Utrecht University

Auditdienst Rijk
Ministerie van Financiën

MASTER THESIS

Speech-based Depression Prediction with Symptoms as Interpretable Intermediate Features

Author:
Floris van Steijn

First / Daily Supervisor:
Dr. Heysem Kaya

Daily Supervisor:
Gizem Soğancıoğlu

External Supervisor:
Fré Vink

Second Examiner:
Dr. Aleksei Nazarov

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

June 27, 2022

Abstract

Speech-based Depression Prediction with Symptoms as Interpretable Intermediate Features

by Floris van Steijn

Mood disorders in general and depression in particular are common and their impact on individuals and society is high. Roughly 5% of adults worldwide suffer from depression. Commonly, depression diagnosis involves using questionnaires, either clinician-rated or self-reported. Due to the subjectivity in questionnaire methods and high human-related costs involved, there are ongoing efforts to find more objective and easily attainable depression markers. As is the case with recent audio, visual and linguistic applications, state-of-the-art approaches for automated depression severity prediction heavily depend on deep learning and black box modeling without explainability and interpretability considerations. However, for reasons ranging from regulations to understanding the extent and limitations of the model, the clinicians need to understand the decision making process of the model to confidently form their decisions. In this work, I focus on speech-based depression severity level prediction on DAIC-WOZ corpus and benefit from PHQ-8 questionnaire items to predict the symptoms as interpretable high level features. I show that using a multi-task regression approach with state-of-the-art text-based features to predict the depression symptoms, it is possible to reach a viable test set Concordance Correlation Coefficient performance comparable to the state-of-the-art systems, while improving on the interpretability of the overall prediction system.

Acknowledgements

First an foremost, I would like to thank Heysem for his extensive supervision of this thesis project. He has always been able to push the quality of my work, while being a very supportive supervisor. I would also like to thank Gizem for her support during this thesis, both in terms of content as well as practical advice. I have enjoyed working together, and have especially learned a lot when working on the paper. Next, I would like to thank Fré for his all his valuable advice, the pleasant experience at the Auditdienst Rijk, as well as for the fun and interesting discussions about explainability, interpretability and other topics. Finally, I want to thank Aleksei for his interest in the project, his positive feedback and the constructive remarks.

Contents

List of Figures

List of Tables

Acronyms

BERT Bidirectional Encoder Representations from Transformers.

BLSTM Bidirectional Long Short-Term Memory.

CCC Concordance Correlation Coefficient.

CNN Convolutional Neural Network.

DAIC-WOZ Distress Analysis Interview Corpus Wizard of Oz.

E-DAIC Extended Distress Analysis Interview Corpus.

EBM Explainable Boosting Machine.

Grad-CAM Gradient-weighted Class Activation Mapping.

GRU Gated Recurrent Unit.

KELM Kernel Extreme Learning Machine.

LIME Local Interpretable Model-Agnostic Explanations.

LIWC Linguistic Inquiry and Word Count.

LLDs Low-Level Descriptors.

MFCC Mel-Frequency Cepstral Coefficients.

ML Machine Learning.

PDP Partial Dependence Plots.

PHQ-8 Eight Item Patient Health Questionnaire.

RMSE Root Mean Square Error.

S-BERT Sentence BERT.

SHAP SHapley Additive exPlanations.

SoA State of the Art.

SVM Support Vector Machine.

UAR Unweighted Average Recall.

Chapter 1

Introduction

This thesis aims to contribute to the field of automated depression diagnosis. In Section 1.1, I will formulate the problem, and discuss how an interpretable approach driven by Machine Learning (ML) could contribute to improvement of the diagnostic practice. In Section 1.2, I discuss the objectives of this thesis, and I will provide the thesis outline in Section 1.3.

1.1 Problem statement

Mood disorders in general and depression in particular are common and their effect on individuals and society is large. Roughly 5% of adults worldwide suffer from depression [8]. In the Netherlands, the percentage of adults with depressive symptoms have increased from 40% to 48% between 2012 and 2020, with roughly 15% of this group having severe symptoms [57]. Especially among young adults, prevalence of depression is high. For example, according to very recent research, which is done in the context of the recent COVID-19 pandemic, 26% of students in the Netherlands have had death wishes, which is highly indicative for major depression [12]. The same study found that 51% of students suffer from psychological problems.

Commonly, depression diagnosis involves using questionnaires, either clinician-rated or self-reported. Interviews or questionnaire methods can be subjective, which potentially induces bias in diagnosis [39]. Therefore, there is an effort among researchers to find more objective depression markers, and automated depression diagnostic methods have gained increasing attention. Since mood disorders often manifest in behavioural cues, using speech expressions as objective markers is one promising field of research [5].

A popular way of using speech expressions as predictors for depression is using ML. One of the common approaches is to develop a multi-modal deep learning model that processes the acoustic and linguistic features of speech, sometimes also accompanied by visual features. Deep learning techniques are capable of learning almost any interaction between input features, making them very flexible predictors. In case of depression prediction from speech, this is a very desirable property,

because behaviour is often represented by a large amount of features which can interact in complex nonlinear ways. However, because of their complexity, deep learning models are notorious for being non-transparent; when all its learned weights are known, it is still unclear how the model's output is causally related to the model's input. Moreover, the features in hidden layers often hardly contain any humanly intelligible meaning [36], although meaning extraction or visualisation of individual neurons or layers exist [40, 37].

Some models operate in a low-risk environment where the opacity is no problem (e.g. movie recommendation systems) and a high performance provides enough basis for trust. However, in health care, good model performance is not the only desirable objective. Stakeholders like clinicians are also interested in validating the model with their own expertise [60]. For example, clinicians want to know which parameters the model takes into account such that the model's limitations can be assessed. In addition to that, a confident prediction accompanied by a convincing explanation can add towards trust in the model. In case of an implausible prediction, or a plausible prediction accompanied by an implausible explanation, improving the model via feedback could be possible. Rudin [50] has argued on moral grounds that black-box models are undesirable for any high-stake decision at all. Since depression misdiagnosis or late diagnosis can bring enormous health-related risks and costs [39], depression prediction is definitely a high-stake decision.

1.2 Research Objectives

This thesis intends to contribute towards interpretable yet accurate prediction of depression severity. As discussed earlier, in the context of ML, there is generally a trade-off between predictive performance and model interpretability. It is hypothesised that this trade-off mechanism can be reduced when including the predicted presence or severity of depression symptoms as intermediate features. This should provide clinicians with performance comparable to state-of-the-art methods, but adding the benefit of instance-specific explanations. This hypothesis will be tested by developing and comparing two-tier depression prediction systems, where the first tier involves prediction of depression symptoms from speech-based features, and the second tier uses these predicted symptoms for total depression severity prediction. Basically, this would force the whole system to use a smaller and more intelligible feature space. In this thesis, different two-tier system designs are compared to each other and to a direct prediction approach. This is done on the basis of performance (of symptoms as well as the final depression score) and interpretability.

There are different reasons for the approach of predicting depression via its symptoms. First of all, symptoms are more observable than depression itself. This means that clinicians can compare depression predictions based on predicted symptoms to their own assessments. This idea is related to the approach discussed in Rudin [50],

where a comparison is made to how people explain things to each other. This often happens by dividing a concept into its simple elements, and providing evidence for these elements. In the case of depression, these elements would be symptoms, amongst others.

Moreover, depression comes in a wide variety of symptom combinations, both in terms of theoretical combinations as well as practically [18], which might result in direct depression prediction models being biased towards certain symptoms. In addition, symptom prediction is a valuable objective in itself, since depression symptoms can be symptoms for other (mental) health problems as well.

To the best of my knowledge, this thesis research is the first to propose predicted depression symptoms as predictors of depression severity.

1.3 Thesis Outline

In the remainder of this thesis, I will first discuss some theoretical background and previous research related to the subject in Chapter 2. Then I will shortly describe some shortcomings in the current literature, the proposal to fill these theoretical gaps and state the research questions in Chapter 3. Chapter 4 deals with the methodology of the research, and Chapter 5 discusses the results of the experiments. Using these results, a the research questions are answered, some limitations are discussed and context about the findings are provided in Chapter 6.

Chapter 2

Literature Review

In this chapter, an overview of relevant prior work is given. In Section 2.1, I provide some theoretical background on a range of different topics that touch upon this research: mood disorder diagnostics, paralinguistics for state and trait detection of speakers, explainable ML, interpretable ML and the desire of interpretability in clinical contexts. Section 2.2 provides an overview of relevant research in automated emotion, mood and mood disorder prediction, followed by a discussion of attempts to incorporate interpretability in the context of mood (disorder) prediction and speech features and a description of the AVEC'19 data set.

2.1 Theoretical Background

In order to understand how mood disorders are usually identified by clinicians, and how automated diagnostic could complement this process, this section provides some necessary background on these matters.

2.1.1 Diagnostics for Mood Disorders

In diagnosing mood disorders, interviews by psychiatrists or mental health professionals are regarded as the gold standard [7]. However, this method is time-consuming and expensive. Next to that, the cost raises the bar for many people to reach out to these specialists. In some cases, people seek help in less specialised health care, where incorrect diagnosis happens more frequently. Moreover, diagnosis can involve a degree of subjectivity. Therefore, there is an increasing interest in the search for objective mood disorder markers [5]. Objective markers can reduce diagnosis and monitoring time and costs, as well as increasing patient well-being, for example by providing remote and accessible advice. Gathering mood (disorder) information on a larger population would also be possible using objective markers.

Therefore, personal as well as societal costs can be enormously reduced when accessible, timely, low-cost and less subjective diagnosis is possible. According to Cummins et al. [7], ML can play an important role in these challenges. ML techniques are able to find patterns that are generalisable in data where manual analysis would not succeed. Therefore, with ML there is the potential of finding objective diagnostic markers.

Speech is known to contain cues for a person's mood. Mood disorders are potentially causing cognitive and physiological changes, which have their effect on the motoric actions that form speech [5]. Depressed speech often contains lower energy and lower acoustic variability, although it is possible that this is the case only for a subset of depressed persons [6]. Still, this indicates that paralinguistic speech cues could be used as proxies for mood disorders. From a linguistic perspective, depressed speech has been found to contain more first-person words, and more words related to negative emotions [7].

2.1.2 Paralinguistics

Paralinguistic speech analysis deals with extracting information that does not regard the message a speaker conveys, but states or traits of a speaker [51]. Vocal and linguistic aspects can both be part of this analysis. Since these modalities consist of different features, respectively acoustic and textual, they can be processed differently.

For acoustics, raw audio files are commonly denoised and chunked into units of analysis so that feature extraction can be performed. A wide variety of different features can be found in the literature, ranging from spectral and cepstral (e.g. Mel-Frequency Cepstral Coefficients (MFCC)) to voice quality features (e.g. jitter and shimmer) [51]. From these Low-Level Descriptors (LLDs), functionals such as extremes, percentiles, variation and many more are often computed, summarizing the LLD over a large window of time, e.g. and utterance or a session. The LLDs and their functionals can then be fed to a learning algorithm in order to train a model for state and trait prediction.

Pre-processing of the linguistic modality often involves techniques like word tokenisation, stemming and lemmatisation, where non-linguistic utterances like laughter or sighs can be represented as 'words' too [24, 51]. Words can be represented in n-gram format using a bag-of-words feature representation, and word or sentence embeddings (e.g. Bidirectional Encoder Representations from Transformers (BERT) [27] and S-BERT [44]) can be created, which is often done in the context of deep learning models [16, 49].

2.1.3 Explainable Machine Learning

Machine Learning (ML) is a collection of methods capable of learning a model by feeding an algorithm with sample data. When the learning process is effective, the resulting model is good at predicting unseen data. However, given that ML models are always meant to operate in real-world tasks, good predictive performance is often not the only metric we are interested in. Among several other objectives (like bias avoidance or safety), explainability can be very desirable. If the operational context demands that an ML model not only predicts well, but also provides an answer to *why* a certain prediction or set of predictions is made, this can increase its

applicability in many circumstances. Several reasons for explainability can be the human or scientific curiosity for finding meaning in data, detecting bias, assessing safety, promoting social acceptance, managing social interactions, debugging and auditing [36].

Some ML models intrinsically 'explain themselves', while there are others that do not. For example, decision trees provide explanations for their decisions, because each decision is based on a decision rule. The same goes for linear and logistic regression, to name a few. There is also the category of models that do not inherently provide explanations, often called black-box models. An example of these are (deep) neural networks. The difficulties in explaining such a model class lies in the fact that a full description of the model (in case of a neural network: its nodes, connections, weights and biases) does not give us information on the reason behind a prediction, while for decision trees and linear regression a full description of the model does provide us with these reasons. Deep neural networks contain so many hidden features and nested non-linear functions, that it is not even remotely possible to render the mapping from input data to prediction [36].

Nonetheless, some methods exists that are able to extract explanations, although these are actually proxy-explanations, because 'the' ground-truth explanation does not exist. This is illustrated by the fact that, over the years, many model-agnostic explanation methods have been proposed. Global explainer methods Partial Dependence Plots (PDP) [21] and Permutation Feature Importance [17]) are designed to provide information on the global working of the model, while local methods like Local Interpretable Model-Agnostic Explanations (LIME) [45] or SHAP [33] are able to provide explanations on instance-level in the form of individual feature contributions to the model output. SHAP is also able to provide the globally most important features and pairwise feature interactions over instance-wise explanations. Explainer methods specifically designed to provide explanations for convolutional networks also exist, like Gradient-weighted Class Activation Mapping (Grad-CAM) [52] and SmoothGrad [54]. The explanation of an image then consists of a saliency map that can be understood as the pixels that the CNN paid most attention to in its prediction.

Whether the above-mentioned proxy-explanations are actually good or not, depends on a few factors. Robnik-Šikonja and Bohanec [48] describe such explanation quality factors, one of which is the fidelity of an explanation, i.e. whether the explanation follows the inner logic of the model correctly. The fidelity of an explanation is often not straightforward to quantify, because the ground-truth explanation does not exist. Moreover, more explanation quality factors exist, like accuracy (how well the explanation generalises to unseen data), stability (how similar explanations are for similar data instances) and many more.

2.1.4 Interpretable Machine Learning

In this thesis, I will distinguish between the concepts of explainability and interpretability, and put the focus on the latter. As aforementioned, an explanation is conceived as a piece of information on the reason behind a prediction in terms of input features. Interpretability, however, is the degree to which explanations make the model intelligible to a human. The distinction between these concepts might seem small, but the fact that a model can be explained does not necessarily entail that this explanation is interpretable by a human being. Interpretability depends on many factors: the limited cognitive capacities of humans or a person's understanding of the model and its features. A decision tree is very well capable of providing the reason behind a prediction, since this reason is just the decision rule, but if the rule contains a large amount of features, the rule is not very interpretable. If a human being is not familiar with the basic mathematics behind an inherently explainable model like linear regression, the weights do not make sense and the model can not be interpreted by this person. The type of input features is important in order to assess interpretability. In the example of the saliency map, the features (pixels) are very intelligible, because they form a humanly intelligible complex (the image). In case of large amounts of features, or features do not contain intelligible meaning, the interpreting a model that uses these features becomes more difficult.

If interpretability is so subject-dependent, how can we make an assessment of the interpretability of a model? Doshi-Velez and Kim [13] conceptualise three methods for assessing interpretability. The best, but most costly testing method would be an application-grounded evaluation, in which humans that would actually use the system, test it in a real-life application setting. For example, psychiatrists might be asked to use a depression prediction system on some actual patients. Second-to-best testing is called human-grounded evaluation, where humans are asked to do tasks that simplify the target task, for example by rating different explanations of a model. The most simple type of interpretability assessment is functionally-grounded evaluation. This method is the least costly in terms of time and money, since it does not involve any humans or a real-life environment. Functionally-grounded evaluation only makes use of proxies of interpretability: the number of cognitive chunks (how many basic units does the explanation consists of?), the form of the cognitive chunks (do these cognitive chunks carry any meaning?), interactions between the cognitive chunks, the level of compositionality of the chunks (are certain units defined in terms of other units?) and the degree to which a human understands probabilities, are just some proposed proxies for the degree of interpretability.

In general, there is a trade-off between model complexity and inherent explainability. Complex models are better at recognising more intricate patterns in data, but these patterns are harder to make explicit. However, more recent research has tried to combine complexity with explainability. An example of this is the Explainable Boosting Machine (EBM) [38], which is inherently able to provide explanations. However, EBM capabilities are very limited on tasks where neural networks are

state-of-the-art, such as image or sequential data. Moreover, the EBM can only guarantee partial interpretability, through global feature importance in the form of PDP.

This thesis deals with depression prediction from speech features. In principle, speech has a sequential nature. As discussed in 2.1.2, the acoustic part of speech can be represented using spectral features, which can be excellently processed by CNNs or RNNs and explained using the model-agnostic or model-specific explainer methods mentioned above. However, a common challenge is to make these explanations interpretable, since spectral input features might not contain any humanly intelligible meaning at all. The same goes for state-of-the-art word embeddings like BERT [27]. This problem is further discussed in section 2.2.3.

2.1.5 Interpretability in a Clinical Context

Tonekaboni et al. [60] have studied clinicians' desires regarding explanations when interacting with ML models. Although the interviewed clinicians work at intensive care units (ICU) and emergency departments (ED), insights about when an explanation is a good one are expected to be transferable to psychiatric and mental health settings, since in both cases there is a diagnosis and treatment. The following points were found to be important for building trust towards an ML system:

Feature importance Knowing the subset of features most important for the model outcome is crucial, since discrepancies between clinical judgement and model decisions can be quickly detected. This can include global as well as local (patient-specific) feature importance.

Explanation by similarity When a new patient is seen, knowing what similar patients the model has seen (and how these were handled) during training can be helpful.

Certainty Providing the certainty of a model decision gives clinicians something close to an explanation. For example, whether or not acting upon a model's outcome is sensible might depend on the accompanying confidence score.

Temporal explanations There is interest in knowing the temporal changes in a patient's state that are responsible for a certain prediction.

Design transparency Model transparency is found important by clinicians, because of easy comparison to existing diagnostic methods.

It should be noted that in some cases not all of the listed properties are feasible. For example, for many prediction tasks model transparency is at odds with model accuracy.

2.2 Related Work in Predicting Emotion and Mood Disorders

Research in mood and mood disorder recognition has been approached in different ways, with different modalities, features and goals. Although this thesis does not deal with mood or emotion prediction, since depression is a mood disorder, previous research in this field might be relevant for the current goal. This section deals with previous research that has dealt with decision support in the medical domain, more specifically focused on themes relevant to this thesis: mood, mood disorders, and interpretable prediction. This section is divided thematically, such that first emotion and mood recognition is discussed, then mood disorder recognition. Consequently, interpretability methods that are relevant to these goals and finally the specific dataset that this thesis will use, are discussed.

2.2.1 Emotion and Mood

In this section, a short description of previous research on mood and emotion recognition is given, where the focus lies on research that is relevant for prediction from speech. Mood can be generally considered to be persisting over longer periods of time (minutes to hours), while emotions typically last shorter (seconds to minutes) [23]. Also, emotions might be provoked by distinct events and recognised by identifiable expression, while the causes of a certain mood are more distributed over time and thus less distinct.

Since humans greatly rely on facial expression to detect emotions in others, one of the common approaches in automated emotion detection is using visual (image or video) features. Including audio features next to visual features has shown to have some predictive benefit for specific emotions [25, 53]. Regarding mood, audio can be good predictors of the arousal dimension of mood, while video and linguistic features generally predict the valence dimension better [55]. Here, simple linguistic features such as valence, arousal and dominance scores per word and text entity per second are used. For elderly emotion recognition from speech, it has been found that such simple linguistic features also perform better than state-of-the-art word embeddings like BERT [56, 27].

2.2.2 Mood Disorders

Next to physiological, behavioural, visual, graphological and neurological approaches to psychiatric disorders, using speech cues is also a promising field of study [32]. Depression (severity) prediction literature indicates that cues like shimmer, jitter and F0 variability tend to increase with depression severity, while F0 mean and range are found to be significantly lower for depressed patients.

Automated mood disorder recognition has shown to benefit from multiple feature modalities, often done with acoustic, linguistic and visual (facial) modalities.

The winners of the AVEC'19 Detecting Depression with AI Subchallenge (DDS), Ray et al. [43], have used attention networks to combine video, audio and linguistic modalities from videos of participants. By analysing the attention weights per modality, it was found that the linguistic modality is weighted more than the video and audio modality features combined. A unimodal approach using only the linguistic approach performed somewhat less but remarkably comparable to the multimodal approach. In general, approaches on the AVEC'19 depression data set have focused on various modalities: linguistic and speech behaviour features [26], acoustic-only features [63], acoustic and textual features [49, 16], visual and acoustic features [59] and acoustic, textual and visual features [62, 43]. To encode temporal information, some studies employed a Convolutional Neural Network (CNN) [49, 16] or transformer networks [59], while others used recurrent neural networks such as a Gated Recurrent Unit (GRU) [26] or a Bidirectional Long Short-Term Memory (BLSTM) [62, 63, 43], in order to encode information in both temporal directions.

This illustrates that state-of-the-art methods on the AVEC'19 data have been focusing mostly on methods that are quite complex in order to performs well. Moreover, input features often are spectrogram, MFCC, BERT-encodings and other features that are hard if not impossible to interpret. Thus, even when model-agnostic methods would be employed for relating the output to the inputs, the model is still not interpretable. The next section discusses some approaches to overcome these problems in the context of mood (disorder) prediction and acoustic or linguistic features.

2.2.3 Interpretability in Speech and Affect Recognition

Interpretability partly depends on the intelligibility of input features. Since the topic of this thesis is prediction from speech, previous research on interpretability in this context is relevant. Efforts to add interpretability to effective but complex models for emotion or mood (disorder) recognition have increased over the past years.

A thorough feature selection study for depression prediction has indicated that there is a high amount of redundancy in speech and visual features [2]. Using only 9 carefully selected features from 815 speech prosody, eye behaviour and head behaviour features, an SVM was able to perform similar or better on different depression data sets when compared to using all features. These 9 features consisted of F0 (average, minimum and first derivative), Harmonics-to-Noise Ratio (HNR) (minimum and range), shimmer (maximum second derivative) and second formant (F2) (minimum) features, next to 2 eye behaviour features. The 9 features have been selected by taking the features that scored well on at least 2 of the 3 data sets according to 38 feature selection methods. Since the datasets have a different setup, it is expected that these features generalise better.

Using explainable methods, the same redundancy is found using a CNN-based dynamic attention model for depression prediction from video, audio and text features [4]. The (globally) most important features for the model were extracted using

SHAP [33]. Using only 5% of the best features based on the SHAP ranking, the F1 score dropped with only 1 percentage point in comparison to using all features. Also, the SHAP explanation showed that text features appear relatively often as important features for the model. In addition, text-based unimodal model performed much better than unimodal models of audio or video features. Important textual features were certain affective features (like valence, dominance, subjectivity and arousal) and word polarity.

These studies on global depression interpretability indicate that relatively intelligible features can provide quite some predictive power. They also indicate that acoustic and linguistic features are potentially important modalities.

A recent attempt to interpretable multimodal speech recognition is done by explaining the inner working of a model that predicts emotions from audio and text [30]. In this research, the authors propose a method to determine how (quickly) the deep learning model learns to separate the different emotion classes in the final dense layers. Another work has focused on interpretable prediction of affects using physiological signals [31].

In a study that predicts COVID-19 coughs from spectrograms using a CNN [3], the instances are explained using Score-CAM [61] saliency maps. However, spectrograms might not provide good performance on speech data, since spectral features are not designed to represent variance in vocal behaviour well. Also, unlike facial images, interpreting saliency maps of spectrograms is quite difficult, since spectrograms themselves are arguably hard to interpret, depending on the prediction task and expertise of the interpreter.

Another study has found that predicting personality impressions from a small set of mood and likability features (high and low valence, arousal and likability, thus only 6 features in total) is done quite effectively using an EBM [55, 38]. In addition, inspecting the model shows which mood features correlate with personality features.

2.2.4 The AVEC'19 Challenge Data

In this thesis, the dataset provided at the AVEC'19 Detecting Depression with AI Sub-Challenge (DDS) [47] will be used for experimentation. This is known as Extended Distress Analysis Interview Corpus (E-DAIC) [9], an extension from the Distress Analysis Interview Corpus Wizard of Oz (DAIC-WOZ) dataset [20]. This dataset consists of semi-clinical interviews, some of which are conducted by a virtual agent, which is controlled by a human in another room, while others are conducted by a fully autonomous AI. From these interviews, raw audio as well as pre-extracted audio features like eGeMAPS and MFCC are available. Also, pre-extracted visual features are provided, as well as text transcribed by the Google Automatic Speech

Recognition (ASR) tool[1]. Eight Item Patient Health Questionnaire (PHQ-8) scores of every interviewee are known.

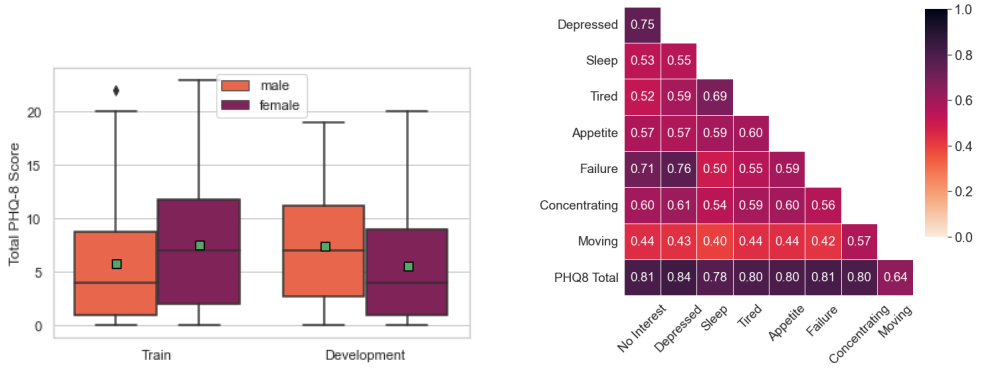
The data set consists of 275 interviews, of which 163 instances are meant for training, 56 instances are allocated for development and 56 for testing. The skewness in the total PHQ8 score labels and two PHQ items can be seen in Figure 2.2. In total, there are 8 PHQ items, of which all are skewed towards lower scores. The 8 items are: no interest, depressed, sleep, tired, appetite, failure, concentrating and moving. The exact questions pertaining to these items can be seen in Table 2.1.

Over the last 2 weeks, how often have you been bothered by any of the following problems?	*Not at all*	*Several days*	*More than half the days*	*Nearly every day*
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself - or that you are a failure or have let yourself or your family down	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8. Moving or speaking so slowly that other people have noticed? Or the opposite - being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3

TABLE 2.1: The Patient Health Questionnaire with 8 questions (PHQ-8) [29]. Symptom scores are added up to obtain the PHQ-8 score.

The distribution of total depression scores per gender group are shown in Fig. 2.1 (a). As can be seen, depression severity mean and maximum scores for the female group are quite higher than the male group for training set. This observed difference between gender groups found in the dataset is inline with the psychological literature [11]. However, the opposite gender correlation with depression severity can be observed on the development set. The development set is quite small and consists of 56 sessions, which apparently does not represent the real-world. Fig. 2.1 (b) illustrates the Pearson correlation matrix of symptoms and final depression severity score. As expected it can be observed that all correlations between depression score and symptoms are positive. The second observation is that inter-symptom correlations are mostly either moderate (higher than 0.4) or strong (higher than 0.6). While

[1]https://cloud.google.com/speech-to-text



(a) Distribution of total PHQ-8 depression score per male and female groups on the train and development set. Green square indicates mean.

(b) Pearson correlation between depression symptoms and total depression score in the combined train + development set.

FIGURE 2.1: The relation between depression and (a) gender and (b) symptoms

No interest, depressed and failure have remarkably strong correlations (higher than 0.7) among them, *moving* symptom has relatively weaker correlations with other symptoms and the PHQ-8 score.

Despite the fact that this dataset plays a crucial role in the paralinguistic research field, there are some limitations to it. First, the loudness is in general low and the audio quality of the interviewees varies largely. This causes problems with the automatic transcriptions, which pose a real-life challenge. Second, unlike the DAIC-WOZ version used in the AVEC 2017 challenge [46], here we do not have the time-stamped and action tagged manual transcriptions of the virtual agent and the participants. Having these detailed annotations could have boosted the symptom predictions via selecting relevant responses from the participants, as done by Gong et al. [19] and Sun et al. [58] to predict the PHQ-8 score. Third, even though I use the extended version, the size of the dataset is small, which is the case for most of the clinical sets due to high confidentiality and difficulty of collecting this type of personal data. Combined with the class imbalance, this causes low prediction performances by the machine learning models trained on these sets. Moreover, when the experimental dataset is not a good representation of the real-world, then models may carry undesired biases and/or unrealistic associations between variables. Interpretable modeling is also a necessity to overcome these limitations [50].

Table 2.2 shows some participants of the AVEC'19 DDS challenge, as well as the best baseline system created by the challenge providers. Challenge guidelines state the Concordance Correlation Coefficient (CCC) as the scoring metric, where a score of 1 means a perfect (positive) correlation between predicted and actual values, and 0 means no correlation at all.

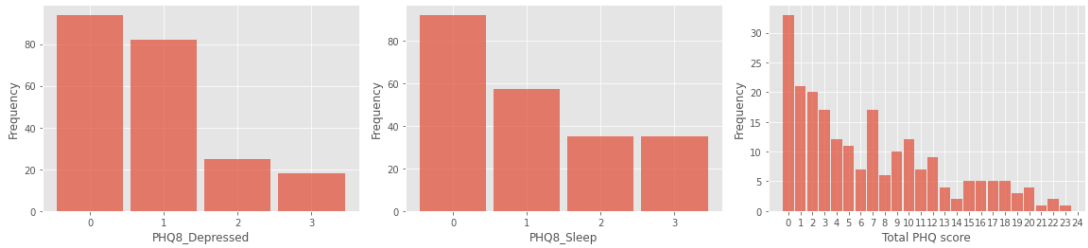


FIGURE 2.2: The distribution of PHQ scores for the *depressed* syptom, *sleep* symptom and total depression severity score in the training + development data, taken from [9].

Paper	Modalities used	Methods used	Test set CCC
Multi-level Attention Network using Text, Audio and Video for Depression Prediction [43]	Text	Multi-layer attention network	0.67
A Multi-Modal Hierarchical Recurrent Neural Network for Depression Detection [62]	Video, Audio, Text	Hierarchical BiLSTM	0.44
Multi-modality Depression Detection via Multi-scale Temporal Dilated CNNs [16]	Audio, Text	Ensemble of Multi-scale Temporal Dilated CNNs	0.43
Predicting Depression and Emotions in the Cross-roads of Cultures, Para-linguistics, and Non-linguistics [26]	Text	Equal Weighted Fusion of Kernel Extreme Learning Machines	0.34
AVEC'19 Official Baseline [47]	Video	Single layer GRU	0.12

TABLE 2.2: Overview of relevant papers that participated in the AVEC'19 DDS challenge.

2.3 Background on Methods Used

In this section, I discuss some background on speech-based features and model class that are used in this thesis.

2.3.1 Speech-based Features

In this section, I discuss some speech-based features that can be extracted from audio or transcriptions. Features from transcriptions are also called speech-based, since they originate from speech, as opposed to written text. The AVEC'19 challenge data provides a large set of acoustic features, amongst others MFCC, eGeMAPS and deep audio features, extracted from pre-trained audio CNNs [47]. Since this thesis will mainly make use of textual features, I will focus my description on these. However, since some eGeMAPS features are quite intelligible, and some experiments relate to this, I will discuss these as well.

Bidirectional Encoder Representations from Transformers (BERT) [10] is a State of the Art (SoA) contextualized word embedding method, which is commonly used in different Natural Language Processing (NLP) problems such as sentiment analysis, text classification, and semantic textual similarity. The pre-trained S-BERT [44], more specifically the 'all-MiniLM-L6-v2' [2] network, which is a modification of the

[2]https://www.sbert.net/docs/pretrained_models.html

BERT network to sentence-level and was pre-trained on a Natural Language Inference (NLI) dataset, is used in the experiments. Each utterance was processed by the Sentence-BERT encoder to construct 384-dimensional vectors, then we apply functionals (mean, variance, median and 75th percentile) to summarize the whole dialogue. The 'all-MiniLM-L6-v2' is chosen since it turned out to be performing best in the baseline setting in comparison to word-level BERT embeddings and larger embedding S-BERT embeddings. Also, the four chosen functionals were chosen from a larger set of summarising functionals, based on a preliminary performance analysis on the baseline model. It should be noted that the S-BERT features are not intelligible at all.

Linguistic Inquiry and Word Count (LIWC) is a tool commonly used to extract textual features that were shown to be significantly correlated with affective dimensions [34] such as personality traits. The LIWC tool was developed by Pennebaker et al. [42] and it allows doing text analysis by means of rich dictionaries and predefined categories. I used the LIWC 2015 tool to extract information from the given dialogue for 93 LIWC categories (e.g., affect information, language measures, informal speech). Since transcripts were automatically transcribed by Google ASR tool, some of the LIWC categories are not relevant for this dataset as they are not present at all in any utterance. Therefore, counts for semicolon, question mark, exclamation mark, quote and parenthesis use were removed from the feature set. The LIWC features are intelligible, since they indicate how often a certain word category a person is using. For a detailed discussion of word categories, see Pennebaker et al. [42].

Hand-crafted feature set (HS) consists of sentiment, speech rate, repetition rate and transcription confidence score, which were extracted per Automated Speech Recognition (ASR) transcription turn. As sentiment analysis is a hot research topic in Natural Language Processing, various pre-trained models and tools have been made available for research purposes. In this study, I use Flair [1] sentiment analysis library to extract sentiment features from the transcripts. Speech rate is computed by dividing the number of words to speech duration in seconds. To compute the repetition rate, the number of words per turn are divided by the number of unique words. The transcription confidence score, as computed by Google ASR tool, is provided in the dataset. As some of these features are extracted from each ASR transcript rather than the whole session, the features are summarized over the interview session by applying mean, standard deviation, sum, minimum and maximum functionals over the turn-level scores. This results in a set of 20 features that are intelligible to humans.

eGeMAPS acoustic LLD features are used in a small set of experiments. The AVEC'19 dataset provides eGeMAPS [15] LLD features, 25 in total. These features are for a large part unintelligible, but F0, jitter, shimmer and loudness are quite intelligible acoustic features. Summarising functionals are applied over these features, to obtain single values for each interview. These summarizers are mean, standard deviation, variance, median, range ($maximum - minimum$), and robust range

(80*th percentile* − 20*th percentile*). The experiments in which these features are used are discussed in Chapter 4.

2.3.2 Kernel Extreme Learning Machine

Kernel Extreme Learning Machine (KELM) is a fast and accurate learning method, whose efficacy is shown in a range of paralinguistic and affective computing challenges [22, 25, 26]. Since a Support Vector Machine (SVM) is comparatively slower, and neural network are more difficult to train on relatively smaller datasets like provided by the AVEC'19 challenge, the KELM seems to be a promising model.

Given a training dataset $\mathbf{D} \in \mathbb{R}^{N \times d}$ with N instances and d features, KELM solves a regularized least squares regression problem between the kernel (similarity) matrix $\mathbf{K} \in \mathbb{R}^{N \times N}$ and the target variable / matrix $\mathbf{T}$:

$$\beta = \left(\frac{\mathbf{I}}{C} + \mathbf{K} \right)^{-1} \mathbf{T}, \tag{2.1}$$

where C is the regularization coefficient optimized via cross-validation and $\mathbf{I}$ is an $N \times N$ identity matrix. The prediction for a test set instance x is obtained via $\hat{y} = K(\mathbf{D}, x)\beta$, where $K(,)$ denotes the kernel function.

In the case of regression, $\mathbf{T}$ can be the target variable vector and can be extended for multi-task regression, where each column can represent a separate regression task. In the case of classification, the categories can be one-hot encoded. In case of ordinal classification, it is possible to use ordinal encoding as in [14, 28]. The flexibility of coding the target matrix $\mathbf{T}$ allows to benefit from the internal correlations of the target variables via multi-task learning and also by means of ordinal encoding of the originally ordinal scores of PHQ-8 symptoms. Thus, I compare single vs. multi-task regression, as well as regression vs. classification alternatives for symptom modeling.

Class imbalance, which is typical in mental healthcare corpora, may mislead the model in favor of the majority class. Using weighted models is one solution to the imbalanced learning problem. In weighted ELM [64], we define a $N \times N$ diagonal weight matrix $\mathbf{W}$, where N is the number of samples. Each diagonal element stores the multiplicative inverse of the number of training samples N_i of the corresponding class i. Integrating $\mathbf{W}$ into the formula, β is calculated as:

$$\beta = \left(\frac{\mathbf{I}}{C} + \mathbf{WK} \right)^{-1} \mathbf{WT}. \tag{2.2}$$

2.3.3 SHAP

SHAP is designed by Lundberg and Lee [33] to provide model-agnostic local and global explanations. In this thesis, I make use of the the fact that SHAP is considered to have a solid foundation in game theory, which enables a user to see which features have contributed to a prediction. SHAP is based on Shapley values. Shapley values

give insight into how a prediction, the 'payout' in game theoretical terms, are fairly distributed among the features (the players) [36]. In this way, for every data instance and its prediction, a feature importance can be computed. By summing these feature importances over multiple data instances, the global feature importance can be obtained. In case of classification, SHAP computes feature importances over all class probabilities, so in that case the feature importances should also be summed over classes.

Chapter 3

Research Approach

In this chapter, the literature gap is described in Section 3.1, the research question(s) that are formulated to provide an answer to this gap are presented in Section 3.2.

3.1 Literature Gap

Current research on automated prediction of depression from speech (and video) has focused mostly on good performance. Next to this, some attempts to add interpretability have been done. Still, these attempts have not yet succeeded to include both 1) actual (instead of proxy) explanations of predictions and 2) expression of input features in a way that is intelligible by clinicians. Therefore, this thesis is aimed at incorporating both criteria into a depression prediction model. In other words, the thesis aims at creating a depression level predictor model that performs well, but is overall interpretable for humans. As a variation on this, as second proposed approach will include intelligible speech-based features next to PHQ8 symptoms as intermediate features, as discussed in Section 4.

3.2 Research Questions

In an attempt to fill the described gap in the literature, the following research question (RQ) is posed:

Research Question: *How far can the predictive performance of a state-of-the-art depression prediction system be approximated (on the AVEC'19 test set in terms of CCC), while keeping the model interpretable by using a limited set of depression symptoms, predicted from acoustic-linguistic features, and interpretable text or acoustic features as intermediate observable features?*

In order to answer this research question, the following research subquestions (RSQ) and subquestions to the RSQs are answered:

RSQ 1: *As a baseline model, can a direct depression prediction system from speech-based features be created that performs on state-of-the-art level on the AVEC'19 test set, in terms*

of CCC score?

RSQ 2: *Which system, with PHQ items (depression symptoms) predicted from speech-based features as intermediate features, maximizes the AVEC'19 CCC score, while improving over the interpretability of the baseline model?*

RSQ 2.1: *How does single-task symptom prediction compare to multi-task symptom prediction, in terms of final depression CCC score and in terms of predicted symptom performance?*

RSQ 2.2: *How does single-task symptom prediction compare to multi-task symptom prediction?*

RSQ 2.3: *How does mere summation of predicted symptoms compare to a second stage regressor, in order to predict the total PHQ-8 score?*

RSQ 2.4: *Which features are important for each of the symptoms?*

RSQ 3: *Can interpretable speech-based features be added to predicted symptoms, in order to boost performance of the final depression score prediction?*

Chapter 4

Methodology

The original AVEC'19 DDS challenge prescribes that only five attempts on the test set could be done, so for comparability reasons, this thesis intends to keep the amount of test set attempts limited. Optimisation is done using the development set performance, and for models that are promising in terms of performance and/or interpretability, the test set is probed.

In this chapter, I first introduce the proposed framework with the proposed prediction pipelines. Then, I describe the features that are used in the experiments, provide a discussion of the model that I use, and and give further details about the experimental setup. Finally, the evaluation of interpretability will be discussed.

4.1 Prediction Pipelines

Although a baseline model was provided for the challenge, I define another baseline model for the experiments, corresponding to RSQ 1. This baseline is used to compare to the proposed (more) interpretable systems, where the model predicts the total PHQ8 depression score directly from speech features. The approach is outlined in Figure 4.1, where a pipeline of the system is shown. For this baseline, different options for different blocks in the pipeline are tried.

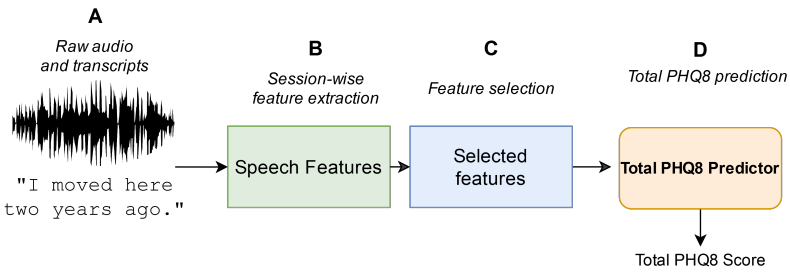


FIGURE 4.1: Pipeline of the baseline system, where selected speech-based features immediately predict the total PHQ8 depression score.

The first proposed interpretable system uses depression symptoms as intermediate features, as shown in Figure 4.2. From selected speech-based features, the depression symptoms are predicted, which are used to predict the total depression score.

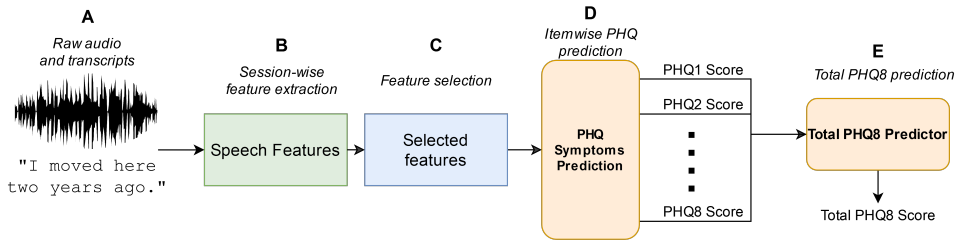


FIGURE 4.2: Pipeline of the proposed prediction system, where first symptoms are predicted, which are used to predict the total PHQ8 score.

The third prediction pipeline, seen in Figure 4.3, is intended to make use of intelligible speech features to complement the predicted symptoms. In this way, it might be possible to add predictive power to the final depression score predictor, without losing interpretability.

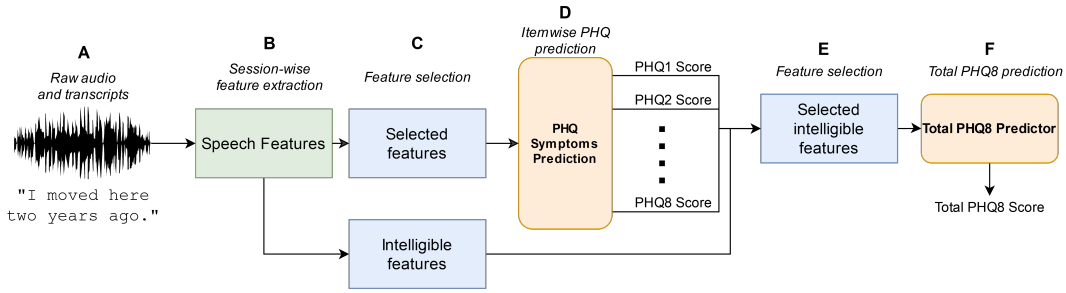


FIGURE 4.3: An extended version of the proposed pipeline, where interpretable speech features are used to complement predicted symptoms for the second tier predictor.

4.2 Setup of the Experiments

An extensive set of experiments are conducted, trying different combinations of aforementioned sets of features, alternative cascaded normalization schemes as well as kernel options and regularisation levels for each model. The options for cascaded normalization for all experiments include combinations of z-normalization with power (signed square root of each feature value) and L2-normalization. For kernelization, I experimented with linear, polynomial, sigmoid and Radial Basis Function (RBF) kernels. I experimented the KELM complexity parameter in the set {0.0001, 0.0003, 0.001, 0.003, ..., 100, 300}.

For the baseline model, a number of system options have been tried. First of all, a feature set selection has been performed. From the S-BERT, LIWC and handcrafted feature sets, combinations of feature sets have been tried and optimised on the development set. This feature selection is done for a KELM model as well as for an SVM. As described in Section 2.3.2, the KELM is a very fast learner and able to be used in a multi-task setting by design, while an SVM is not. Since depression symptoms are highly correlated (see Figure 2.1), the KELM has an advantage. Whether or not these

advantages are compromising the predictive performance of the KELM, I compared its performance to that of an SVM, using the same hyperparameter and feature set settings. As will be discussed in Section 5, the KELM did perform somewhat worse than the SVM, but not so much as to exclude it from subsequent experiments. Thus, for comparibility reasons, the following experiments into interpretable models have been conducted using KELM.

For the indirect prediction model using only predicted symptoms, the same feature sets, kernel options, and regularisation parameters as described above have been used. For the symptom prediction (block D in Figure 4.2), several options are available. First of all, symptom prediction can be done in a classification or regression setting. For regression, the multi-task option is available.

For the second model, the total PHQ8 predictor, some different options are tried as well. A first and simple approach is to sum the predicted symptoms, as in the PHQ8 questionnaire, to obtain the predicted depression severity score. Secondly, it is possible to train a second tier regressor with predicted symptoms as features. This should be done in a setting where symptom predictions are stacked, where the training protocol should be modified for the symptom prediction models, in order to provide unbiased predictions to the second tier regressor. K-Fold Cross-Validation (CV) on the challenge training set is used to obtain symptom predictions of this set. Subsequently, all K trained models are used to predict the challenge development and test sets, where their (rounded) average is taken as features for the second tier.

When experimenting with the third pipeline using interpretable features next to symptoms for the second tier (as seen in Figure 4.3), the only option is to use the stacking CV setting. In order to see whether the prediction can be boosted by intelligible speech-based features next to symptoms, we experiment with different combinations of LIWC, handcrafted text features, and intelligible eGeMAPS features (functionals of F0, jitter, shimmer and loudness, as discussed in Section 2.3.1).

4.3 Evaluation of Systems

For the evaluation of systems, two measures are of importance: performance on the prediction of the depression severity score and the degree of interpretability. Predictive performance of the whole system is evaluated in terms of CCC score, as per AVEC'19 guidelines, however, Root Mean Square Error (RMSE) is also provided. The interpretability of systems is assessed based on the following factors:

- **Model transparency** White-box models are highly preferred over less transparent models, because they can inherently provide explanations.
- **Amount of input features** Since the cognitive abilities of human beings are limited, less input features make a model more interpretable. A rough estimation is that seven (plus or minus two) categories can be held in short-term memory [35]. This rule of thumb will set an aim for the amount of features that

are acceptable in the second tier predictor, limiting this amount to preferably less than 10 features.

- **Intelligibility of input features** Especially for the second tier predictor, the input features should be intelligible. We can choose from predicted symptoms, LIWC, handcrafted features and some eGeMAPS features.

- **Accuracy of predicted symptoms** The prediction system is less interpretable if symptoms are not accurately predicted. For interpretability reasons, inaccurate prediction of symptoms but accurate prediction of total depression severity should be avoided. Accuracy of symptom predictions is evaluated in terms of Unweighted Average Recall (UAR) for a classification setting, and in terms of CCC for a regression setting. The choice for UAR is based on the fact that the data is quite skewed towards lower depression (symptom) values. In order to avoid inaccurate prediction of higher symptom scores, every symptom class is weighed equally in the UAR.

It should be noted that the model transparency and intelligibility of input features have the character of harder constraints, while the constraints for the amount of input features and accuracy of predicted symptoms are somewhat softer, because they act on a scale. The softness of these constraints is one drawback of such a functionally-grounded assessment, and future human-grounded or application-grounded evaluation should circumvent this ambiguity.

Chapter 5

Results

In this chapter, the results of experiments done are discussed. The chapter is divided in sections on the baseline model (Section 5.1), interpretable models using only symptoms (Section 5.2) and interpretable models using symptoms as well as other intelligible features (Section 5.3).

5.1 Direct prediction models

For the baseline model, a number of different feature sets have been tried for both the KELM and SVM models, as can be seen in Table 5.1. These models have been optimised on the development set, optimising over the range of pre-processing steps, kernels and regularisation strengths. Using all S-BERT, LIWC and handcrafted features (HS), the KELM performs somewhat worse than the SVM regressor. Different options with only intelligible features (LIWC and handcrafted, or only LIWC or handcrafted) or combinations of intelligible features with S-BERT have been tried, of which the KELM using only LIWC came out best. Still, this model performed on a par with the best KELM model using all features for CCC, but worse on RMSE. These findings indicate that all feature sets likely contribute towards accurate prediction. I continue the two-stage experiments, described in the following sections, with the KELM model, since it is capable of modelling in a multi-task setting (where is SVM is not). Moreover, the KELM model is trained so much quicker that it is expected that more experiments with a more extensive hyperparameter optimisation can be done for the KELM.

	KELM		**SVM**	
Features	**CCC**	**RMSE**	**CCC**	**RMSE**
S-BERT	0.55	4.40	0.56	4.39
LIWC	0.57	5.10	0.40	5.43
HS	0.42	6.45	0.27	5.45
LIWC + HS	0.55	5.99	0.45	4.70
S-BERT + LIWC + HS	0.57	4.41	0.59	4.31

TABLE 5.1: Performance of baseline models on the development set.

		Symptoms									Total PHQ
		1: No Interest	2: Depressed	3: Sleep	4: Tired	5: Appetite	6: Failure	7: Concentrating	8: Moving	Average	
Multi-task Regression	CCC	0.46	0.44	0.42	0.43	0.50	0.37	0.43	0.28	0.41	0.61
Single-task Regression	CCC	0.56	0.44	0.40	0.35	0.37	0.40	0.46	0.18	0.39	0.61
	Feature set	LIWC+HS	S-BERT+HS	S-BERT	S-BERT+ LIWC+HS	S-BERT+ LIWC+HS	S-BERT+HS	S-BERT+ LIWC+HS	HS		

TABLE 5.2: Best symptom CCC scores on the development set for multi-task and single-task symptom regression, with corresponding feature sets. Multi-task regression approach uses S-BERT features.

5.2 Prediction via Symptoms

In the initial experiments with modelling options, I noticed that ordinal coding for classification gives exactly the same outputs with regression in both single and multi-task settings. Thus, I excluded the ordinal classification from the remainder of the experiments and used the regression alternative, which is simpler to implement and interpret.

I consequently experiment with single- and multi-task regression alternatives. These options are tried in the setting where the symptoms are summed to obtain the total PHQ8 score. The comparative results of this experiment is shown in Table 5.2. The best performance for the multi-task symptom regression was obtained using only the S-BERT representation, while in the single task setting the best performing feature combinations varied from a single use of handcrafted set (Symptom 8: Moving), to combination of S-BERT, LIWC and the handcrafted set (Symptoms 4, 5 and 7). As can be seen, the *moving* (moving or speaking very slowly or unusually often) symptom is poorly recognized, maybe due to the fact that this symptom is most skewed towards lower scores. On the overall, it can be seen that despite the different symptom-wise predictions, both alternatives give the same PHQ-8 CCC score up to two decimal digits (0.61) on the development set, when a simple symptom summation model is used. I subsequently probe the test set performance of these two alternatives. On the test set, the multi-task regression shows very close CCC performance (0.62) to the development set, while the single-task regression performance drops to 0.45 CCC (see Table 5.6 for comparative results).

In the next batch of experiments, I compare multi-task regression with single-task Weighted KELM symptom classification in summation and stacking alternatives. In order to stack the symptom predictions to second tier regressors, I switch to K-Fold CV approach as explained in Section 4.2, using K=4. K-Fold CV symptom prediction Unweighted Average Recall (UAR) performances are shown in Fig. 5.1. As can be seen, WKELM models trained for single-task classification gives higher UAR score for all symptom predictions with a large gap. This might be due to the weighting trick used to overcome the class imbalance. Moreover, the sanitization that is applied to normalize the predictions might affect the performance of regression models. Although classification models perform quite higher than regression models, I evaluate both symptoms models for comparing the effect of both models'

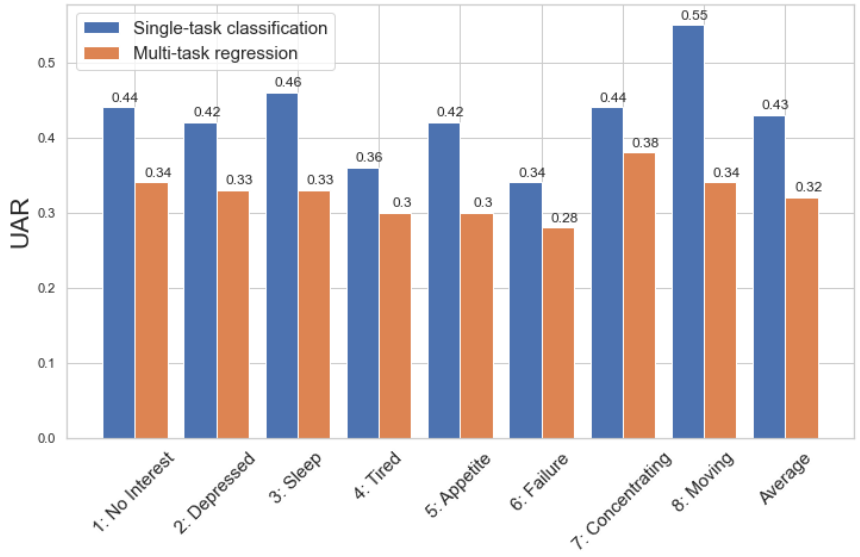


FIGURE 5.1: UAR scores for every symptom as predicted by the single-task classifier and the multi-task regressor.

predictions to the final depression severity prediction.

For the second tier RF experiments, I experimented with stacked regression and classification based symptom predictions both separately and in combined fashion. I experimented with the number of trees, this was varied in steps of 10 within range [10, 300]. The number of features was left at the default value (using all features) in scikit-learn library in Python [41]. Separate stacking experiments have shown that weighted KELM classification based symptom predictions can reach up to 0.62 CCC on the challenge development set, while regression KELM based predictions cannot exceed 0.35 CCC. Further, when I applied a feature selection algorithm (Recursive Feature Elimination using RF as learner) to the combination of the classification and regression based symptom predictions, I observed that the top 8 selected features are exactly the symptom predictions from the classification based models. Therefore, I have not probed the test set performance for the regression predictions based stacking approach. For the sake of direct comparison with the regression based method however, I also probed the summation of classification based symptom predictions.

In order to understand which features have been important for the separate symptoms, a SHAP analysis has been made. Every symptom is classified separately in the single-task classification setting, since for single-task learning, classification can benefit from the weighted KELM option. All S-BERT, LIWC and handcrafted features have been used to train a weighted KELM for each symptom, where kernel choice, regularisation parameter and preprocessing were optimised on the development set in terms of UAR. Then, these 8 optimised models were analysed using SHAP, in order to obtain an ordered list of most important features (globally). Surprisingly, for the *moving* symptom, the SHAP analysis resulted in a zero feature importance for all features. It was hypothesized no feature adds information to the class prior probabilities, which would mean that the *moving* symptom classifier is

Top 10 Features	1: No Interest	2: Depressed	3: Sleep	4: Tired
1	'i'	SBERT_224 variance	SBERT_224 variance	SBERT_128 variance
2	SBERT_224 variance	SBERT_128 variance	SBERT_128 variance	SBERT_224 variance
3	SBERT_128 variance	'i'	'i'	'i'
4	SBERT_342 median	SBERT_305 75th_perc	SBERT_305 median	SBERT_158 median
5	SBERT_337 mean	SBERT_286 75th_perc	SBERT_15 75th_perc	SBERT_256 75th_perc
6	flair_sentiment_min	SBERT_147 75th_perc	SBERT_4 mean	SBERT_367 75th_perc
7	SBERT_266 median	SBERT_15 median	SBERT_172 median	SBERT_180 median
8	SBERT_176 mean	SBERT_172 median	SBERT_53 variance	SBERT_244 variance
9	SBERT_265 75th_perc	SBERT_218 variance	SBERT_92 variance	SBERT_131 median
10	SBERT_205 75th_perc	SBERT_27 variance	SBERT_333 median	SBERT_291 75th_perc

Top 10 Features	5: Appetite	6: Failure	7: Concentrating
1	'i'	'i'	'i'
2	SBERT_224 variance	SBERT_128 variance	SBERT_128 variance
3	SBERT_128 variance	SBERT_224 variance	SBERT_224 variance
4	SBERT_93 median	SBERT_135 median	SBERT_171 median
5	SBERT_268 75th_perc	achieve	SBERT_120 mean
6	SBERT_100 75th_perc	SBERT_101 median	SBERT_330 75th_perc
7	SBERT_6 variance	friend	SBERT_270 variance
8	SBERT_317 75th_perc	SBERT_279 median	SBERT_41 75th_perc
9	SBERT_64 75th_perc	SBERT_186 variance	SBERT_368 mean
10	SBERT_259 median	SBERT_285 variance	SBERT_254 median

TABLE 5.3: Top 10 features for every symptom, according to SHAP feature importance. 'Moving' symptom is excluded, since the SHAP analysis gave implausible results.

predicting the same (majority) class for every data instance. However, upon inspection of predicted symptom scores of the development set, this turned out not to be the case, since development set predictions ranged from 0 to 3. This makes the zero feature importance likely to be a bug of the code rather than actual importance and therefore the *moving* symptom has been excluded from further analysis.

The top 10 most important features for the symptoms are listed in Table 5.3. The first thing that stands out is that the LIWC category '*i*', which refers to the relative use of first person singular words like 'I', 'mine' or 'me', features in the top 3 of all symptoms, as well as the variance of S-BERT features 128 and 224. Most features are S-BERT features, but we see that the minimum flair sentiment score is important for the *no interest* symptom, and the LIWC categories of *achieve* (words like 'win', 'success' or 'better') and *friend* (words like 'buddy', 'neighbor') are important for the *failure* symptom.

Training and optimising based on the obtained top N features is repeated, where N is increasing from 3 to 1644 (all features) exponentially. The resulting UAR scores on the development set are shown in Figure 5.2. The plot shows that for some symptoms, there is quite a high redundancy in features. For example, the orange *depressed* symptom line peaks at only 8 features, which is 0.48% of all features. The second peak in UAR is at all features. These results motivate to sum up all predicted development set symptoms at their best UAR setting, to see whether good total depression severity scores can be obtained. This score is compared to summing up predicted symptoms from using all features. The results are shown in Table 5.4. For

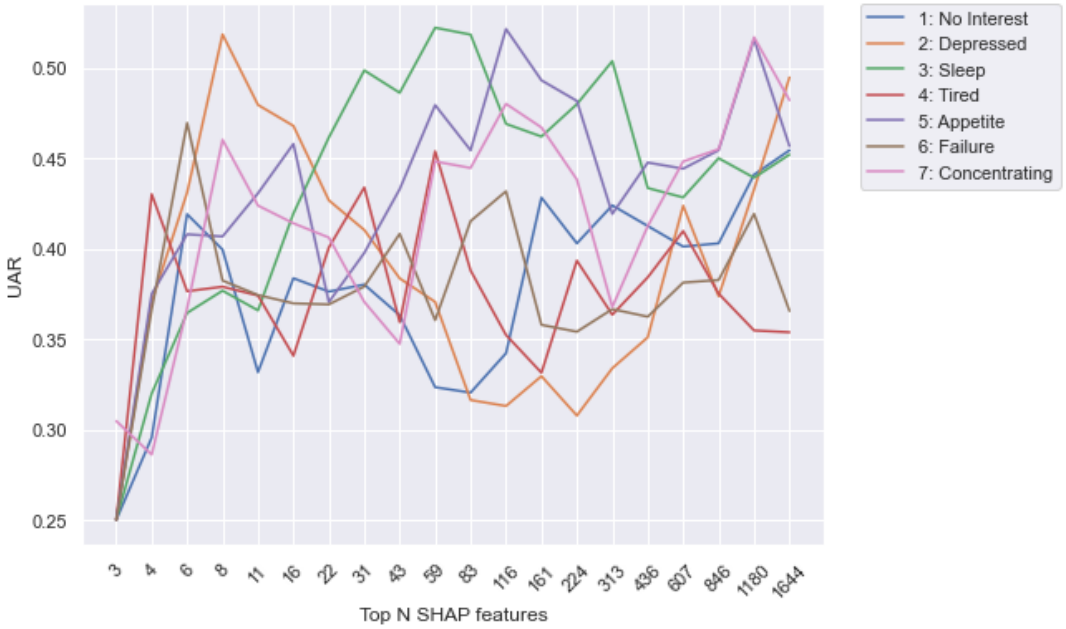


FIGURE 5.2: Top N selected features and their UAR scores for the single-task classification predicted symptoms individually.

the *moving* symptom, the average on the training set is chosen as predicted value. A large drop in performance can be seen. This might be due to the fact that optimising on UAR leads to overestimating of symptom score. Figure 5.3, which shows the predicted versus actual total depression score of the model that sums predictions of the best performing classification models, illustrates this point.

Features used	**CCC**	**RMSE**
All features	0.55	5.48
Best feature combintion	0.37	7.05

TABLE 5.4: Summation of single-task predicted symptoms, where using all features for all symptoms are compared to using the best feature set (in terms of UAR). CCC scores are on the development set.

5.3 Prediction via Symptoms and Other Intelligible Features

Until now, approaches with only predicted symptom as intermediate features have been discussed. However, since there are intelligible speech features available, in the form of LIWC, handcrafted features and eGeMAPS, it might be valuable to add these to the symptoms in order to boost performance. This is only possible in a stacking setting with cross-validation, so that the second tier model can train on predicted symptoms. It is chosen to use the single-task classified symptoms, as well as LIWC, handcrafted features and intelligible eGeMAPS features, which adds up to 140 features. Using all of these features for the second tier regressor would definitely result in an uninterpretable system. Therefore, feature selection is applied to select

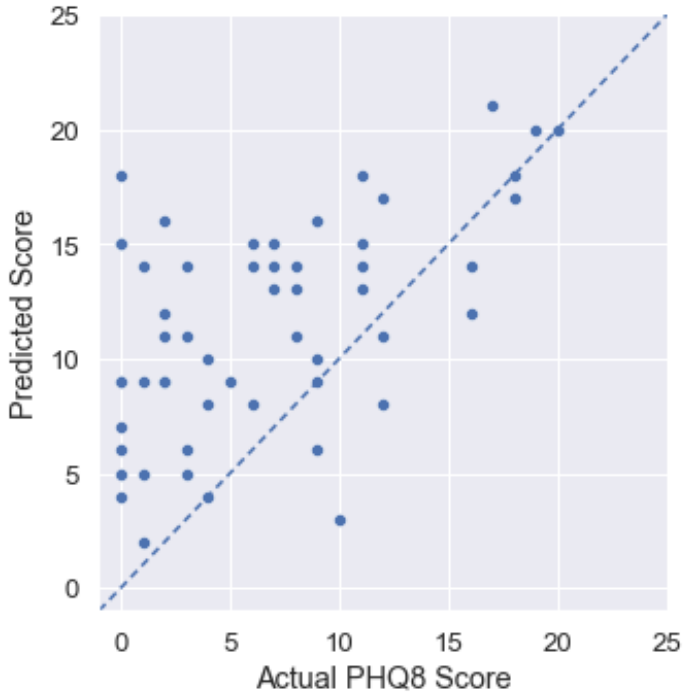


FIGURE 5.3: Predicted vs. actual PHQ8 score on the development set for the model that sums the best performing single-task classification models, obtained from the SHAP analysis.

the most important features. This is based on the performance on the development set. First the interpretable Ridge regressor model is used, where the regularisation factor is varied in the range of {0.0001, 0.0003, 0.001, 0.003, ..., 100, 300}. Inspection of the first results showed that eGeMAPS features are never selected, so in order to optimise the process, these features are removed from the feature selection method.

The best result on the development set was a CCC score of 0.71 and uses 18 features, of which 5 were symptoms, and the rest contains a mix of LIWC and handcrafted features. For interpretability reasons, the number of features would preferably be lower. The best Ridge regressor that uses less than 10 features, resulted in a development set CCC of 0.61, which is significantly lower. This model uses 8 features, where the *tired* and *failure* symptoms were replaced by the 'adverb' and 'sad' LIWC features, so the model uses mostly symptom features.

To see whether the best model (CCC of 0.71) can generalise well, I probed the test set with this. This resulted in a CCC score 0.48, which indicates that the Ridge regressor with a rich variety of other intelligible features next to symptoms does not generalise well. Therefore, a few experiments with a less transparent Random Forest as second tier model have been performed, to see whether non-linear modelling is needed for better generalisation.

These experiments have resulted a best model that performed with a 0.75 CCC score on the development set, using almost all symptom features, but feature selection replaced the *tired* symptom with the LIWC category 'adverb'. The combination of a small feature set and high performance gave reason to probe the test set once

Model	**Features**	**CCC**	**RMSE**
Ridge	1:No Interest, 2:Depressed, 3:Sleep, 7:Concentrating, 8:Moving, flair_sentiment_std, flair_sentiment_mean, speech_rate_std, speech_rate_mean, Confidence_mean, adverb, sad, cogproc, power, reward, space, death, netspeak	0.71	3.82
Ridge	1:No Interest, 2:Depressed, 3:Sleep, 5:Appetite, 7:Concentrating, 8:Moving, adverb, sad	0.61	4.25
RF	1:No Interest, 2:Depressed, 3:Sleep, 5:Appetite, 6:Failure, 7:Concentrating, 8:Moving, adverb	0.75	3.61

TABLE 5.5: An overview of models where feature selection of intelligible features have been applied. Scores are on development set.

	Model	**Setting**	**Stacking?**	**Second-tier features**	**CCC**		**RMSE**	
					Dev	**Test**	**Dev**	**Test**
Previous works	Challenge Baseline [47]				0.26	0.12	7.72	8.01
	AVEC'19 Winner [43]				-	**0.67**	4.37	4.73
	Yin et al. [62]				0.40	0.44	4.94	5.50
	Fan et al. [16]				0.47	0.43	5.07	5.91
	Makiuchi et al. [49]				0.70	0.40	3.86	6.11
	Kaya et al. [26]				0.48	0.34	-	5.88
Ours	Baseline (KELM)	Direct prediction	-	-	0.57	0.49	4.41	5.57
	Summation	ST regression	No	Symptoms	0.61	0.45	4.18	5.69
	Summation	MT regression	No	Symptoms	0.61	**0.62**	5.10	6.06
	Summation	ST classification	Yes	Symptoms	0.52	0.52	5.69	6.00
	Random Forest	ST classification	Yes	Symptoms	0.62	0.53	4.23	5.37
	Random Forest	ST classification	Yes	Intel.	0.75	0.54	3.61	5.59
	Ridge Regressor	ST classification	Yes	Intel.	0.71	0.48	3.82	5.79

TABLE 5.6: PHQ-8 prediction performance of our models and previous works in terms of CCC (challenge measure) and RMSE. ST = Single-task, MT = Multi-task, Intel. = a selection of all intelligible features, including predicted symptoms.

more, however, performance dropped again quite drastically to a CCC of 0.54.

The results of these models on the development set, including their RMSE value can be seen in Table 5.5. The performance of all models that have been probed on the test set, including those using only symptoms as intermediate features, are shown in Table 5.6. The best interpretable model's predictions on the test set, versus the actual depression score, is shown in Figure 5.4. Two points in the lower right corner draw the attention. These predictions are undesirable, since low depression scores are predicted but high depression severity is present.

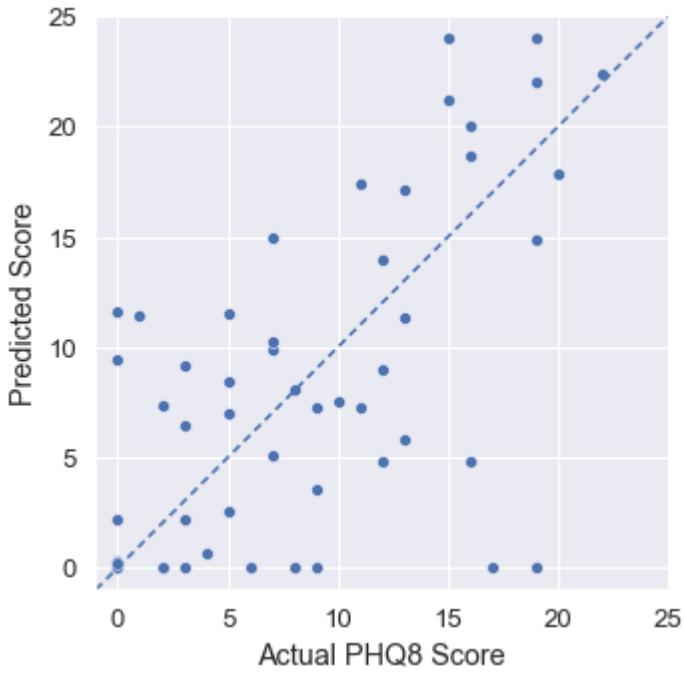


FIGURE 5.4: Predicted vs. actual PHQ8 score on the test set for the best interpretable model (summation of multi-task prediced symptoms).

Chapter 6

Discussion and Conclusion

In this chapter reflect on the research objectives and discuss these. First, an overview of the results is given in Section 6.1, then the research questions are answered in Section 6.2 and finally some limitations of the current research and suggestions for follow-up research are discussed in Section 6.3.

6.1 An Overview of the Results

In this study, I investigated whether interpretable depression prediction models are achievable, which have a competitive performance over state-of-the-art systems that employ deep architectures. To achieve this, it was hypothesized that a prediction system would benefit from first modelling symptoms which are used in the PHQ-8 questionnaire (namely, *sleep quality, lack of interest, feeling tired, appetite, failure, moving, concentrating, depressed feelings*), and then use these observable items to predict depression. I validated the proposed approaches experimentally on the AVEC 2019 dataset [47]. For prediction of eight symptoms' scores from the conversational text, I performed a set of single-task and multi-task regression, as well as single-task classification experiments with varying set of features. The results show that Weighted Kernel ELM (WKELM) classification UAR performance per symptom, outperforms the regression performance using the same learner, even though multi-task learning is employed for the regression treatment. The higher UAR performance of WKELM is partly due to the weighting trick used in the presence of class imbalance. However so, the average development set UAR performance across eight symptoms is around 0.43, which presents a large room for improvement in future research.

In addition, predicted eight symptoms are used to predict the final depression severity score (PHQ-8). The most straightforward approach is simply summing up the predictions of symptoms, as originally this gives the final depression severity score. The summation approach with the multi-task regression based predictions, obtains the second best test set performance (0.62 CCC) to date using E-DAIC corpus with AVEC 2019 challenge protocol. Surprisingly, while the average UAR performance of the single task classification based symptom prediction (0.43) is markedly higher than that of regression based prediction (0.32), the corresponding sum of the scores performs markedly poorer (0.52 CCC) on the test set compared to the sum

of regression based symptoms. This can be explained by the fact that Weighted KELM optimizing UAR performance generally causes overestimation of the symptom score, due to the imbalance of symptom scores that are skewed towards lower values. Then, final summation of symptoms also leads to overestimation of the total PHQ score.

Since the performance of symptom predictor models is not perfect, I alternatively proposed to stack the symptom predictions to a Random Forest regressor to overcome the short-comings of symptom predictors. In order to avoid overfitting and generate the symptom predictions for the training set, I used K-Fold CV and combined the respective predictions from K folds. For the development and test sets, the predictions from K-Folds are averaged and then sanitized. Here, RF stacking performance of the multi-task regression on the development set was found to be way poorer (0.35 CCC) compared to a simple summation of the predicted symptom scores (0.61 CCC). The classification based symptom predictions with RF stacking performed relatively better on the development set (0.62 CCC), while the test set performance was only slightly better (0.53 CCC) compared to the summation approach.

Also, a second tier classifier other than mere summation did not improve on the predictive performance, both in case of using symptoms only, as well as when other interpretable features can be selected by a feature selector. Moreover, the risk overfitting on the development set increased, whereas some summation models were better at generalising from development to test set.

It can be noted that all proposed direct and indirect approaches outperform the first runner up performance on the test set in terms of CCC score. In terms of RMSE, only one approach, with Random Forest as second tier regressor, performed better than the runner up of the challenge.

6.2 Answering the Research Questions

Regarding **Research Subquestion 1**, which asked *whether a direct prediction baseline can be created that performs on state-of-the-art level*, it should be answered that this was not completely achieved, although both the best KELM and SVM models would have been the first runner-up in the AVEC'19 challenge.

Research Subquestion 2 asked *whether a system with depression symptoms as the only intermediate features is capable of approximating the state-of-the-art in predictive performance, while improving the interpretability of the system.* The answer to this question is that approximation of state-of-the-art CCC score is very close, although in terms of RMSE, which punished outliers more, there is room for improvement. The symptoms-only approach turned out to be the best proposed approach, both in terms of overall test set performance as well as generalisation power.

Research Subquestion 3 is an extension to the second research subquestion, and *asks whether interpretable speech features can be added to symptoms in order to boost the*

predictive performance of the total system. This turned out not to be the case. This can be partly due to the fact that summation of symptoms as a second stage regressor is more effective and robust, but also because symptoms generalise better than interpretable speech features.

The main research question was the following: **Research Question:** *How far can the predictive performance of a state-of-the-art depression prediction system be approximated (on the AVEC'19 test set in terms of CCC), while keeping the model interpretable by using a limited set of depression symptoms, predicted from acoustic-linguistic features, and interpretable text or acoustic features as intermediate observable features?*

The answer to this question can be formulated as such: a close approximation of state-of-the-art depression prediction can be achieved in terms of CCC score, while keeping the model interpretable. The approximation in terms of predictive performance is quite close, since the test set CCC score (0.62) of the best model is much closer to the state-of-the-art on this dataset (0.67) than to the first runner-up (0.44). The interpretability has been defined according to 4 factors, of which model transparency, amount of input features and intelligibility of input features of the best prediction model improved greatly, when compared to the state-of-the-art. The accuracy of predicted symptoms leaves a large room for improvement.

6.3 Limitations and Future Research

It is concluded that state-of-the-art depression prediction systems can be quite closely approximated in terms of predictive performance, while keeping the system interpretable. Some context can be given to this conclusion.

With respect to my own baseline, and to the state-of-the-art, the best performing system is performing better in terms of model transparency, amount of input features and intelligibility of input features. Whether the accuracy of predicted symptoms is enough, is not definitively answered, but there is certainly a lot of room for improvement. An application-grounded or human-grounded evaluation of the system should provide more information on this.

Furthermore, this thesis has not succeeded in creating a system that is interpretable from beginning to end, since the predicted symptoms are predicted using a large amount of input features that are not intelligible. Further research could determine whether clinicians find this two-stage strategy, where the first stage predicts observable symptoms uninterpretably and the second stage is fully interpretable, acceptable for their daily work or not.

In a broader context, this thesis has shown that, following the approach (inspired by Rudin's [50]) of dividing a problem into its parts is a promising direction in the field of automated mental health diagnostic support. This has several advantages, ranging from bias detection (does the system recognise one symptom better than the other?) to scientific discovery (which cues correlate with a symptom?).

Here it should be added that in this thesis, I have applied the knowledge that summing up the ground-truth symptoms gives the depression score. For other mental health problems, the relation between symptoms and disorder might not be as well-validated as for depression.

Some ethical considerations should be noted. Part of the motivation for the presented research is to gain more insight in a high-stake decision system. However, depression (symptom) prediction is also a sensitive and potentially compromising activity when consent is absent. Nowadays, devices that can record voice are in most people's pocket, while there is generally little knowledge about the relative ease of obtaining such sensitive information. Therefore it should be noted that any research in the field of depression prediction, thus also this thesis, brings these problems closer to the present day.

This thesis has only trained and evaluated models on the AVEC'19 data set. A more systematic approach using different depression data sets should provide more insight into the viability of the current approach.

Furthermore, one of the major points of improvement lies at the symptom prediction. In contexts where interpretability of predicted symptoms is less of a problem, state-of-the-art models like neural networks could be used. These have as a major advantage that multi-task classification is naturally built in. If interpretability of symptom predictions is more important, there should be more focus on the search for relevant intelligible speech features.

Using UAR as a symptom optimisation metric has its drawbacks. Summation of symptoms results in overestimation of the total depression score. A weighted average of UAR and Weighted Average Recall (WAR) could therefore be promising, although there is the risk of putting too much focus on low symptom scores.

Part of the original intention of this thesis was to include predicted emotions or mood as features for depression prediction, either on their own or together with the predicted symptoms. Since emotion or mood was not available in the provided data, and annotation turned out to take up time beyond the limits of this thesis, this direction of research was abandoned. Still, since depression is a mood disorder, this could be a viable line of future study.

Bibliography

[1] Alan Akbik et al. “FLAIR: An easy-to-use framework for state-of-the-art NLP”. In: *Proc. NAACL-HTL (Demonstrations)*. 2019, pp. 54–59.

[2] Sharifa Mohammed Alghowinem et al. “Interpretation of depression detection models via feature selection methods”. In: *IEEE Transactions on Affective Computing* (2020).

[3] Flavio Avila et al. “Investigating Feature Selection and Explainability for COVID-19 Diagnostics from Cough Sounds”. In: *Proc. Interspeech 2021*. 2021, pp. 951–955.

[4] Tathagata Banerjee et al. “Predicting Mood Disorder Symptoms with Remotely Collected Videos Using an Interpretable Multimodal Dynamic Attention Fusion Network”. In: *arXiv preprint arXiv:2109.03029* (2021).

[5] Nicholas Cummins et al. “A review of depression and suicide risk assessment using speech analysis”. In: *Speech Communication* 71 (2015), pp. 10–49.

[6] Nicholas Cummins et al. “Analysis of acoustic space variability in speech affected by depression”. In: *Speech Communication* 75 (2015), pp. 27–49.

[7] Nicholas Cummins et al. “Artificial intelligence to aid the detection of mood disorders”. In: *Artificial Intelligence in Precision Health*. Elsevier, 2020, pp. 231–255.

[8] *Depression*. URL: https://www.who.int/news-room/fact-sheets/detail/depression.

[9] David DeVault et al. “SimSensei Kiosk: A virtual human interviewer for healthcare decision support”. In: *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. 2014, pp. 1061–1068.

[10] Jacob Devlin et al. “BERT: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018).

[11] Lynn V Doering and Jo-Ann Eastwood. “A literature review of depression, anxiety, and cardiovascular disease in women”. In: *Journal of Obstetric, Gynecologic & Neonatal Nursing* 40.3 (2011), pp. 348–361.

[12] JM Dopmeijer et al. “Monitor Mentale gezondheid en Middelengebruik Studenten hoger onderwijs. Deelrapport I. Mentale gezondheid van studenten in het hoger onderwijs”. In: RIVM (2021).

[13] Finale Doshi-Velez and Been Kim. “Towards a rigorous science of interpretable machine learning”. In: *arXiv preprint arXiv:1702.08608* (2017).

[14] Sergio Escalera, Oriol Pujol, and Petia Radeva. "Separability of ternary codes for sparse designs of error-correcting output codes". In: *Pattern Recognition Letters* 30.3 (2009), pp. 285–297. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2008.10.002. URL: https://www.sciencedirect.com/science/article/pii/S016786550800305X.

[15] Florian Eyben et al. "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing". In: *IEEE transactions on affective computing* 7.2 (2015), pp. 190–202.

[16] Weiquan Fan et al. "Multi-modality depression detection via multi-scale temporal dilated cnns". In: *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. 2019, pp. 73–80.

[17] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. "All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously." In: *J. Mach. Learn. Res.* 20.177 (2019), pp. 1–81.

[18] Eiko I Fried et al. "Depression is more than the sum score of its parts: individual DSM symptoms have different risk factors". In: *Psychological medicine* 44.10 (2014), pp. 2067–2076.

[19] Yuan Gong and Christian Poellabauer. "Topic modeling based multi-modal depression detection". In: *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. 2017, pp. 69–76.

[20] Jonathan Gratch et al. "The distress analysis interview corpus of human and computer interviews". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. 2014, pp. 3123–3128.

[21] Brandon M Greenwell, Bradley C Boehmke, and Andrew J McCarthy. "A simple and effective model-based variable importance measure". In: *arXiv preprint arXiv:1805.04755* (2018).

[22] Furkan Gurpinar et al. "Kernel ELM and CNN based facial age estimation". In: *Proc. CVPRW*. 2016, pp. 80–86.

[23] Dini Handayani et al. "Systematic review of computational modeling of mood and emotion". In: *The 5th International Conference on Information and Communication Technology for The Muslim World (ICT4M)*. IEEE. 2014, pp. 1–5.

[24] Daniel Jurafsky and James H. Martin. *Speech and language processing - an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall series in artificial intelligence. Prentice Hall, 2000. ISBN: 978-0-13-095069-7.

[25] Heysem Kaya, Furkan Gürpınar, and Albert Ali Salah. "Video-based emotion recognition in the wild using deep transfer learning and score fusion". In: *Image and Vision Computing* 65 (2017), pp. 66–75.

[26] Heysem Kaya et al. "Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics". In: *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. 2019, pp. 27–35.

[27] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of NAACL-HLT*. 2019, pp. 4171–4186.

[28] Kyoung-jae Kim and Hyunchul Ahn. "A corporate credit rating model using multi-class support vector machines with an ordinal pairwise partitioning approach". In: *Computers & Operations Research* 39.8 (2012), pp. 1800–1811.

[29] Kurt Kroenke et al. "The PHQ-8 as a measure of current depression in the general population". In: *Journal of affective disorders* 114.1-3 (2009), pp. 163–173.

[30] Puneet Kumar, Vishesh Kaushik, and Balasubramanian Raman. "Towards the explainability of Multimodal Speech Emotion Recognition". In: *INTERSPEECH 2021*. ISCA. 2021, pp. 1748–1752.

[31] Jionghao Lin et al. "An explainable deep fusion network for affect recognition using physiological signals". In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2019, pp. 2069–2072.

[32] Daniel Low, Kate Bentley, and Satrajit Ghosh. "Automated assessment of psychiatric disorders using speech: A systematic review". In: *Laryngoscope Investigative Otolaryngology* 5.1 (2020), pp. 96–116.

[33] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Proceedings of the 31st international conference on neural information processing systems*. 2017, pp. 4768–4777.

[34] François Mairesse et al. "Using linguistic cues for the automatic recognition of personality in conversation and text". In: *Journal of artificial intelligence research* 30 (2007), pp. 457–500.

[35] George A Miller. "The magical number seven, plus or minus two: Some limits on our capacity for processing information." In: *Psychological review* 63.2 (1956), p. 81.

[36] Christoph Molnar. *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. 2019.

[37] Anh Nguyen et al. "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks". In: *Advances in neural information processing systems* 29 (2016), pp. 3387–3395.

[38] Harsha Nori et al. "Interpretml: A unified framework for machine learning interpretability". In: *arXiv preprint arXiv:1909.09223* (2019).

[39] Judith Norman. "Gender bias in the diagnosis and treatment of depression". In: *International Journal of Mental Health* 33.2 (2004), pp. 32–43.

[40] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. "Feature Visualization". In: *Distill* (2017). https://distill.pub/2017/feature-visualization. DOI: 10.23915/distill.00007.

[41] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[42] James W Pennebaker et al. *The development and psychometric properties of LIWC2015*. Tech. rep. 2015.

[43] Anupama Ray et al. "Multi-level attention network using text, audio and video for depression prediction". In: *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. 2019, pp. 81–88.

[44] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proc. EMNLP*. 2019.

[45] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016, pp. 1135–1144.

[46] Fabien Ringeval et al. "AVEC 2017: Real-Life Depression, and Affect Recognition Workshop and Challenge". In: *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. AVEC '17. Mountain View, California, USA: Association for Computing Machinery, 2017, 3–9. ISBN: 9781450355025. DOI: 10.1145/3133944.3133953. URL: https://doi.org/10.1145/3133944.3133953.

[47] Fabien Ringeval et al. "AVEC 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition". In: *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. 2019, pp. 3–12.

[48] Marko Robnik-Šikonja and Marko Bohanec. "Perturbation-based explanations of prediction models". In: *Human and machine learning*. Springer, 2018, pp. 159–175.

[49] Mariana Rodrigues Makiuchi et al. "Multimodal fusion of BERT-CNN and gated CNN representations for depression detection". In: *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. 2019, pp. 55–63.

[50] Cynthia Rudin. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead". In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215.

[51] Björn Schuller. "Voice and speech analysis in search of states and traits". In: *Computer Analysis of Human Behavior*. Springer, 2011, pp. 227–253.

[52] Ramprasaath R Selvaraju et al. "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626.

[53] Karan Sikka et al. "Multiple kernel learning for emotion recognition in the wild". In: *Proceedings of the 15th ACM on International conference on multimodal interaction*. 2013, pp. 517–524.

[54] Daniel Smilkov et al. "Smoothgrad: removing noise by adding noise". In: *arXiv preprint arXiv:1706.03825* (2017).

[55] Gizem Sogancioglu, Heysem Kaya, and Albert Ali Salah. "Can mood primitives predict apparent personality?" In: *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE. 2021, pp. 1–8.

[56] Gizem Sogancıoglu et al. "Is Everything Fine, Grandma? Acoustic and Linguistic Modeling for Robust Elderly Speech Emotion Recognition". In: (2020).

[57] Centraal Bureau voor de Statistiek. *Psychische gezondheid in Nederland*. 2021. URL: https://www.cbs.nl/nl-nl/maatwerk/2021/43/psychische-gezondheid-in-nederland.

[58] Bo Sun et al. "A random forest regression method with selected-text feature for depression assessment". In: *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. 2017, pp. 61–68.

[59] Hao Sun et al. "Multi-modal adaptive fusion transformer network for the estimation of depression level". In: *Sensors* 21.14 (2021), p. 4764.

[60] Sana Tonekaboni et al. "What clinicians want: contextualizing explainable machine learning for clinical end use". In: *Machine learning for healthcare conference*. PMLR. 2019, pp. 359–380.

[61] Haofan Wang et al. "Score-CAM: Score-weighted visual explanations for convolutional neural networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2020, pp. 24–25.

[62] Shi Yin et al. "A multi-modal hierarchical recurrent neural network for depression detection". In: *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. 2019, pp. 65–71.

[63] Ziping Zhao et al. "Hierarchical attention transfer networks for depression assessment from speech". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 7159–7163.

[64] Weiwei Zong, Guang-Bin Huang, and Yiqiang Chen. "Weighted extreme learning machine for imbalance learning". In: *Neurocomputing* 101 (2013), pp. 229–242.