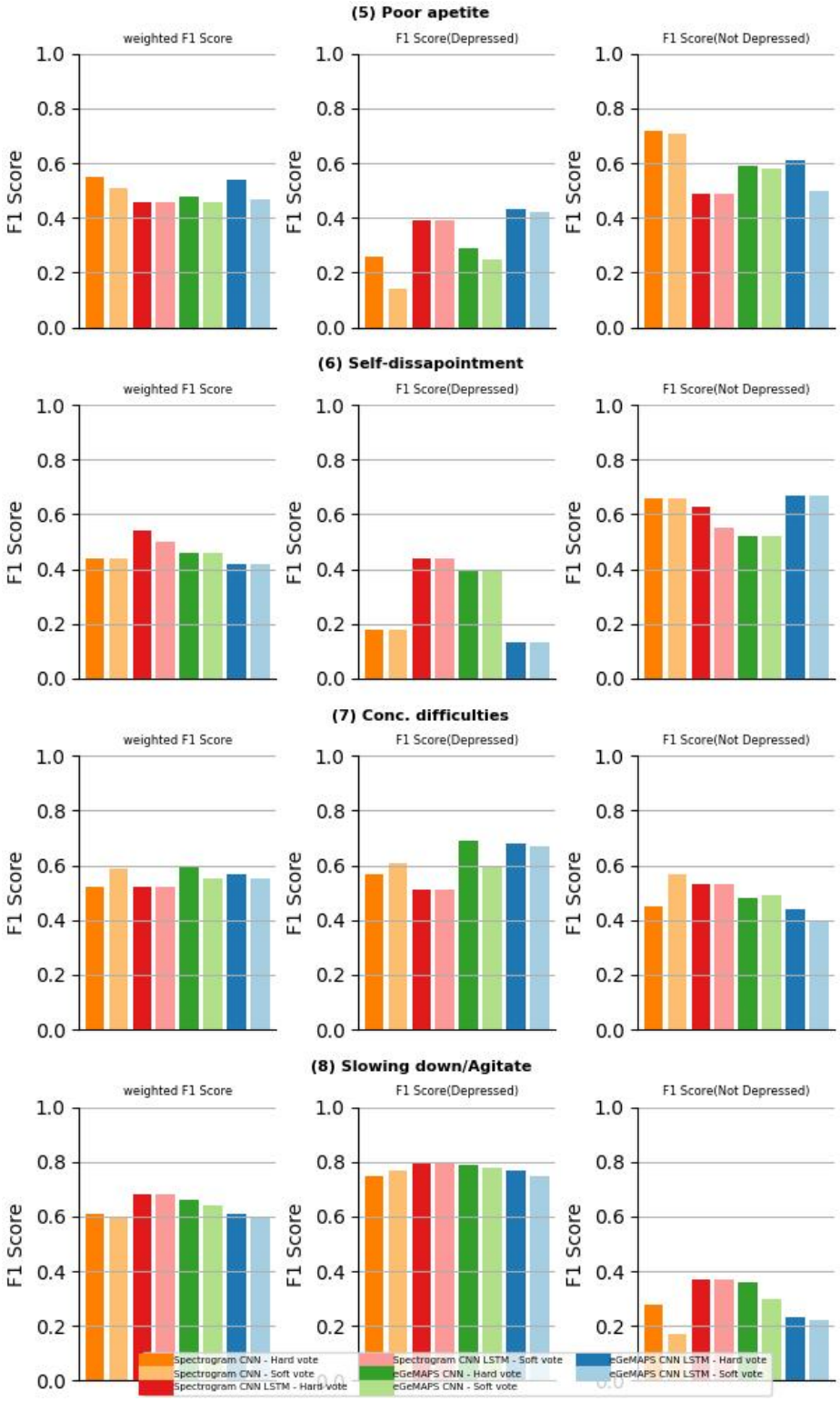


Figure 6.4: DAIC-WOZ F1 Score comparison for hard and soft voting, questions 5 to 8.

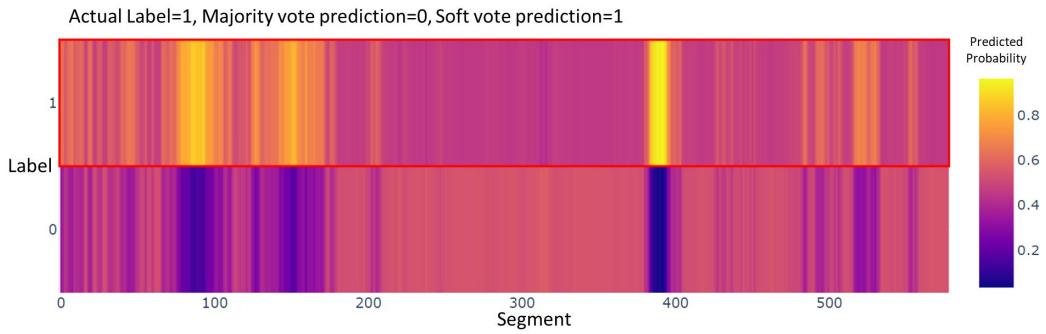


Figure 6.5: Example of the prediction probabilities of a recording (sample) labelled differently when using majority vote and soft vote for item (3) Inner tension of MADRS on the eGeMAPS CNN model.

Probabilities of predicting 1 (Depressed) and 0 (Not Depressed). In this example, is it clear that the predicted probabilities across most of the segments are close to 0.5, however, some specific segments show signals of Depression and obtain a high predicted probability for that class. However, the final prediction using majority/hard vote is wrong (0 - Not Depressed) and using soft vote is correct (1 - Depressed)

the prediction changes between voting schemes. Figure 6.5 contains the visual representation of the predicted probabilities of the segments of a recording. In the figure, the highest predicted probabilities belong to label 1 (Depressed). However, in most of the recording segments, the model is not that confident when it predicts and estimates probabilities close to 0.5. When applying the majority/hard vote, the predicted class is 0 (Not Depressed). When using the soft vote the prediction is 1 (Depressed). The soft vote strategy makes the correct prediction due to the influence of higher predicted probability segments in the 1 (Depressed) class.

The change from majority voting to soft voting does not always bring improvements. Figure 6.6 shows another example of a recording with conflicting predictions between majority/hard vote and soft voting. This time around, the model predicts much more confidently (higher probabilities), even when the predictions are incorrect. After applying the majority/hard voting across the segments, the prediction turns out to be 1 (Depressed), while soft voting predicts 0 (Not Depressed). In this case, the majority/hard vote obtains the correct prediction.

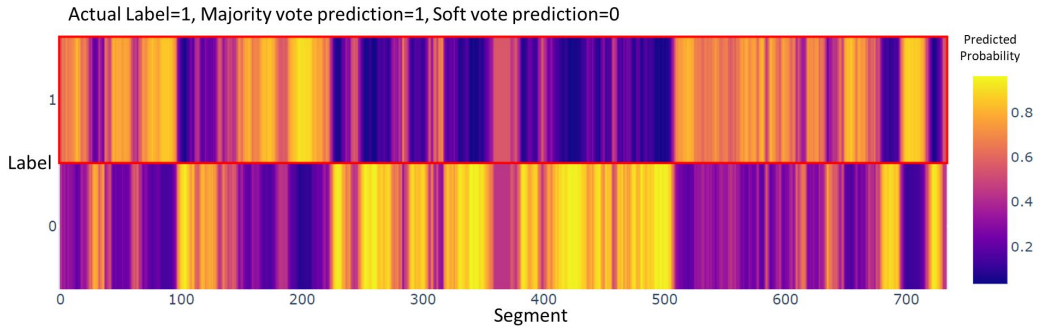


Figure 6.6: Example of the prediction probabilities of a recording (sample) labelled differently when using majority vote and soft vote for item (1) Apparent sadness of MADRS on the eGeMAPS CNN model.

Probabilities of predicting 1 (Depressed) and 0 (Not Depressed). In this example, the model confidently predicts that some segments have a high propability of No Depressed, while others get predicted a high probability of Depressed. In this scenario, the final prediction using majority/hard vote is correct (1 - Depressed) and using soft vote is wrong (0 - Not Depressed)

Chapter 7

Conclusion

We have introduced and explored a new task in the field of automated depression assessment from speech, where the aim is to automatically predict the presence of individual depressive symptoms as defined by clinical depression estimation instruments. We do this by using neural network architectures whose inputs are the acoustic features of speech signals, i.e., paralinguistic characteristics of speech.

Our experiments show interesting and varied results when looking at individual items from clinician-rated and self-rated depression assessment instruments. From our results, we observed that each of the proposed approaches has its strengths and weaknesses. In summary, most of our methods obtain better performance than an empirical majority-based baseline on almost all the items of both depression assessment tools. In general, specific items from questionnaires, such as inability to feel, pessimistic thoughts, and suicidal thoughts, achieved better performances overall with one of our model architectures. Other items, such as reduced appetite, were not easily detectable with our experiments and are, most likely, difficult to predict using acoustic features from speech.

Throughout our approaches, we also tested two different alternatives for aggregating segment-level predictions into recording-level predictions on each of the symptoms. As part of our results, we show that both strategies, majority/hard voting, and soft voting, obtain similar performance. After analyzing specific examples, it was clear that their results are inherently bound to perform differently depending on the segment-level predictions.

7.1 Future work

Depressive disorders are and will continue to be a challenge for humanity. Our work contributes to the early detection of depressive symptoms, as defined in depression assessment questionnaires. However, it also enables opportunities for future progress.

Some areas to explore are:

- Extend the proposed binary classification task into more categories that bring more information about each specific item. For example, categories might indicate mild, moderate, or severe presence of a symptom.

- Use verbal and video features for the proposed task of depression detection on individual items from depression assessment instruments. Multimodal approaches can clarify the complementary (or substitutable) nature of speech acoustics, verbal content, and video for detailed depressive assessment while improving the results obtained in this work.

- Use the predictions obtained for the items of the questionnaires as an intermediate step that can help improve depression detection or depression severity estimation as a whole while still having the added benefits of explainability brought by predicting all the items in the depression assessment instruments.

- Try more alternatives for the process of aggregating segment-level predictions into track-level predictions. Using majority/hard voting or soft voting has shown to be a good approach. However, it is a simple method that, when replaced, might bring improvements in the overall results.

- Predicting the individual scores obtained for self-rated and clinician-rated depression assessment scales for samples assessed with both instruments. Doing so would give a better understanding of which type of scales are more predictable on the same samples. Results from such an approach would provide a good comparison between the performance on each scale. Also, this approach would provide useful information about the presence of more symptoms.

Bibliography

[1] Tuka Alhanai, Mohammad M Ghassemi, and James R Glass. Detecting depression with audio/text sequence modeling of interviews. In *Interspeech*, pages 1716–1720, 2018.

[2] Jont B. Allen and Lawrence R. Rabiner. A unified approach to short-time fourier analysis and synthesis. *Proceedings of the IEEE*, 65(11):1558–1564, 1977.

[3] Alina Arseniev-Koehler, Sharon Mozgai, and Stefan Scherer. What type of happiness are you looking for? - a closer look at detecting mental health from language. pages 1–12, 01 2018.

[4] A. T. Beck, C. H. Ward, M. Mendelson, J. Mock, and J. Erbaugh. An Inventory for Measuring Depression. *Archives of General Psychiatry*, 4(6):561–571, 06 1961.

[5] Christopher Olah. Understanding lstm networks. `http://colah.github.io/posts/2015-08-Understanding-LSTMs/`, 2015.

[6] COVID-19 Mental Disorders Collaborators. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *The Lancet*, 396(10258):1204–1222, 2020.

[7] Nicholas Cummins, Jyoti Joshi, Abhinav Dhall, Vidhyasaharan Sethu, Roland Goecke, and Julien Epps. Diagnosis of depression by behavioural signals: A multimodal approach. In *AVEC 2013 - Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, AVEC 2013 - Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, pages 11–20, United States of America, January 2013. Association for Computing Machinery (ACM). ACM International Workshop on Audio/Visual Emotion Challenge, AVEC 2013 ; Conference date: 21-10-2013 Through 21-10-2013.

[8] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep: A collaborative voice analysis repository for speech technologies. 05 2014.

[9] S. Pavankumar Dubagunta, Bogdan Vlasenko, and Mathew Magimai.-Doss. Learning voice source related information for depression detection. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6525–6529, 2019.

[10] Sri Harsha Dumpala, Sheri Rempel, Katerina Dikaios, Mehri Sajjadian, Rudolf Uher, and Sageev Oore. Estimating severity of depression from acoustic features and embeddings of natural speech. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7278–7282. IEEE, 2021.

[11] F. Eyben, M. Wöllmer, and B. Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proc. ACM conference on Multimedia*, pages 1459–1462, 2010.

[12] Florian Eyben, Klaus Scherer, Björn Schuller, Johan Sundberg, Elisabeth Andre, Carlos Busso, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth Narayanan, and Khiet Truong. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7:1–1, 01 2015.

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

[14] Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[15] Max Hamilton. A rating scale for depression. *Journal of Neurology, Neurosurgery & Psychiatry*, 23(1):56–62, 1960.

[16] C.J. Hawley, T.M. Gale, and T. Sivakumaran. Defining remission by cut off score on the madrs: selecting the optimal value. *Journal of Affective Disorders*, 72(2):177–184, 2002.

[17] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.

[18] Rachelle Horwitz, Thomas Quatieri, Brian Helfer, Bea Yu, James Williamson, and James Mundt. On the relative importance of vocal source, system, and prosody in human depression. pages 1–6, 05 2013.

[19] Zhaocheng Huang, Julien Epps, and D Joachim. Exploiting vocal tract coordination using dilated cnns for depression detection in naturalistic environments. In *ICASSP*, pages 6549–6553. IEEE, 2020.

[20] Institute of Health Metrics and Evaluation. Global burden of disease study results tool. `http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/d780dffbe8a381b25e1416884959e88b`, 2019.

[21] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015.

[22] John F. Kelley. An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.*, 2(1):26–41, January 1984.

[23] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[24] W. Koenig, H. K. Dunn, and L. Y. Lacy. The sound Spectrograph. *Journal of the Acoustical Socierty of America*, 18(1):19–33, Jul 1946.

[25] Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B.W. Williams, Joyce T. Berry, and Ali H. Mokdad. The phq-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, 114(1):163–173, 2009.

[26] Anders Krogh and John A. Hertz. A simple weight decay can improve generalization. In *NIPS*, 1991.

[27] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

[28] Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity*. *Bulletin of Mathematical Biology*, 52(1):99–115, 1990.

[29] Hongying Meng, Di Huang, Heng Wang, Hongyu Yang, Mohammed AI-Shuraifi, and Yunhong Wang. Depression recognition based on dynamic facial and vocal expression features using partial least square regression. In *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, AVEC '13, page 21–30, New York, NY, USA, 2013. Association for Computing Machinery.

[30] Stuart A. Montgomery and Marie Åsberg. A new depression scale designed to be sensitive to change. *British Journal of Psychiatry*, 134(4):382–389, 1979.

[31] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. World: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, E99.D(7):1877–1884, 2016.

[32] Alan V. Oppenheim. Speech spectrograms using the fast fourier transform. *Spectrum, IEEE*, 7:57 – 62, 08 1970.

[33] Daniel Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin Cubuk, and Quoc Le. Specaugment: A simple data augmentation method for automatic speech recognition. pages 2613–2617, 09 2019.

[34] Srinivas Parthasarathy and Ivan Tashev. Convolutional neural network techniques for speech emotion recognition. 09 2018.

[35] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, and Maja Pantic. Avec'19: Audio/visual emotion challenge and workshop. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 2718–2719, New York, NY, USA, 2019. Association for Computing Machinery.

[36] A.John Rush, Madhukar H Trivedi, Hicham M Ibrahim, Thomas J Carmody, Bruce Arnow, Daniel N Klein, John C Markowitz, Philip T Ninan, Susan Kornstein, Rachel Manber, Michael E Thase, James H Kocsis, and Martin B Keller. The 16-item quick inventory of depressive symptomatology (qids), clinician rating (qids-c), and self-report (qids-sr): a psychometric evaluation in patients with chronic major depression. *Biological Psychiatry*, 54(5):573–583, 2003.

[37] Björn Schuller, Stefan Steidl, and Anton Batliner. The interspeech 2009 emotion challenge. pages 312–315, 01 2009.

[38] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[39] Alex Sherstinsky. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *CoRR*, abs/1808.03314, 2018.

[40] Malcolm Slaney. Auditory toolbox: A matlab toolbox for auditory modeling work. Technical report, Interval Research Corporation, 1998.

[41] Sound eXchange project contributors. SoX - Sound eXchange. `http://sox.sourceforge.net/`, 2015.

[42] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958, January 2014.

[43] S. S. Stevens, J. Volkmann, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America*, 8:185–190, 1937.

[44] Fuxiang Tao, Anna Esposito, and Alessandro Vinciarelli. Spotting the traces of depression in read speech: An approach based on computational paralinguistics and social signal processing. *Proc. Interspeech 2020*, pages 1828–1832, 2020.

[45] Andrea Trevino, Thomas Quatieri, and Nicolas Malyska. Phonologically-based biomarkers for major depressive disorder. *EURASIP Journal on Advances in Signal Processing*, 2011, 08 2011.

[46] Michel Valstar, Björn Schuller, Kirsty Smith, Florian Eyben, Bihan Jiang, Sanjay Bilakhia, Sebastian Schnieder, Roddy Cowie, and Maja Pantic. Avec 2013 - the continuous audio/visual emotion and depression recognition challenge. pages 3–10, 10 2013.

[47] Michel F. Valstar, Jonathan Gratch, Björn W. Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. AVEC 2016 - depression, mood, and emotion recognition workshop and challenge. *CoRR*, abs/1605.01600, 2016.

[48] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification, 2020.

[49] James Williamson, Thomas Quatieri, Brian Helfer, Rachelle Horwitz, Bea Yu, and Daryush Mehta. Vocal biomarkers of depression based on motor incoordination. *AVEC 2013 - Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, 10 2013.

[50] Le Yang, Dongmei Jiang, Lang He, Ercheng Pei, Meshia Cédric Oveneke, and Hichem Sahli. Decision tree based depression classification from audio video and language information. In *Proc. ACM workshop on Audio/visual emotion challenge*, pages 89–96, 2016.

[51] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *CoRR*, abs/1710.09412, 2017.

[52] Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels, 2018.

[53] Ziping Zhao, Zhongtian Bao, Zixing Zhang, Jun Deng, Nicholas Cummins, Haishuai Wang, Jianhua Tao, and Björn Schuller. Automatic assessment of depression from speech via a hierarchical attention transfer network and attention autoencoders. *IEEE Journal of Selected Topics in Signal Processing*, PP:1–1, 11 2019.

[54] Ziping Zhao, Qifei Li, Nicholas Cummins, Bin Liu, Haishuai Wang, Jianhua Tao, and Björn W. Schuller. Hybrid Network Feature Extraction for Depression Assessment from Speech. In *Proc. Interspeech 2020*, pages 4956–4960, 2020.

[55] Mark Zimmerman, Michael A. Posternak, and Iwona Chelminski. Defining remission on the Montgomery-Asberg depression rating scale. *Journal of Clinical Psychiatry*, 65(2):163–168, Feb 2004.