MSc Artificial Intelligence - Utrecht University

MSc Thesis

TITLE

François Blom 5988918

Supervisor: Dr. A. Akdag

Daily Supervisor: PhD M.F. Pessanha

Second Examiner: Dr. H. Kaya

December 22, 2022

Contents

Abstract

1 Introduction

1.1 Research Motivation

Depression is one of the most common mental health disorders affecting over 300 million people globally [1]. Over the past 15 years, the number of people diagnosed with depression has increased tremendously [2]. Research by Ettman et al. suggests that the COVID-19 pandemic has amplified the general trend of increase in depression rates even further [3]. It is estimated that on a global level depression is the number one cause of disability [2].

Depression negatively affects individuals. Symptoms occur on the mental level as well as the physical level (psycho-motor effects) [4]. Depression is often characterized by symptoms such as negative thoughts, tiredness, low self-esteem and disinterest. It can cause severe negative consequences to the individual if not treated properly. The most prevalent consequence of depression is, as stated before, disability. However, depression can also lead to suicide in severe cases. Taking these negative effects together it becomes clear that depression negatively affects individuals, but also society as a whole due to the disability it causes and the grief that is caused by the loss of a relative or friend. Therefore, correctly diagnosing and subsequently treating depression is of great importance.

Currently, diagnosing depression occurs mainly through two channels: self-assessment and interviews with a clinical psychologist. However, in this process a lot of misdiagnosis occurs [5]. Such misdiagnoses cause individuals with depression to receive the wrong treatment and may also cause people without depression to receive anti-depressants, which in turn might cause undesirable consequences. In order to reduce the amount of misdiagnoses of depression, it is important to decrease errors caused by subjective interpretation and missed cues. Therefore, finding a more objective and precise approach towards the diagnosis of depression is of great value as that would allow for a decrease in misdiagnoses. Recent advancements in technology have enabled us to quantify and measure behavioral cues that are hard or impossible to notice by humans. Because of the fact that these cues can be quantified they can also be used in predictive models. Therefore, such models are no longer restricted to using only transcripts of conversations with (possibly) depressed people. Combining domain knowledge from the field of psychology and advancements made in the field of artificial intelligence has enabled researchers to make substantial progress in the process of building models that can predict the presence of depression in an individual. The use of multiple types of data in predictive models is referred to as multi-modal analysis in the literature.

Although the introduction of multi-modal analysis for depression of detection has gained a lot of attention in recent years, multiple issues remain with respect to general consensus regarding which types of data (predictive features) yield high predictive power for the detection of depression among patients. This is

partly attributed to the fact that the best performing models are often ensemble models that incorporate deep learning model architectures. Such architectures attain high performance, but are less interpretable by humans. This low interpretability causes researchers to be unable to validate the model against psychological theory and also creates a risk where researchers lose the ability to explain on what grounds a patient is diagnosed as depressed. The goal of developing models that can automatically predict the presence of depression among patients is not to replace the human clinicians. Instead, it aims to create a new dynamic where the clinician is empowered by the new technology and is able to use these predictive tools to improve the overall quality of work delivered by clinicians. Thus, these predictive models should not be seen as stand-alone technologies but as tools that can be leveraged by professionals. Again, it is therefore very important that these models remain to be understandable for the clinicians as they are the ones who need to use them and (at least partially) rely on the predictions made by these models.

Additionally, as has been stated above, the process of diagnosing patients with depression occurs based on subjective interpretations of clinicians. In practice, this leads to there being disagreement between clinicians regarding what the correct diagnosis is for a patient. In such situations, the presence of a tool that can suggest a baseline diagnosis and also indicate on which grounds the diagnosis is made can facilitate the process of multiple clinicians reviewing patient cases together. In short, it is important to focus on the development of robust, explainable and well-performing predictive models.

1.2 Research Objectives

This research aims to contribute to the field of research dedicated to the automated detection of depression by investigating the role that audio-based features play. The motivation to focus on these two comes from the intuition that recording clinical interviews on video brings along multiple complications. Firstly, clinical interviews tend to include sensitive and personal topics. Therefore, it is of great importance to ensure the safety of the matters discussed in such interviews. Video recording cannot be anonymized easily. Instead, software is needed to do this. This complicates the entire process of using video-based data for depression detection while respecting the anonymity and privacy of patients. Secondly, the presence of a recording camera in a clinical setting might have undesirable consequences. In a clinical interview, it is important that the patient feels comfortable and that a setting is created in which everything can be shared. If during a clinical interview the patient knows that they are being recorded on camera, the answers they give might be influenced by the presence of the camera and possibly lead to social-desirability bias present in the answers given by the patients. This complicates the process of correctly diagnosing patients.

Furthermore, this research intends to evaluate which paralinguistic features yield the highest predictive

power in order to create a benchmark of features that are indicative of the presence of depression. Additionally, The aim is to also investigate which breathing-related features are the most indicative of the presence of depression. Lastly, this research aims to evaluate the relationship between paralinguistic features and breathing-related features.

1.3 Thesis Outline

The rest of this thesis will be organized as follows: In the second section, a literature review in which related work as well as relevant theoretical background will be provided. Next, An overview of the research question (and sub-questions) will be provided. In the fourth chapter, a more elaborate explanation will be given regarding the methodologies as well as the experimental setups that have been used. This will be followed by the fifth chapter, in which the results of the experiments will be addressed. The last chapter will conclude by drawing conclusions, summarizing the findings and limitations of the study providing directions for further research.

2 Literature Review

2.1 Affective Computing

Affective computing is a field of research that lies at the intersection of psychology, computer science and cognitive science. In the literature, affective computing is defined as the overarching term for the automation of practices involving simulation, classification and interpretation of (human) emotions [6] [7]. This research will focus specifically on the branch of affective computing that involves the automated classification of depression based on data originating from different modalities.

2.2 Depression Cues

In order to be able to diagnose individuals with depression correctly, cues must be identified that indicate depression among individuals. In this section an overview will be given of recent research towards possible cues indicating the presence of depression. However, different measures of depression exist. Therefore, before going into details about the different cues, the different measures and the differences between them will be discussed.

The depression cues are split into two different modalities. First, the paralinguistic features will be addressed, these revolve around the non-linguistic properties of an individual's communication. Secondly, the linguistic cues will be discussed, these cues revolve around the language that an individual uses while communicating.

2.2.1 Measures of Depression

There are multiple ways of classifying depression. The classification occurs mainly through two processes. Either a clinical psychologist classifies the patient in accordance with a chosen framework, or the patient takes a self-assessment test. Furthermore, it should be noted that the classification of depression often occurs on a scale (discrete variable). Occasionally, depression is classified on a binary scale. This section provides an overview of the most relevant measures of depression as used in the literature.

Firstly, the Hamilton Depression Rating Scale (HDRS), this instrument is intended to be used by clinical psychiatrists after interviews with the patient. The instrument consists of 17 questions, each question is ranked on a scale (mostly three or five points). In order to maximize accuracy of the instrument, it is recommended to have the questions of the instrument filled in by two clinicians independently after the interview and consequently sum the two scores. The general classification of scores suggests that scores below 7 indicate the absence of depression. Scores between 7 and 17 indicate mild depression, scores between 18 and 24 suggest moderate depression and scores above 24 indicate severe depression [8].

Secondly, there is the Beck Depression Inventory (BDI). This instrument main difference with the HDRS is that it is intended to be a self-report questionnaire. The instrument consists of 21 questions, aiming to identify the presence and severity of depression amongst individuals. Each item is scored on a three point scale, with the exception of two questions, which are scored on a seven point scale. The total score is computed by summing the scores for each individual question. Scores between 0 and 13 indicate minimal presence of depression, scores between 14 and 19 indicate mild depression, scores between 20 and 28 indicate moderate depression and scores above 28 indicate severe depression [9].

Thirdly, the Quick Inventory of Depressive Symptomatology (QIDS) is an instrument containing 16 items. The instrument aims to classify severity of depression among patients. The items are all ranked on a 3 point scale. The 16 items in the questionnaire are selected based on diagnostic symptoms of depression as described in the Diagnostic and Statistical Manual of Mental Disorders (DSM). It should be noted that the instrument exists in two different formats: one intended to be used by the clinician (referred to as the QIDS-C), the other one is intended to be filled in by patients themselves, thus self-reporting (referred to as QIDS-SR). Scores below 6 suggest the absence of depression, scores between 6 and 10 suggest mild depression, scores between 11 and 15 indicate the presence of moderate depression, scores between 16 and 20 indicate severe depression and scores above 20 suggest very severe depression [10].

Lastly, there is the Patient Health Questionnaire 9 (PHQ-9) this instrument aims to enable patients to self-report their symptoms and consequently come to a severity classification. The questions are based on each of the 9 criteria described in the DSM. The questions are scored on a 3 point scale. It should be noted that the scoring occurs based on frequency of symptoms, a score of 0 indicates the absence of a symptom whereas a score of 3 indicates the daily presence of the scored symptom. The instrument contains 9 questions in total and the final score is computed by summing the scores of all questions. Total scores below 5 suggest minimal depression, scores between 5 and 9 indicate mild depression, scores between 10 and 14 suggest moderate depression, scores between 15 and 19 indicate moderately severe depression and scores above 19 suggest severe depression [11].

2.2.2 Paralinguistic Features

Previous research in the field of psychology suggests that changes in cognitive as well as physical states can cause differentiation in acoustic properties [12]. At the same time, research by Christopher & MacDonald [13] suggests that depression affects all components of the working memory of a patient. More specifically, the experiments find substantial impairments in the phonological loop and the visuospatial sketch pad of depressed people. Furthermore, research by Baddeley [14] suggests that disorders in the working memory can cause implications with respect to the production and processing of language. Thus, previous research

in the field supports the notion that depression affects one's ability to speak and to communicate.

By combining the fact that depression affects one's cognitive and physical state with the knowledge that changes in cognitive and physical states can cause differentiation in acoustic properties of speech and communication, it is hypothesized that acoustic features can serve as objective indicators of the presence of depression among humans [15]. This hypothesis has been fundamental in the use of vocal features as predictors for depression. The non-verbal properties of a person's communication are often referred to as 'paralinguistic features', these features are generally extracted from vocal (audio) data. This subsection will review the most relevant contributions with respect to the predictive power of vocal features for the detection of depression.

Research by Hönig et al. [16] has investigated the role of specific acoustic features in the prediction of depression by performing statistical analyses on a multitude of features. The analysis starts by building a multiple linear regression model that incorporates 3805 acoustic features. This model reaches a performance of 0.44 and is used as a baseline model. Secondly, the features are manually grouped together based on what they measure. The subset of features grouped together as 'voice quality' consists of 7 features and achieves an accuracy of 38%. Next, the group 'prosody', consisting of 17 features achieves an accuracy of 36%. The third group consists of 10 spectral features. The regression model containing these features achieves an accuracy of 29%. Combining the features of the three groups (34 features into a single regression model leads to an accuracy of 39%. Lastly, a model has been created with the same number of features (34) but in this case, the features were automatically selected through a greedy search algorithm. This model attained an accuracy of 36%. From this, the authors conclude that the model containing features that were selected on a theoretical basis outperforms the model containing automatically selected features. This can be attributed to the fact that a greedy search algorithm was used, such an algorithm does not always find a global maximum in terms of features that lead to the highest possible accuracy.

The second analysis performed by the authors focuses on the importance of specific features instead of groups of features. Features are considered as important if their correlation to the target variable (degree of depression, as measured by the BDI) exceeds or is equal to 0.25 (either positive or negative). This selection procedure results in 11 features. The exact descriptions of these features can be found in the original paper. In summary, this analysis finds that a lower speaking-pace is positively correlated with increasing levels of depression. Furthermore, high variation in loudness of speech is negatively correlated with increasing levels of depression. In other words, monotonic speech (no change in loudness) is positively correlated with increasing levels of depression.

Other independent research conducted by Cannizzarro et al. [17] has investigated the effect of similar features. The analysis conducted by the authors consists of a multivariable linear regression. The target

variable used is the degree of depression. In this case, the degree of depression was expressed in terms of HDRS rating. Furthermore, it should be noted that the data on which the regression was done contains 7 subjects, of which 6 are classified as moderate/severely depressed and 1 as mild/moderately depressed. The independent variables used in the analysis are speaking rate, pitch variation and percent pause. Out of these three variables, only speaking rate showed a significant (negative) correlation with the dependent variable; severity of depression. Pitch variation showed a negative correlation with the independent variable but did not meet the significance threshold ($p = 0.0581$) Lastly, percent pause showed a positive, but very insignificant correlation with the dependent variable ($p = 0.1997$).

Additionally, research by Mundt et al. [18] finds similar effects. The study conducted by Mundt et al. aims to verify the generalizability of acoustic markers (for depression) found in previous studies. In the study, 105 depressed people have been interviewed and their speech during these interviews has been analyzed. For every speech sample, 12 acoustic properties have been extracted. 6 of these acoustic measures returned to be significantly correlated with severity of depression (measured with the QIDS-C framework). The correlated dependent variables suggest that there is a negative effect of speaking rate and severity of depression. This indicates that severe depression is related to lower speaking rate. Furthermore, the correlated variables suggest that depressed people take longer and more breaks while speaking.

Taking these independent studies together, it becomes clear that the same acoustic features play a role in the detection of severity of depression. This leads to the conclusion that there are two main indicators of depression with respect to the acoustic properties of speech. The first one being monotonic speaking. The second one is slow speaking. Especially this second indicator aligns with the theory. As has been stated before, depression affects the working memory, which is (partially) responsible for the production of speech. Therefore, the impairments in the working memory caused by depression may indirectly cause a decrease in speed of production of speech.

In this paragraph, an overview will be given of previous work in the field of sentiment analysis towards non-verbal paralinguistic features, specifically breathing. Boiten et al. [19] suggest that different emotions result in different breathing patterns and that there is a correlation between emotional states and breathing patterns. Looking specifically at depression, depression is often seen as a state of emotion or as a cause for a change in emotional state. Therefore, it can be hypothesized that depression also causes changes in breathing patterns. This hypothesis is strengthened by the fact that depression is in some cases the undesirable consequence of a traumatic experience. Research by Ogden & Minton [20] finds that individuals who have experienced traumatic events can encounter changes in breathing patterns when discussing the trauma. Therefore, breathing-related features should theoretically be able to be used in automating the detection of depression with machine learning models. Anxiety can be one of the symptoms of depression,

Masaoka & Homma [21] have researched the effects of increased anxiety on breathing patterns. From this research, they find that individuals with high anxiety tend to increase the rate at which they breath.

2.2.3 Linguistic Features

In addition to acoustic features, research has also been done on the predictive power of linguistic features. The analysis of linguistic features falls under natural language processing (NLP) and will be discussed in this section. Although the focus of this thesis will not lie on the usage of text-based features, it is important to briefly address the previous work done as previous research has shown that text-based models tend to be the best performing in the automated detection of depression.

The hypothesis supporting that depressed people use different language in their communication stems from the cognitive theory of depression which claims that depressed people experience their surroundings through a negative scope, leading to the use of more negative language [22]. Additionally, Pyszczynski & Greenberg [23] suggest that depressed people tend to focus more on themselves compared to non-depressed individuals. This is attributed to the fact that depression arises from a decrease in self-worth, leading to a vicious cycle of self-focus. Consequently, the depressed individuals tend to talk more about themselves than non-depressed people as self-focused thoughts are more prevalent, which translates to more self-focused use of language.

These hypotheses has been extensively evaluated through empirical research. A study conducted by Rude et al. [24] researched the differences in language used by depressed people, people who were depressed in the past and people who have never been depressed. The study analyzed the language used in essays written by college students. The students were instructed to write about their feelings regarding the start of college. As has been stated before, the participants were split into three groups: depressed, previously depressed, not depressed. The groups have been made based on the BDI and the IDD-L. The written essays were consequently analyzed using the LIWC software. The results from the conducted analysis suggest that there is a significant difference in usage of the first person word 'I' between depressed and non-depressed people. No significant difference was found between the previously depressed and non-depressed people with respect to use of first person words.

Research by Stirman & Pennebaker [25] find similar results in their study analyzing the words used by poets in their written poems. The authors analyze the poems of both suicidal as well as non-suicidal poets in order to evaluate whether there are specific words that are used more by one group than the other.

The analysis conducted by the authors consists of a comparison between the types of words used (specific words as well as categories of words) by both non-depressed and depressed poets. Furthermore, the chosen poems from the poets have been grouped together based on career phase. From the results of the analyses the

authors conclude that suicidal poets use more words related to themselves (self-references) than non-suicidal poets. However, no difference was found in the amount of references to other people. Furthermore, the authors conclude that there is no significant difference in the amount of negative (or positive) words used between suicidal and non-suicidal poets.

Ramirez-Esparza et al. [26] have conducted similar research with the main objective being to shine light on linguistic markers of depression. In their study, two experiments have been conducted. The first experiment aims to validate linguistic markers found in previous studies. The second experiment aims to investigate whether there is a difference in themes that are brought up by depressed people from different nationalities (and thus, languages). The authors analyzed pieces of text posted to online forums, both English and Spanish written texts were used.

The results from the first study, focusing on linguistic markers of depression suggest that both in English and in Spanish, depressed people use more self-referencing pronouns. Furthermore, a significant difference was found between depressed women and non-depressed women with respect to the usage of negative words. Depressed women tend to use more negative words (again both in English and in Spanish) than the non-depressed women in the control group. This aligns with the cognitive theory of depression and the self-focus theory of depression discussed earlier in this section.

The second study addresses the difference in topics brought up by depressed people from different nationalities. The results from this experiment provide a granular list of words used per nationality. Generally speaking, one can conclude that English depressed individuals tend to focus on the medical implications and concerns they have. On the opposite, Spanish depressed individuals focus much more on relational and emotional topics in their discourse. This difference between nationalities suggest that linguistic markers for depression differ per language and nationality. The research that will be conducted in this specific study focuses on English-speaking individuals. Therefore, the markers that apply most (in accordance with this study) are those that fall under topics as medication and health.

Lastly, Wang et al. [27] propose a way of detecting depression among blog-posts on social networks. The study aims to provide a model that is able to detect depression based on sentiment analysis. The feature selection for the model is done based on underlying psychological theory. The significance of each feature is analyzed in order to address the theoretical validity. From this significance analysis it becomes clear that the amount of first person pronouns is a valid predictor for depression as well as the polarity of sentences used. Sentence polarity refers to the general connotation of an entire sentence. Sentence polarity can be seen as closely related to the amount of negative words used in piece of text or transcript. Lastly, it should be noted that other features also returned to be significant predictors of depression. However, these features are closely related to the usage of emoticons in written text. As the textual analysis of this research relies

only on conversation transcripts, no emoticons are present. Thus, such features cannot be considered in this specific study.

In summary, psychological theory seems to indicate that there are two main factors that differentiate language used by depressed people as opposed to non-depressed people. Firstly, due to the negative scope through which depressed people often perceive the world, their language consequently also contains more words with a negative connotation. Secondly, depressed people think more about themselves than non-depressed people. This translates to a difference in number of self-referencing pronouns being used between depressed and non-depressed people. More specifically, depressed people refer more often to themselves when speaking than non-depressed people do. The multiple empirical studies described above find results that align with these psychological theories. By combining the two we can conclude that there are mainly two linguistic markers for depression. The presence of self referencing pronouns as well as the presence of negative words indicate the presence of depression. Furthermore it should be noted that different languages and cultural backgrounds cause for a variety of additional linguistic markers as trends can be found regarding specific topics brought up by depressed individuals from different nationalities (and different languages).

2.3 Automated Depression Detection

This section provides an overview of previous research done in the field of automated depression detection. In general, the research can be split into two categories: uni-modal predictive models and multi-modal predictive models. This research will focus on predictive models that rely on textual and paralinguistic features. Therefore, firstly, text-based research will be discussed. Secondly, paralinguistic-based research will be addressed. Lastly, the combination of these two modalities in predictive models will be considered.

2.3.1 Text-based Analysis

Text-based models work such that a predictive model is trained on a set of features that are extracted from pieces of text. Research by Amanat et al. [28] has investigated the performance of deep learning model architectures on the detection of depression based on features extracted from text excerpts. The conducted text pre-processing consists of removing stopwords and non-words. Next, the text excerpts have been tokenized. For each token, the words have been lemmatized and stemmed, to reduce the number of different words in the data. Lemmatizing helps in reducing the number of different words in a piece of text as it converts words to their 'basic' form. Stemming is a technique that also decreases the number of different words in a text. This is done by converting every word in a text into its root word. Converting words into more general forms with these techniques aids in reducing the number of different words as slightly different words (e.g. worked, working into work) are converted to the same word. The stemmed and lemmatized

words then become the textual tokens. The tokens are subsequently used as input features for the predictive model.

After the data has been pre-processed and the features have been extracted, a deep learning model is trained. Due to the fact that textual data has temporal components, a model architecture needs to be chosen that can incorporate temporality. The authors have chosen to use a recurrent neural network (RNN) to do this. Additionally, regular RNNs tend to suffer from the vanishing gradient problem, which makes it harder for the network to update weights over time. Therefore, the authors have chosen to add LSTM-units to the network. These units take care of the memory of important words such that the network can handle long textual sequences without there being a vanishing gradient problem.

The split between data used for training and testing of the model is 90% to 10%. Furthermore, the authors have used 10-fold cross-validation to obtain the highest results. The target variable of the used dataset is a binary variable, each instance is thus classified as either depressed or non-depressed. The authors report that the final model has an accuracy of 99%, a precision of 98%, a recall of 99% and a F1-score of 98%.

However, it should be noted that the used textual data originates from Twitter and that the used tweets originate from users that were hand-picked. Therefore, some bias may be present in the data. Furthermore, no clear information is given regarding how the tweets have been labelled. Thus, in order to validate the performance of the model it would be good to test it on data that originates from clinical interviews and that has been labelled in a clinical setting. Nevertheless, this research has shown that deep learning model architectures are suitable for depression detection.

2.3.2 Paralinguistic Analysis

This section focuses on previous research towards paralinguistic-based depression detection models. These models rely mostly on the extraction of features through external software as inputs for predictive models.

He & Cao [29] have researched the automation of depression detection through audio data with deep learning model architectures. The authors aim to predict severity of depression among patients (expressed in terms of the BDI-II) based on audio fragments containing speech of the concerned patients. The process of feature extraction done is two-fold, both hand-crafted features and deep-learning features are extracted from the raw audio data

Firstly, the hand-crafted features can be split into two groups. One set of features is extracted using the openSMILE software (this software will be addressed extensively in a later section). In short, openSMILE extracts low-level descriptors from audio files and allows the user to apply statistical functionals to these descriptors to finally obtain a set of predictive features. The other set of hand-crafted features is obtained by applying Median Robust Extended Local Binary Patterns (MRELBP) to the spectogram of the audio

file. The MRELBP is used to determine the spatial structure of images (measured in different terms, which become the predictive features in this case), due to the fact that a spectogram is an image-representation of audio, it can also be applied to this scenario.

Regarding the deep-learning based features, again a split can be made into two groups of features. The first set contains features that are extracted from the audio files. The raw audio files are inputted into the CNN and the network is trained to extract feature representations that become more abstract the more layers are used. The second group consists of features that originate from inputting a spectogram (which can be treated as an image) of an audio file in to a CNN and during training, the network is then trained to obtain different level representations that serve as predictive features.

Next, four CNNs are trained, each on a separate set of features (the ones described above). Lastly, the final prediction is computed by taking the average of the four predictions. Regarding the performance of the single models the authors report that the models trained on deep-learning features outperform the models trained on hand-crafted features. However, the model that fuses the predictions from all four individual models performs the best. The authors argue that this is an indication that a combination of hand-crafted features and deep-learning features yields the best results to detect depression.

Chlasta et al. [30] have performed different research towards the efficiency of deep learning towards depression detection. The authors focus on the use of pre-trained CNNs to predict depression. The main difference between this research and the research conducted by He & Cao is that Chlasta et al. focus solely on pre-processing the data such that it can be inputted in a pre-existing model whereas he & Cao focus on the extraction of different types of features and consequently train a new model with the extracted features as inputs.

Regarding the pre-processing of the data, in order to be able to predict the presence of depression with the pre-trained CNNs the data format needs to be altered from an audio file to an image of size 224 x 224 pixels. This has been done by creating wavelets out of the .WAV files (the raw audio files). Additionally, spectograms have been extracted from each audio file. These spectograms subsequently serve as the input images for the pre-trained CNNs (different versions of the ResNet model). Although the CNNs have been pre-trained, the authors perform hyperparameter tuning to ensure that the model performs optimally on the spectogram images. After hyperparameter tuning, the best performing model achieves an accuracy of 77%. However, each model was trained on a different subset of the data. Therefore, the different models cannot be fairly compared as differences in performance might be caused by higher or lower variance in the training data used for that specific model.

2.3.3 Multi-modal Analysis

This section will focus on previous research towards models that rely on features from the two modalities discussed before, paralinguistic and text-based features.

Alhanai et al. [31] combine the audio and text modalities to create a depression prediction model without explicit performing topic modeling of the data. This implies that no manual pre-processing is done regarding which data (from clinical interviews) is used to predict depression through the model. The authors have conducted three experiments, first a logistic regression model has been fitted regardless of the type of questions asked. Next, a logistic regression model was fitted while considering which questions have been asked and the answers given to specific questions. Lastly, a LSTM model was trained without prior information regarding which questions have been asked. Instead, the LSTM model relies on the sequences of responses that have been given by the patient.

For the first experiment, the logistic regression model incorporates text features that have been extracted from the raw interview transcripts through Word2Vec, this is a technique that vectorizes chunks of text in order to be able to use them in predictive models. Each question and each response was vectorized. The audio features used as inputs for the first experiment have been extracted from the raw audio file using the COVAREP software. The features consist of statistical functionals of descriptors that have been extracted through the COVAREP software. Next, all features that have shown an insignificant correlation were left out of the final model.

The features used as inputs in the second model are the same as in experiment one. However, in experiment two, the authors have aimed to assign weights to certain parts of the interviews. This was done to check if the predictive power of different features changes based on the type of question that is being asked (the topic of the question, and thus also of the response). In order to do this, each answer given by the patient was evaluated based on its predictive performance. If the predictive power was above a specific threshold it is given a weight in accordance with the computed predictive power. This leads to the assignment of weights to answers that are most indicative of depression without manually evaluating the importance of specific queries.

Lastly, the LSTM model was trained. This allowed the authors to incorporate temporality and relations between different queries in the model. For both modalities, a separate LSTM model was trained and the outputs of the two models were concatenated and inputted into a final feed forward network which computes a final prediction score.

The authors find that when comparing the performance of the three models, the first model outperforms the second model in terms of RMSE and MAE and the LSTM model performs the best as it yields the lowest

RMSE and MAE. This suggests that the approach chosen for the automation of content modeling might not have been effective. In the literature there are multiple examples of selecting text-features based on topic modeling. However, the results vary between different instances of research. Therefore, no conclusive statement can be made regarding the effectiveness of the approach. The authors suggest that the lower performance of the second model might be attributed to the fact that the model was not calibrated to perform well om multi-class problems.

Ye et al. [32] have also conducted research towards multi-modal depression detection. Their research focuses on building a model that predicts a depression score for Chinese native speakers based on audio and text. Two experiments have been conducted, the first one is a reading experiment in which the participants were supposed to read a Chinese text out loud. The text consists of both negative and positive parts. The audio fragments containing the positive parts of the text were separated from the negative ones and for both audio features have been extracted with the openSMILE toolkit. In addition, deep-learned features are extracted by inputting the raw data in a pre-trained VGG model. The second conducted experiment consists of a conversation between the participant and the clinician, in the conversation the clinician poses questions related to the HAM-D depression scoring framework and the participant answers them. Next, the answers given by the participant are transcribed using the Baidu software. Features are extracted from these conversation transcripts by vectorizing each word with the Word2Vec model. Next, two uni-modal models are trained. For the audio-based model, a temporal convolutional network (TCN) is trained. Opposed to regular CNNs, TCNs do not have the problem of vanishing gradient and are therefore more suitable for the modeling of long-term temporal relations between words. For the text-based model a tranformer model (also deep-learning model architecture) is trained. The transformer model takes the word vectors (text-based features) and the locations of the words as inputs. After having trained the two uni-modal models, they are combined into a single multi-modal model. This is done by concatenating the outputs of the two models and using it in another neural network that outputs a final binary classification (either depressed or non-depressed. The fused model achieves an accuracy of 91.2% and a F1-score of 90.6%.

2.3.4 Evaluation of Models

Having discussed the previous work done in the field of depression detection, this paragraph will provide an assessment of the chosen approaches. With respect to the text-based model, the pre-processing steps work well as they aid in reducing the number of features and thereby decrease the probability that the model will overfit on the training data. However, regarding the chosen model architecture, a deep learning architecture is chosen. This makes it hard for humans to interpret how the model makes predictions. Therefore, it would be better to use a model architecture without a black-box component. In terms of quantitative performance,

this model performs the best out of all models that have been addressed. However, due to the fact that the data originates from Twitter and was labelled in a non-verifiable setting, these results might not be indicative of the performance of the chosen approach.

Secondly, looking at the audio-based models, He & Cao choose to extract both hand-crafted and deep learning features. Their results suggest that using both type of features in a model yields the best result. This aligns with the goals set of for this specific research; to focus on explainability as well as performance. However, the features extracted from the application of MRELBP only focus on the spatial structure of the spectograms, these features are less interpretable and cannot be easily linked to psychological theory. Therefore, it would be a good choice to focus on explainable hand-crafted features and only incorporate deep-learning features if the performance is not high enough without them. With respect to the chosen model structure, CNNs are trained, as was the case for the text-based models, CNNs contain black-box model components, making it impossible for humans to follow the model's decision-making process. Again, choosing a simpler more interpretable model architecture would align better with the goals set for this specific research. The research by Chlasta et al. focuses on the tranformation of audio data to images in order to be able to efficiently make use of CNNS (which work best on images). Again, this approach yields impressive results. However, by changing the original data format, the interpretability of the model decreases vastly. One could argue that a spectogram is still interpretable to some degree but after the input layer, the features (represented by nodes in the neural network) become uninterpretable for humans. All in all, this approach is inventive and achieves a high accruacy. However, it does not hold up well in terms of interpretability and linkability to psychological theory.

Thirdly, the multi-modal model will be evaluated. The pre-processing of the data is not done manually, instead the authors propose an algorithm that identifies which fragments are important based on feature importance. Although this works in this specific case, for this research it would be better to identify the data (from all modalities) that revolves around depression related topics. Instead of doing this based on feature importance, it would be good to focus on certain topics that are brought up. Because the used dataset stems from a wizard of Oz set-up it is easier to identify topics of given answers based on the virtual interviewer's question. Secondly, the authors use multiple model architectures to be able to compare performance. This is also desirable for our research as it allows for comparison between model performance, interpretability and linkability to psychological theory. Furthermore, the authors propose a manner of feature extraction (Word2Vec) that has shown to be effective in other research as well. Therefore, it would be a good idea to try this technique as well and compare it to other text-based feature extraction techniques in order to determine which techniques yields the most optimal results.

Lastly, it should be mentioned that none of the studies incorporate breathing-related paralinguistic fea-

tures. Instead, they focus mostly on audio-related paralinguistic features For this research, it is feasible to research the possible predictive power of breathing-related features as there seemingly has been little research towards the predictive power of these type of features.

2.3.5 Early vs. Late Fusion

In the previous section multi-modal analysis has been discussed. In the literature, multi-modal analysis can be conducted through two different approaches, early fusion and late fusion. Multi-modal models by definition incorporate (at least) two sets of predictive features originating from different modalities. In the case of early fusion, features are extracted from the uni-modal data streams (e.g. video, audio) and grouped together. Next, a single predictive model is trained on the features from all different modalities. A schematic overview can be seen of an early fusion multi-modal model below in Figure 1.

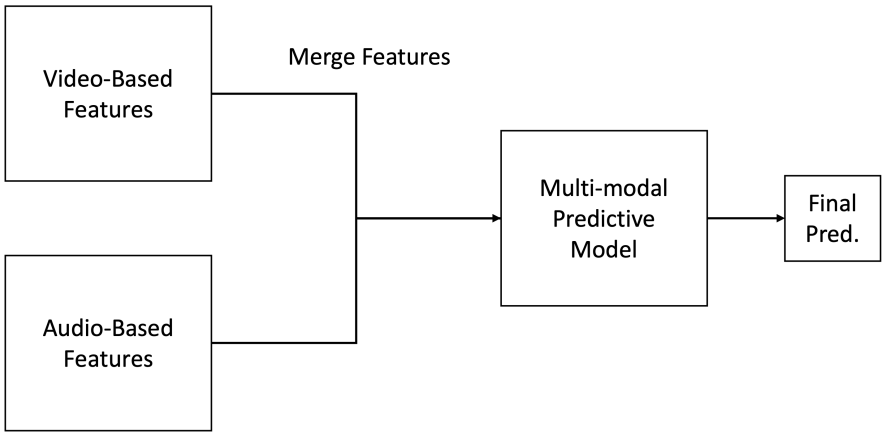

Figure 1: Early Fusion Schematic Overview

On the opposite, a late fusion approach consists of the extraction of different sets of features from uni-modal data streams. Subsequently, multiple predictive models are trained, each on a set of predictive features originating from a different modality. The outputs of these uni-modal models are then used as inputs for a final predictive model that outputs a final prediction. A schematic overview of a late fusion multi-modal model can be seen below in Figure 2.

Research by Snoek et al. [33] has investigated whether there is a substantial difference in performance between the two approaches to multi-modal analysis. In the study, the authors train both an early fusion model and a late fusion model to predict the topic (out of 20 possible categories) of a video from multi-modal data. The features used in the models originate from three different modalities: video, audio and text. The results suggest that there is a difference in performance between the model which was trained directly on all extracted features (early fusion) and the late fusion model. The late fusion model outperforms the early

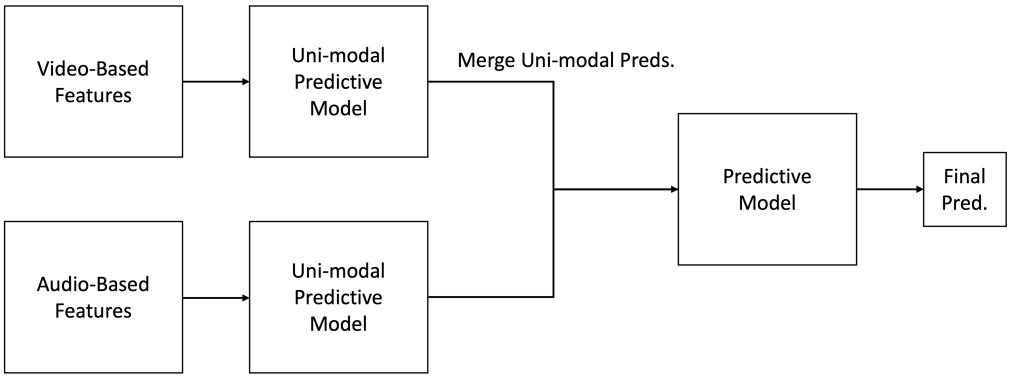

Figure 2: Late Fusion Schematic Overview

fusion model on the classification of 14 (out of 20) categories. Therefore, the authors conclude that late fusion tends to yield a higher performance than early fusion. However, it should be taken into account that a late fusion architecture is more computationally expensive as more models have to be trained.

2.4 Feature Extraction

In order to be able to train predictive models based on data from different modalities, features have to be extracted from the data streams. The raw data streams cannot be used directly as inputs. This process is called feature extraction. In this section an overview will be given of the different techniques used for feature extraction with regards to both text-based data and audio-based data. First, multiple techniques regarding the feature extraction of text data will be addressed. Secondly, feature extraction for audio data will be discussed.

2.4.1 Text Features

For the text-based feature sets, four different sets have been created: a bag-of-words representation of the interviews' transcripts and secondly, two dictionary-based feature-sets and lastly, a pre-trained sentence embedding algorithm was used to vectorize the content of the transcripts.

The first feature-set is a bag-of-words representation of the data. This means that all words in the dataset are tokenized and used as input features for a predictive model. The tokens present in the training set make up the vocabulary and determine the total number of input features (length of vocabulary). Next, for every word in the vocabulary the number of occurences of that word are determined for every part of the data set. For every feature (token) the numerical value used corresponds to the number of occurences of that word. It is worth mentioning that words that are not present in the vocabulary but are present in the development and test set are left out of the analysis as they cannot be counted. An advantage of this approach is that

it is computationally inexpensive, the resulting model is interpretable as the input features are existing words. However, the main disadvantage of this approach is that by using a bag-of-words representation, the temporality and relationships between words is lost. Only the occurences of every word are counted but their places in the text are not being accounted for.

The second text-based feature set that has been extracted from the transcripts consists of features that aim to represent the affective dimensions of a piece of text. These features have been extracted with the LIWC (Linguistic Inquiry and Word Count) software tool [34]. The tool works such that it takes in a piece of text and scores it on a maximum of 116 dimensions. The scoring occurs based on a matching algorithm between a pre-set dictionary and the words in the piece of text that is being analyzed. The dictionary consists of a set of words with scores on a multitude of dimensions. The final scores are then computed based on how often certain words (and thus the associated scores) occur in the piece of text.

Thirdly, the NRC-VAD Lexicon has been used to extract numerical information from the written transcripts of the data. This dictionary brings forward a list of 20.000 words that are all scored on three dimensions; valence, arousal and dominance. Every word in the list is scored on a scale between 0 (the lowest) and 1 (the highest). The NRC-VAD Lexicon can be used to compute an affective score for single words or sentences. The notion behind using these three dimensions as effective indicators of affect stems from previous research conducted by Russell which indicates that valence, arousal and dominance are to be considered as the most important dimensions of word meaning [35] [36].

The fourth and last text-based feature set that has been extracted from the audio data is a combination of a pre-trained BERT (Bidirectional Encoder Representations from Transfomers) model and the application of multiple functionals to summarize over entire interviews(as mentioned in the previous paragraph. The pre-trained model that was used is called S-BERT, more specifically; 'all-MiniLM-L6-v2'. This model maps sentences or pieces of text to a 384 dimensional vector space. This model effectively works the same as a traditional BERT model, but is adapted such that it can provide embeddings for entire sentences instead of just words. The (S-)BERT models have shown to reach State-of-the-Art performance [37]. However, it should be noted that the 384 S-BERT features extracted for each sentence purely indicate a point in the vector space and are unintelligible. Again, in order to ensure standardized lengths of feature vectors for every interview, multiple statistical functionals are applied to the original features. The entire set of used functionals will not be reported as a selection has been made on the most well performing functionals, these will be reported in the Results section together with the models' performance.

This section has elaborated upon the feature sets that have been extracted. However, the experimental setup in which these features have been used will be discussed in the fourth section ('Methodology')

2.4.2 Paralinguistic Features

In this section, the used feature extraction techniques for audio data will be discussed. Raw audio data cannot be used as input for machine learning models. Therefore, the features need to be extracted from the raw data. Different approaches are possible that focus on different properties of the audio data. Three feature extraction tools are discussed in depth. Firstly, the OpenSMILE software will be addressed. Secondly, MFCC (mel-frequency cepstral coefficients) will be discussed. Lastly, VGGish, an embedding model will be brought forward.

OpenSMILE is a software tool that allows users to extract features from speech audio files as well as music audio files. This research is solely concerned with the analysis of speech audio. Therefore, openSMILE's applications regarding the analysis of music audio will not be discussed. OpenSMILE's capacities can be split into two groups: the extraction of Low-Level Descriptors and secondly, the application of statistical functionals to these Low-Level Descriptors. Thus, for every Low-Level Descriptor, the openSMILE toolkit provides a set of statistical functionals that can be applied (to create more specific features) to the extracted Low-Level Descriptors. The amount of features that can be extracted through openSMILE has changed and increased over time due to the fact that it is an open-source project [38]. Still, the features can be grouped together based on what they measure. The OpenSMILE toolkit allows for the extraction of different featuresets, a rough split can be made between the ComParE2016 set (65 LLDs) and the GeMAPS set (18 LLDs). Most of the related research makes use of the OpenSMILE toolkit use the GeMAPS set. However, in order to ensure completeness this research has tested both feature sets (ComParE2016 and GeMAPS1b). The LLDs are extracted per set timestep, then summarized through functionals per audio fragment (in this case; answer to every question). Lastly, the entire conversation is summarized by taking the functionals over every audio fragment.

The second set of features that has been extracted are the MFCC (mel-frequency cepstral coefficients) features. These coefficients are computed by slicing an audio fragment into tiny frames, sliding a window across these frames, computing the discrete fourier transform for every frame, this results in a frequency spectrum. A frequency spectrum essentially gives an overview of all present frequencies in a frame. From this frequency spectrum, the power spectrum is computed. Next, the filterbank energies are computed. The next step in the process consists of taking the logarithm (base 10) of these computed filterbank energies. Lastly, in order to de-correlate the log filterbank energies (coefficients) discrete cosine transform is performed on all coefficients. The result is a set of coefficients for the entire original analyzed audio-fragment.

A visualization of this can be seen in Figure 3. As we are dealing with multiple audio fragments that need to be analyzed (due to the nature of the dataset), statistical functionals have been applied to summarize the

individual MFCCs over the entire conversation. The specifics of this will be discussed in the Methodology chapter.

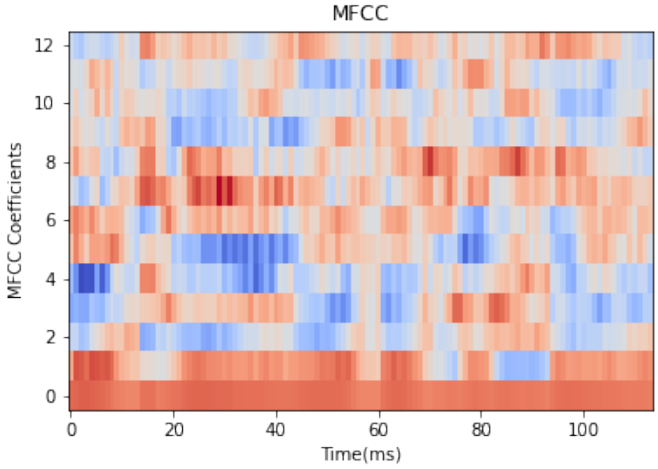

Figure 3: MFCC Visualization Example

The last feature vector on which predictive models have been trained are extracted through the VGGish model.The VGGish model is used for topic classification of audio files. The model architecture is a convolutional neural network (CNN) and strongly resembles the VGG model, which has shown to be very successful on the classification of images (and video) [39]. The VGGish model has been introduced during research towards the possible use-cases of CNNs towards the content classification of audio fragments. This research has shown that CNNs in general outperform deep neural networks (DNN) on the classification of audio fragments. VGGish specifically has been trained on the audio of 70 million Youtube videos. These videos (and their audio components) have automatically been assigned (multiple) labels. In total, there are 30.000 labels citehershey2017cnn. The VGGish model has been trained on an immense amount of data and is therefore mostly used as pre-trained model. It should be noted that the VGGish model takes log mel spectograms as inputs. Therefore, the audio should be pre-processed into that format in order for the model to work. The output of the model is a 128-dimensional vector, which is comparable to the word embeddings mentioned before. One of the main advantages of VGGish compared to other deep-learning models is that it produces a relatively compact feature vector, allowing for the usage of shallow learners for the final classification or regression task.

2.4.3 Breathing-related features

The last set of features that has been extracted are breathing-related. The used DAIC-WOZ dataset does not contain any ground-truth information with regards to the breathing patterns of the interviewed individuals.

Therefore, a breathing prediction was made based on the raw audio data. These predictions were made by using the predictive model brought forward by Markitantov et al. in their winning submission for the ComParE2020 challenge [40]. The proposed method consists of a 1D CNN (convolutional neural network) + LSTM (long-short term memory) model. For a more in-depth description of the parameters of the model, the original paper can be read. In the challenge, the model achieved a PCC (pearson correlation coefficient) of 76.3% with respect to the ground-truth breathing data. Therefore, the predictions coming forward from the usage of this model might provide us with breathing-related features that are close to the truth and possibly indicative of the presence of depression.

The sections above have elaborated upon the feature sets that have been extracted. However, the experimental setup in which these features have been used will be discussed more thoroughly in the fourth section ('Methodology').

2.5 Model Architectures

The extracted features will be used as inputs into three different model architectures; decision trees, random forest and linear regression. In this section, the mathematical intuition behind these architectures will be explained

2.5.1 Decision Trees

TODO

2.5.2 Random Forests

TODO

2.5.3 Linear Regression

TODO

2.6 Feature Importance

One of the crucial components of machine learning is the ability to interpret a predictive model and derive how the individual input features affect the model's prediction. SHAP is a python library created by Lundberg & Lee that enables the user to identify feature importance [41]. The SHAP library provides multiple functions that enable the interpretation of both linear and non-linear models. The importance of every feature is based on a 'Shapley' value. This value represents the effect that a feature has on the prediction of the model when

included as input feature. For linear models the 'Shapley' value of a feature is computed by training a model in which the feature is present and by training a model in which the feature is left out. This is done for every possible subset of features in the model. Next, the average of the differences between the models in which the feature is present and the model in which it is left out is then calculated and used as Shapley value.

3 Data

The depression prediction models used in this research will be trained on the DAIC-WOZ dataset. The dataset is used in the Audio/Visual Emotion Challenge (AVEC) which is a yearly competition that aims to stimulate the research towards the automation of emotion analysis based on multi-modal input data. The goal for the competitors is to create the best-performing model (on the test set of the DAIC-WOZ dataset).

This section will provide an explanation concerning the origin of the dataset as well as an overview of previous research done with the DAIC-WOZ dataset.

3.1 The DAIC-WOZ dataset

The DAIC-WOZ dataset originates from another larger dataset, the Distress Analysis Interview Corpus (DAIC). The DAIC dataset is made up of a variety of clinical interviews conducted by virtual interviewers. The DAIC dataset has been created to facilitate research towards the diagnosis and classification of psychological distress disorders [42]. The data present in the DAIC dataset originates from a set of interviews conducted by a virtual agent. The motivation to use a virtual agent to conduct these interviews has been to research possible future applications of virtual agents as identifiers of mental disorders [43].

The DAIC-WOZ dataset is the subset of the DAIC dataset which contains all Wizard of Oz interviews. These interviews have been recorded and have been conducted by a virtual agent (named 'Ellie'). The virtual agent interacts with the participant over the course of the interview but does not do this autonomously, instead, it is controlled by a researcher controlling the virtual agent's communication (not present in the room where the interview takes place).

The experimental setup used to gather data during the interviews can be seen below in Figure 1. Participants are placed in a room with the virtual interviewer. The interview is led by the virtual agent, controlled by a researcher in a different room and the participant responds to questions asked by the interviewer. During the entire interview, the participant is recorded on video and the interview's audio is recorded through a microphone. After the interview, both recordings are logged to a server. Subsequently, the audio recordings are manually annotated. Resulting in transcripts of the interviews. Therefore, each interview can be analyzed based on three different modalities; video, audio and text. It should be noted that the data has been anonymized in order to respect the participants' privacy. Therefore, the video data cannot be accessed directly. Instead, software has been used to extract features from the raw data and these features can be accessed. This has been done to prevent the participants from being recognized based on raw video recordings. The raw audio files are available for analysis. Additionally, the dataset contains audio features which have been extracted from the raw audio files using the COVAREP software [44].

The dataset is labeled in two ways. The first one is a binary classification between depressed and non-depressed. The second classification is on discrete values between 0 and 24. This classification represents the degree of depression. It should be noted that the degree of depression is based on the PHQ-8 score of the participant. The PHQ-8 is an instrument used to enable individuals to self-report severity of depression. The instrument resembles the PHQ-9 questionnaire which has been discussed in a previous section. The difference between the two is that the PHQ-8 as opposed to the PHQ-9 does not contain the question concerned with self-harm and death [45]. The binary classification uses the PHQ-8 score and transforms it into a binary value by classifying every PHQ-8 score below 10 as non-depressed and every score equal to and above 10 as depressed.

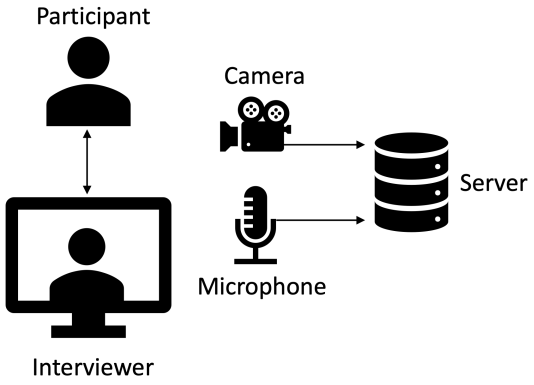

Figure 4: DAIC-WOZ Interviews Experimental Setup

The DAIC-WOZ dataset consists of 189 interviews, this dataset has mainly been created for machine learning applications. Therefore, the dataset is already split into a training, validation and test set. The training set consists of 107 interviews, the validation set consists of 35 interviews and the test set consists of 47 interviews. In total, the dataset contains 54 interviews with depressed people and 132 interviews with non-depressed people. This class imbalance negatively affects the trained models' performances on the infrequent class. In this case, the underrepresented class is depression. However, the predictive models should perform well on the detection of depression. Therefore, depending on the model's performance on the depression class it might be required to re-balance the dataset in order to improve performance. Furthermore, research by Bailey & Plumbley [46] has found that the DAIC-WOZ dataset contains gender bias. The bias comes from the fact that the ratio between depressed and non-depressed people is higher for females than for males. The authors report that this gender bias may cause inaccurate representations of the performance of machine learning models that have been trained on the DAIC-WOZ dataset. To solve this, the authors suggest the implementation of gender 'balancing' techniques. These techniques aim to level the ratios between depressed and non-depressed people for both genders.

3.2 AVEC2017 Challenge

As has been stated earlier in this section, the DAIC-WOZ dataset has been used by researchers in the AVEC2017 challenge to develop emotion predictive models [47]. This section will discuss the most relevant submissions that focus on the detection of depression.

Firstly, Le Yang et al. [48] propose a multi-modal deep learning model to predict depression. The authors of the paper start by selecting the features from each modality that will be inputted into the model. For the text-based features, a selection has been made to include sentences that contain information related to depression symptoms. In addition, these sentences are vectorized through the use of the Paragraph Vector framework. The Paragraph Vector model can be compared to the Word2Vec model, which quantifies the representation of words to vectors in a vector space [49]. The Paragraph Vector framework does the same but for entire sentences or paragraphs. The second modality that is used as input into the model is video data. For this modality, the authors directly use the features provided in the DAIC-WOZ dataset; the 2D landmarks. Secondly, the authors extract a second set of video-based features by measuring the movement or displacement of every landmark over time. This is referred to as Histogram of Displacement Range (HDR). The HDR allows for an in-depth analysis of the movements of the facial landmarks in both the vertical and horizontal directions over time. Lastly, the authors include audio-based features. These features are extracted using the openSMILE software [38]. Each interview's audio file has been segmented into chunks of 60ms in length with a shift of 10 ms per chunk. Subsequently, each chunk is analyzed with the openSMILE software, which extracts 238 descriptors from every chunk. This is followed by an application of multiple statistical functionals to every descriptor. In total, this results in a total of 6902 predictive features for every 60 ms audio chunk.

After having extracted the features for each modality, the authors proceed to train three deep learning models, one for every modality. The model architectures chosen for the uni-modal trained models are deep convolutional neural networks. The three models are trained to output and thus predict the PHQ-8 score. The predicted scores from all three models are then used as inputs into a new deep neural network. This model combines the three predictions of the first three models and outputs a final prediction for the PHQ-8 score.

The results suggest that the model combining the three modalities outperforms the uni-modal models. Furthermore, the model outperforms the baseline models as provided in the AVEC2017 challenge on both the validation and the test set [50].

A different approach has been chosen by Gong & Poellabauer [51] , instead of using deep learning they focus on topic modeling and the use of a simpler model architecture. The hypothesis that the authors use as

the foundation for their research is that when pre-selecting pieces of data that contain information regarding depression, the model's performance will be better than when trained on the entire data corpus (which also includes a lot of noise or irrelevant data).

The interviews making up the DAIC-WOZ dataset are made up of interviews between a patient and a virtual agent, where the virtual agent is controlled by a human in another room. Because of this human-controlled setup, the process of topic modeling becomes relatively trivial as the questions that are being asked all originate from a limited, fixed batch of questions. The authors analyze the virtual agent's entire sentence corpus, consequently getting rid off all statements that are not questions and do not start the discussion of a new topic. Next, the rest of the questions are grouped together based on the topic they address. This results in 83 topics that can be addressed during the clinical interviews. For every topic, three sets of features are extracted from the DAIC-WOZ dataset. The audio features consist of the features that are extracted from the raw audio files with the COVAREP software (in chunks of 10 ms). For the video features a subset of the provided features in the original dataset are used, namely, the action unit related features. Lastly, for the text-based features, the LIWC framework is used to categorize the words used (in the answers to the questions per topic) in 93 emotion categories. These features are taken together and form the initial set of predictive features. However, in order to prevent overfitting and reduce computational costs, the authors perform two steps of feature selection. Firstly, the predictive power of every feature is computed by evaluating the correlation between the feature and the target variable. The features that have a high correlation with the target variable are selected and considered as important. Additionally, features that yield a high correlation with respect to other features are left out of the final selection of features. The second step that is undertaken by the authors is that every feature selected after the first step is evaluated based on the F-value between the feature and the target variable. Subsequently, the final selection of features is a selection of the features that yield the highest F-value.

The model architecture that is used to train the final model one is a stochastic gradient descent regression. In order to be able to contextualize the final model's performance, three baseline models are used. When comparing the final model's performance to three baseline models, the authors find that the final model outperforms all three baseline models. Therefore, the authors conclude that the implementation of topic modeling results in a better performing depression prediction model. Furthermore, when looking at the final set of features used in the predictive model it becomes clear that some of the features are not theoretically related to the detection of depression. The authors suggest that this indicates that there are more indicators of depression than a clinician might be able to observe in an interview. However, the majority of selected features revolve around topics that are theoretically closely related to symptoms of depression.

Research by Syed et al. [52] proposes another solution to the AVEC2017 challenge. The authors present

a model which focuses on bio-markers of psycho motor retardation to predict severity of depression. The motivation for this approach originates from research in the field of psychology. The authors refer to studies indicating a difference in motor activity and movement between depressed and non-depressed people. According to the authors of the paper this notion of there being a difference in psycho motor activity between depressed and non-depressed people suggests that if this difference can be quantified, it can be used in building models that are able to automatically classify severity of depression.

The hypothesis for this research is that psycho motor retardation affects an individual's speech and facial expressions. Therefore, the authors start by extracting features from the different modalities. Firstly, the audio-based features are constructed by taking the parts of the interview in which the participant is speaking and consequently the audio data is split into chunks and the turbulence for each chunk of audio is computed. In this case, the turbulence of the participant's speech is defined as the highest pitch present in a specific chunk of audio divided by the root mean square of all pitches in the specific audio chunk. Additionally, the audio features given in the original DAIC-WOZ dataset are also being used as predictive features. Secondly, the video-based features that are used in the model are extracted from the 3D features initially present in the dataset. From these 3D features, the movement of the mouth, the head and the eyelids are extracted. For all three movements the speed and acceleration are extracted in both the vertical and horizontal directions. Thirdly, the speech spectra of the participants are quantified through Fisher Vector encoding to attain even more audio-based features. This is done by creating a baseline model of all (given) spectral features. The baseline model assumes that the values of all spectral features are randomly chosen from a Gaussian (normal) distribution. Consequently, the participants' data is normalized and transformed such that the number of spectral features in the baseline model and the participants' data are the same. For every participant the spectral feature values are compared to the baseline model's values and the deviation is computed, completing the Fisher Vector encoding.

In order to determine feature importance for each modality, three models are fitted, each containing the features from a different modality. Consequently, the correlation between the features and the target variable (PHQ-8 score) are computed. The features that yield the highest correlation are then determined to be the best predictors of depression. Additionally, another predictive model (a support vector regression) is fitted. The features used in the model are the Fisher Vectors which have shown to yield the highest predictive power. After having fitted the model, its performance is computed on the test set and compared to the baseline model provided by the AVEC2017 challenge. The results show that the support vector regression outperforms the baseline model. Lastly, a second model is fitted which uses pitch-related features on a partial least squares regression, this model slightly outperforms the support vector regression model.

The last submission of the AVEC2017 that will be addressed is a paper by Sun et al. [53]. The paper

proposes an approach in which the transcripts of the interviews are analyzed and features are extracted based on the topics that are addressed in certain parts of the interviews. These extracted features are subsequently used as inputs for a random forest model that aims to predict PHQ-8 scores.

First, the authors describe the feature selection and extraction process that has been done. The audio features that have been used are the features that are provided in the dataset. No further pre-processing of these features is done. The video-based features are created by measuring the speed and acceleration of the (eye) gaze and the position of the head. Furthermore, principal component analysis is done to reduce the number of landmark features. However, most of the feature selection occurs on the transcripts (text-based modality) of the interviews. The transcripts of the interviews are analyzed such that only the responses of the participant are being considered as features and the virtual interviewer's questions are omitted. Next, the text-based features are selected based on parts of the transcript that revolve around topics that are indicative of the presence of depression. The topics that are selected as features in the model are chosen based on a search for the optimal combination of features. This is done by randomly leaving out features and consequently evaluating the performance attained with the selected set of features. This process is repeated until the optimal combination of features is found. The final selection of topics considered for text-based features are successive treatment, depression diagnostic, personal preference, feeling and sleep quality.

The feature selection procedure is followed by a comparison between a multitude of models that are trained on different sets of features. First, three models are trained on the features of the different modalities. Next, two models are created that fuse the data from two different modalities. A video-text fused model is trained and an audio-text fused model is trained. Both these fused models outperform the single-modality baseline models on the validation set.

Lastly, a model is trained on data from all three modalities. This model predicts PHQ-8 scores based on inputs that originate from all three modalities present in the dataset (video, audio, text). The model is designed such that first the video and text streams are fused. Similarly, the audio and text streams are fused. Lastly, the video-text model, the audio-text model and the text stream are merged, resulting in the multi-modal fused model which predicts PHQ-8 scores.

The performance of the trained models is evaluated in terms of root mean square error (RMSE) and mean absolute error (MAE). The authors specifically evaluate the performance of the multi-modal random forest fused model and the selected-text model. These two models are evaluated against each other and against the baseline models. From this comparison it becomes clear that the model trained on selected text-based features outperforms the other trained models and the provided baseline models (with a RMSE of 4.98 and a MAE of 3.87) on the test set.

3.3 Evaluation of AVEC2017 Models

This section will evaluate the chosen approaches in the previous research on the DAIC-WOZ dataset. Le Yang et al. approach the problem by performing topic modeling on the text-based data. This is desirable as it effectively reduces the noise in the data that the model is trained on. However, the authors do not do this for the data from different modalities. In our research it would be better to identify the pieces of text and timestamps that address depression related topics and use this topic modeled data as training data for all three modalities. With respect to the architecture of the final model, the authors train three uni-modal models and merge their predictions into a final model. This is known in the literature as 'late fusion'. As has been stated in a previous section, late fusion has shown to outperform early fusion and also allows for a more granular analysis of the most indicative modalities (and features). Using a late fusion approach is the most suitable for our research. The authors use mostly deep learning architectures for the model. This achieves high results but difficults the process of identifying which features yield high predictive power. Therefore, it might be a better idea to use more shallow models. Furthermore, the authors perform data balancing in order to attain a more even ratio between depressed and non-depressed people. Due to the fact that our research will make use of the same dataset, it would be good to consider some form of re-sampling in order to balance the ratios.

Secondly, Gong & Poellabauer also perform topic modeling. The authors identify 83 topics and select a fixed set of features to be extracted for all topics. These features are then concatenated into a single vector and inputted into a shallow learning model. The notion of using topic modeling on all modalities is good as it reduces the noise for all three modalities instead of focusing solely on text. However, this early fusion makes it harder to identify the individual role that the modalities and features play in the final prediction. Therefore, it would be more suitable to train uni-modal models and merge their predictions in a final model. Furthermore, the authors perform a grid-search between different model architectures to find the best-performing combination of features and model architecture. Using multiple shallow models is also desirable for our research as it allows for cross-model analysis of feature importance.

Thirdly, Syed et al. focus mostly on the use of video-based data to predict features. This research instead focuses more on the role that paralinguistic features play in the detection of depression. Therefore, although the authors propose a novel way of extracting features from video data, their practices are not usable for this research.

Lastly, Sun et al. use a late-fusion random forest model to incorporate video, text and audio-based features. Furthermore, the authors train uni-modal random forest to provide baselines and comparison material. As stated above, the final model also relies on video-based data. However, this research does not

incorporate video-based features. Therefore, the methods chosen for video-based feature extraction are not relevant in this case. Opposed to the other discussed research papers, this research evaluates the models on both binary classification (depressed and non-depressed) and regression classification. This is insightful as it provides more information regarding the relative performance of each model (on slightly different tasks) and the robustness regarding the models' ability to identify depression (even when measured in accordance with different frameworks).

4 Scope of Research

RQ: What is the role of the audio modality in the automated depression of detection in the AVEC2017 dataset?

RQ.1: What is the role of paralinguistic features in the automated detection of depression?

RQ.2 How do breathing-related features contribute to the automated detection of depression?

RQ.3 What is the relationship between breathing and paralinguistic features in the AVEC2017 dataset?

5 Methodology

In this chapter, an explanation will be given regarding the trained models and the features used in those models. First, the different model architectures are discussed, followed by an explanation concerning the models and tools used to extract features from the raw data. Lastly, the methods used for final feature selection will be addressed.

5.1 Model Architectures

This thesis focuses on the role that paralinguistic features and breathing-related features have in the automated detection of depression. Therefore, the model architectures of the models that have been trained for this research all fall under the class of 'shallow' learners. This allows us to be able to interpret the output of the models and the relation between specific (input) features and the target variable (degree of depression). For each set of features, three model architectures have been tested: a decision tree, a random forest and a linear regression.

With regards to the decision tree, the parameters that have been tuned are 'max-depth', 'min-samples-leaf' and 'min-samples-split'. For the random forest models, the hyper parameters that have been tuned consist of 'max-depth', 'min-samples-leaf', 'min-samples-split' and 'n-estimators'. The 'max-depth' parameter refers to the maximum number of depth (in terms of layers) that the model can have, the 'min-samples-leaf' refers to the minimal number of observations that is needed for a split, the 'min-samples-split' parameter dictates how many observations a leaf node should at least contain and lastly, the 'n-estimators' parameter refers to how many decision trees are used to form the random forest model. It should be noted that for this model architecture, these hyperparameters are mostly used to prevent overfitting of the model on the training data. Lastly, for the linear regression, only the combination of input features has been optimized as this architecture does not yield any hyperparameters that can be optimized.

5.2 Feature Sets

In total, models have been trained on feature sets originating from two modalities: text modality and audio modality. The feature sets that are extracted from the raw audio data can again be split into two categories; paralinguistic features and breathing-related features. The focus of this research has been on features extracted from audio. First, the approach used for the text-based models, which will serve as additional baseline models will be discussed. Secondly, the methodology regarding the audio-based models will be discussed.

5.2.1 Text-Based Models

For the text-based models, features are extracted from the interview transcripts and subsequently used as inputs for a regression model. The pipeline that can be seen in the figure below illustrates the process from transcript to depression prediction (PHQ8-Score).

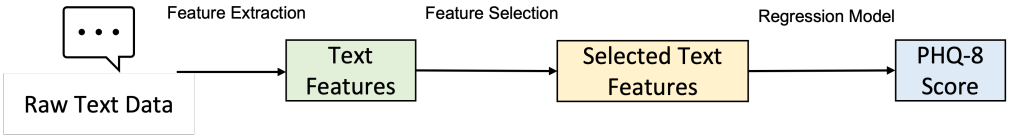

Figure 5: Pipeline for Text-Based Models

Three different regression models have been built. Each model was built using a different set of input features. The first set of features used a B-O-W representation of the transcripts as input features.

For the bag-of-words representation all answers given by the patients in the interviews making up the training set have been tokenized into single words. Next, in order to reduce the number of different words, all tokens have been stemmed. The process of stemming a word (or token) consists of reducing a word to its stem. In machine learning applications this is often done to reduce the dimensionality of the data. Next every unique token extracted from the training set is used as a feature, this is the final set of features and often referred to as the 'vocabulary'. The value assigned to each feature is based on the number of times that token (word) is used by the patient in the entire interview. Next, the patients' answers from the interviews in the development and test set are also tokenized and stemmed. Lastly, for each interview in the development and test set, the number of occurrences of every feature are counted and used as values for those features. This completes the process of feature extraction.

The second feature set that has been used to train predictive models consists of emotion-related features extracted through the LIWC dictionary. Again, for every answer provided by the patient, the LIWC dictionary is applied, this dictionary extracts scores on a multitude of emotional dimensions. In this specific case, 110 features have been extracted from every answer given by the patient. Only the features relating to punctuation have been left out as those are not relevant for this research. Due to the fact that the interviews are not of standardized length and vary in terms of questions asked by the AI-agent, a method is needed that allows for a summarization over the entire conversation as well as a standardization of number of input features. This is required as the chosen model architectures only work if different data points consist of the same dimensions. To solve this, statistical functionals are applied to the set of features (110 features per answer). This causes there to be a fixed number of features per interview regardless of length. The application of statistical functionals has also been done on other feature sets, those will be mentioned separately.

The third and last set of features that has been extracted from the text consists of sentence embeddings extracted with a pre-trained S-BERT model. This model extracts a feature of length 128 for every sentence. Thus, in this specific case, for every answer provided by the patient, a sentence embedding is extracted. Next, in order to create a feature vector of fixed length for every interview, multiple statistical functionals are applied to summarize the entire interview. This is required because the number of questions asked varies per interview, therefore, if all embeddings would be concatenated this would result in feature vectors of different lengths. This does not work for training a shallow regression model.

5.2.2 Audio-Based Models

For the audio modality, different sets of features are extracted. However, the prediction pipeline has been the same, in Figure an abstract visualization of the pipeline can be seen.

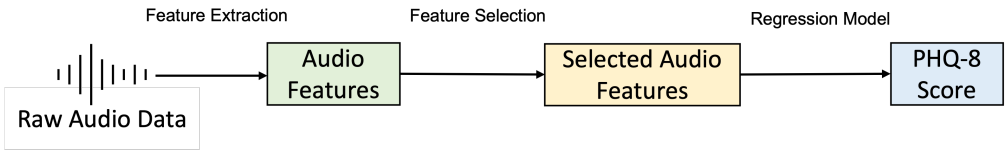

Figure 6: Pipeline for Audio-Based Models

As can be seen, the pipeline starts by taking in the raw audio data (.wav format). Next, a tool or pre-trained model is used to extract a set of audio-based features. After this, another step of pre-processing is done through feature selection. Lastly, a regressor model is fitted or trained to make a final prediction with regards to depression severity.

In the step from raw audio data to audio features, variation occurs based on the featureset that is being extracted. It should be noted that only the patients' audio is used to extract these features. The method employed for this is to first slice every audio file into different groups: answers and questions. As this research is concerned with the detection of depression (among patients) the data that is used to extract audio features is the slices of the audio that correspond to the answers given by the patients. Additionally, it should be noted that features should be extracted per answer and then summarized. The answer audio slices cannot be concatenated to directly extract a single set of features for the whole interview.

First, the OpenSMILE toolkit has been used to extract two featuresets. As has been stated before, OpenSMILE extracts a set of LLDs for every frame in an audio fragment. Both the ComParE2016 LLD set and the GeMAPS1b LLD set have been extracted. These features are extracted per frame on a sliding window. It should be noted that the dataset consists of conversations between an AI agent and a person. These conversations vary in length and in questions asked by the AI agent. Therefore, it is not possible to append all LLDs and use those as inputs for the regressor model. This would lead to there being variation

in the number of input features. To solve this problem statistical functionals are applied to summarize the features extracted per answer given by the individual. This process is then repeated to summarize the entire conversation. Multiple statistical functionals have been tested and evaluated in order to optimize the set of extracted features. Examples of such functionals are: standard deviation, mean, median, min, interquartile range, etc.

The second set of features that has been extracted are the mel-frequency cepstral coefficients. As was the case for the OpenSMILE features, the MFCC features are computed per frame and on a sliding window. In order to be able to extract features the raw audio data has to be quantified. To solve this and go from raw audio to vector, the librosa library has been used, which extracts the signaling rate as well as the vector representation of the signal [54]. The signaling rate for the audio in the DAIC-WOZ dataset is 16.000. Next, the extraction of the MFCC features was done by using the 'Python Speech Features' library [55]. The default number of features extracted per frame is 13, this default setting was adopted for this part of the analysis. In order to create input feature vectors of a consistent length, statistical functionals are used to summarize and standardize the number of features per interview.

Another set of feature vectors on which predictive models have been trained are extracted through the VGGish model. The VGGish model extracts a feature of length 128 for every frame. In order to summarize this embedding for per answer, statistical functionals are applied. Next, to summarize over the entire interview, statistical functionals are applied again. This results in feature vectors of fixed length independent of the length of the interview from which the features are extracted.

Lastly, a set of breathing-related features have been extracted from the raw audio files.The DAIC-WOZ dataset does not contain any ground truth data for the patients' breathing patterns. Therefore, a pre-trained model has been employed which predicts the breathing patterns based on the raw audio files. This pre-trained model is a deep learning model brought forward by Markitantov et al. and consists of a 1D CNN + LSTM model architecture [40]. The original model was trained on a different dataset and thus optimized for that dataset as well. It should also be noted that the model brought forard by Markitantov et al. was trained on the data provided in the INTERSPEECH202 breathing sub-challenge [56]. The data provided for the sub-challenge only contains the ground-truth for the upper respiratory belt. Therefore, Marktitantov et al.'s model also only predicts upper-belt breathing patterns. The optimal window size of the model on the original dataset is 16 seconds. However, previous research by Pessanha et al. has shown that on the DAIC-WOZ dataset using non-normalized audio data and a window size of 6 seconds yield satisfactory results [57]. Therefore, for this research the same settings have been used. This also allows for a comparison between results. It should be noted that the model has thus only been used on the audio fragments that are 6 seconds (or longer) stemming from the answers that the patients have provided on the AI agent's questions. The AI

agent's audio has not been used as the focus lies on predicting depression among the patients. In figure 6 below an example can be seen of a predicted breathing signal (of the upper respiratory belt).

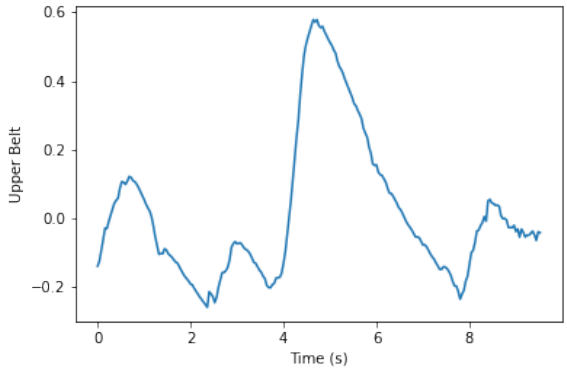

Figure 7: Example of Predicted Breathing Signal

As stated above, the pre-trained model predicts breathing signals per question. However, in order to be able to use the breathing signal in predictive models, statistical properties of the signal have to be extracted and subsequently be summarized to attain a feature vector of consistent length. To do this, two approaches have been tested.

First an approach has been tested focusing on a set of features that can be segmented into three groups: the duration of peaks, zero-crossings and statistical properties of the signal. The duration of peaks measures how long (in s) the peaks are in the signal, for this property, the max, min and mean are calculated. Next, zero-crossing refers to the volatility of the data; how often the signal alters between positive and negative (and crosses the 'zero') for this group, the max, min and mean are calculated. Lastly, to summarize every individual breathing signal the following statistical functionals have been computed: mean, min, max, standard deviation.

The second approach that has been tested focuses on the extraction of a different set of features. These features can, like for the first approach, be grouped. The first group consists of features that relate to peak-to-peak distance, this refers to the distance between peaks in the breathing signal. The second group consists of features that relate to how much is being inhaled (in volume), this refers to how deep the patient is breathing. The last group contains features that describe how steep the slope is of every peak, in practice this refers to the abruptness of an inhale.

Again, to be able to summarize over entire interviews, for every feature mentioned above, multiple statistical functionals have been computed to attain a consistent input feature vector length for every interview.

5.3 Feature Selection

After having extracted the different feature sets, the last pre-processing step before training the final model is the process of selecting the features that will serve as inputs for the regression models. This is especially in scenarios when initially there are a lot of features (high dimensionality of the data). The approach chosen for this research consists of two components. The first component is independent of model architecture. For every feature in the set the PCC is calculated with respect to the target variable, in this case; the PHQ-8 score. Next, for the second component, from the features that are most correlated (highest PCC) with the target variable the performance of different combinations of features are evaluated. The combination that scores highest is then used on the test set to evaluate the overall performance of a certain feature set / approach.

5.4 Correlation Analysis

Part of the goal of this research is to determine the relationship between paralinguistic features and breathing features. Therefore, a correlation analysis has been done between the features that make up the paralinguistic feature sets and the breathing-related feature sets. The methodology chosen for this has been to take all features present in two sets and compute the PCC for every combination. Due to the fact that the used feature sets contain a large number of features only the combinations with the highest PCC are discussed.

5.5 Answer Clustering

The audio-based analysis can be improved upon by incorporating the text-based modality indirectly. Instead of directly extracting text-based features and using those as inputs for the model, the transcripts can be used to segment the interview into different contexts. The interviews are segmented based on emotions present in each answer. For every utterance a score is computed on three dimensions: arousal, dominance and valence. The scores for every dimension are computed through the use of the NRC VAD Lexicon [58]. This lexicon contains over 20.000 words and has set scores for every dimension for every word. In order to compute the score per question/answer the score for every word is summed and then divided by the number of words in order to attain a 'mean' score per sentence. It should be noted that the interviews contain words that do not appear in the NRC VAD Lexicon, those words can thus not be scored and are left out. The scores are solely based on words present in the NRC VAD Lexicon.

For every individual emotional dimension (arousal/dominance/valence) different cutoff values have been tested to segment the answers provided by the patients into segments with different emotional load. The segmentation was done for all three dimensions but in different experiments to research the individual in-

fluence of each of the three mentioned emotional dimensions in combination with breathing features on the prediction of depression.

For the first experiment all patients' answers that have a score between 0-1 are put together. For the second experiment, the answers are split into instances that scored below 0.5 and instances that scored above 0.5. The third experiment did the same as the second experiment but instead split the answers into three categories (0-0.33, 0.33-0.66 and 0.66-1). The fourth and last experiment used cutoff values that split the answers into four equal segments (0-0.25, 0.25-0.5, 0.5-0.75, 0.75-1). No further splits have been used as preliminary tests showed that this drastically reduced the number of datapoints per segment. This segmentation process was performed to determine if training predictive models on a subset of the data with a specific emotional load could improve overall model performance. For every experiment both a linear regression and a random forest model were trained on the different emotional subsets of the data, the results of these experiments are reported in the following chapter.

5.6 Feature Importance

6 Results

In the previous section an overview has been given of the methodology used and the experimental setups. This section will provide the results of the conducted experiments. As has been stated in the previous section, all feature sets models have been trained on a multitude of hyperparameter combinations as well as different subsets of the original feature set. However, in this section the focus will lie on the best performing models and the most relevant results. This chapter is further divided into five subsections: First, the baseline models provided by the AVEC2017 challenge will be discussed. Next, a second set of baseline models will be brought forward in the form of purely text-based shallow models. Thirdly, the audio-based models will be elaborated upon. This will be followed by an explanation of the results of the answer-segmentation approach. Lastly, an overview will be given of all models and the performance that they yield.

6.1 Baseline Models

Before discussing the results of the experiments conducted for this specific research, an overview will be given of the baseline models provided in the AVEC2017 challenge. In addition to the baseline models provided in the original challenge, some of the best performing submissions to the AVEC2017 challenge are also included in the table below. This allows for a comparative analysis between the presented models and the baseline models brought forward in the literature. It should be noted that the most relevant result from the baseline models is the performance on the audio modality, as that modality has been the focus of this specific research as well.

Table 1: Overview of performance of baseline models provided by the AVEC2017 challenge and submissions

	Modality	**Model Architecture**	**Performance (RMSE)**	
			Dev	*Test*
Challenge Baseline	Audio	RF	6.74	7.78
	Video	RF	7.13	**6.97**
	Audio + Video	RF	**6.62**	7.05
Gong et al. [51]	Audio + Video + Text	SGD LR	3.54	4.99
Sun et al. [53]	Selected Text	RF	4.97	**4.98**
Yang et al. [59]	Audio + Video + Text	DL	**3.09**	5.40
Yang et al. [50]	Audio + Video + Text	DL	4.65	5.97

6.2 Text-based Models

The first set of models that have been trained are text-based. These models will serve as a second baseline to see whether audio-based models can outperform text-based models as the original provided baseline does not contain a text-based model that is trained on all data. It should be noted that Sun et al.'s submission in the AVEC2017 challenge did contain a text-based approach. However, this approach employed a complex selected-text approach. Sun et al. split the interviews in sections that discuss different topics. These topics are PTSD, successive treatment, feeling, personal preference, introversion and sleep quality. Next, text-based features are extracted for every topic and experiments are conducted to determine which combination of topics can be used as input to build a predictive model that yields the best performance. This results in Sun et al. finding an optimal set of topics on which a model is trained which then substantially outperforms the baseline models provided in the challenge (RMSE of 4.97 on the development set and 4.98 on the test set) [53].

The approach brought forward by Sun et al. reaches an impressive performance. However, the chosen approach is complex and requires a lot of steps. Additionally, the aim of this thesis is not to find the best text-based model, instead it focuses on the use of paralinguistic features for the prediction of depression. Therefore, the text-based models brought forward in this section are all trained on features extracted from the entire dataset and only make use of shallow model architectures. The aim of this section is thus to provide a text-based baseline to allow for comparison between text-based models and paralinguistic models on the DAIC-WOZ dataset and not to beat the existing best-performing text-based models.

Three different feature sets have been experimented with: B-O-W features, LIWC-based features and S-BERT. A rough split can be made between feature sets that contain intelligible features: the B-O-W feature set and the LIWC feature set and the feature set that consists of unintelligible features. The performance of all three models can be seen in the table below. The results in the table show that the S-BERT model performs the best in terms of RMSE. The second best performing featureset is LIWC and the third featureset is B-O-W. This ranking of performance is in line with what was expected. Due to the fact that the model making use of S-BERT relies on sentence embeddings, it can take into account temporality and contextuality of words in a sentence opposed to both LIWC and B-O-W. This has to do with the fact that both intelligible featuresets purely rely on counting words present in a sentence and basing feature values on those counts, they do not take into account the order of words. This prevents these feature extraction techniques from detecting sentiment in a sentence if a negation is present. However, the fact that S-BERT produces unintelligible features should be considered as a downside of using that technique as it prevents human interpretation on how a certain prediction was reached. This is not directly reflected in the performance as presented in the

table but should be considered in the qualitative comparison between the models.

Feature Set	**Model**	**Performance (RMSE)**	
		Dev	*Test*
B-O-W	LR	5.98	6.40
	RF	6.35	6.52
LIWC	LR	6.06	5.95
	RF	5.94	6.24
S-BERT	LR	**5.50**	**5.75**
	RF	6.24	6.15

Table 2: Performance of text-based models on dev and test set

6.3 Audio-based Models

This section will focus on the experiments that have been done with audio-based predictive models. The experiments are split into two groups, first the results of the experiments revolving around paralinguistic models will be discussed, the experiments done with breathing-based models will be elaborated upon.

6.3.1 Paralinguistic Models

Three different types of features have been used to train paralinguistic-based models: OpenSMILE (ComParE2016 and GeMaps1b sets), MFCC and VGGish. A split can be made between intelligible features and unintelligible features. The OpenSMILE and the MFCC features are intelligible, the VGGish features are unintelligible. The VGGish features are unintelligible because for every audio chunk a feature vector of length 128 is extracted which maps the audio to a 128 dimensional space with no defined dimensions.

For every featureset the initial conducted experiment was to evaluate performance if all features were used and a number of functionals were used to summarize the features over the entire conversation. Examples of used functionals are: 'mean', 'median', 'min' 'max' 'std dev.'. The next step was to test different subsets of every featureset to see if performance could be improved, this optimization was done on the development set and aimed to minimize the RMSE. As an exhaustive search for the optimal featureset is computationally expensive, a heuristic was used to narrow down the combinations to be tested. This heuristic was based on every input feature's PCC to the target variable ('PHQ8-Score'). Then combinations were tested containing the features that have the highest PCC with respect to the target variable. As can be seen in the table below, after feature selection, the OpenSMILE featuresets brought forward the model that performed best on the development set (RMSE of 6.16) and the model that performed best on the test set (RMSE of 6.25).

Table 3: Performance of different acoustic feature-sets for both linear regression and random forest model architecture on the development and test set

Feature Set		**Model Architecture**	**Performance (RMSE)**	
			Dev	*Test*
OpenSMILE	ComParE2016	LR	6.52	7.29
		RF	6.71	**6.25**
	GeMaps1b	LR	**6.16**	7.40
		RF	6.40	6.47
MFCC		LR	6.37	6.92
		RF	6.48	6.34
VGGish		LR	6.67	6.41
		RF	6.92	6.67

For the random forest model, hyperparameter tuning has been performed to optimize the model even further. The hyperparameter tuning for the ComParE2016 based model resulted in the following settings: max-depth = 10, max-features = SQRT, min-samples-leaf = 2, min-samples-split = 5, n-estimators = 50. For the GeMaps1b model, the following parameter settings were found: max-depth = 30, max-features = SQRT, min-samples-split = 5, n-estimators = 50.

The most important features for the ComParE2016 random forest model and the GeMaps1b linear regression model can be seen in the table below. The importance of features is based on Shapley values that have been extracted through a SHAP analysis for every model [41].

For the ComParE2016 model it can be seen that out of the five most important features, two are mel frequency cepstral coefficients. These coefficients describe the general shape of the power spectrum of an audio excerpt. When looking at the GeMaps1b linear regression model it becomes apparent that the three most influential features all originate from the same low-level descriptor 'F0'. This feature refers to the fundamental frequency in an audio excerpt. Previous research has shown that the fundamental frequency feature is correlated with stress [60]. Research has further shown that depression can lead to increased levels of stress. Hence, the results from the GeMaps1b model align with psychological theory and previous work [61]. For a more extensive explanation of every feature the OpenSMILE software documentation is referred to [38].

Table 4: Overview of five most important features in the random forest depression prediction model based on selected ComparE2016 features according to SHAP analaysis

Top #	**LLD**	**Turn Func.**	**Session Func.**	**Explanation Power**
1	MFCC (3)	Mean	Min	
2	Spectral Variance	Mean	Max	
3	MFCC (5)	Mean	Mean	
4	Spectral Roll-Off (90%)	Mean	Max	
5	Auditory Spectrum (R-Filter)	Max	Std Dev.	

It should be noted that for all featuresets, the linear regression outperforms the random forest on the development set. However, when looking at the performance on the test set it becomes clear that the random forest outperforms the linear regression model in all cases except for the VGGish model. This can possibly be explained by the fact that linear regression models can only capture linear relationships between the independent and dependent variables whereas tree-based models (e.g. a random forest) are also able to capture non-linear relationships between the independent and dependent variables. Thus, it could be the case that the used input features are the most indicative of depression but that the features' relationship to

Table 5: Overview of five most important features in the linear regression depression prediction model based on selected GeMaps1b features according to SHAP analysis

Top #	LLD	Turn Func.	Session Func.	Explanation Power
1	Pitch	Mean	Std Dev.	
2	Pitch	50th perc.	Std Dev.	
3	Pitch	20th perc	Std Dev.	
4	Loudness	Std Dev.	Std Dev.	
5	F1 Frequency	Mean	Std Dev.	

the depression score is not linear, therefore resulting in a lower performance when fitted on a linear model.

6.3.2 Breathing Models

For the breathing-based predictive models, different combinations of groups of breathing-related features have been used to train models. The two best performing models are reported in the table below together with the used input features. Both models have been optimized on the development set. It can be seen that the model taking the Peak / Zero-Crossing / Shape features performs best on both the development and the test set. More specifically, the linear regression model provides the model with the best performance on the development set and should thus be regarded as the best performing model. However, when looking at test set performance, the random forest model outperforms the linear regression model. Again, the models are optimized on the development set thus one cannot conclude that because the random forest model outperforms the linear regression model on the test set that the random forest model is the best model.

In order to contextualize the performance of these breathing models, the results from a recent study by Pessanha et al. will be brought forward. The study also researched the role of breathing-based features on the detection of depression in the DAIC-WOZ dataset. Pessanha et al. bring forward the results of a linear regression as well as a random forest trained on breathing features. The linear regression reaches a RMSE of 5.98 on the development set and 7.65 on the test set. The random forest model attains a RMSE of 6.98 on the development set and 6.40 on the test set [57]. When comparing Pessanha et al.'s results to the results of the models trained for this research it becomes clear that these models perform worse on the development set but better on the test set. When comparing the results to the performance of the baseline models it becomes clear that the breathing-based models substantially outperform the audio-only baseline models on both the development and the test set. Lastly, when looking at the performance of the paralinguistic models brought forward in the previous section it becomes clear that the best performing paralinguistic model outperforms the best performing breathing-based model (RMSE of 6.16 vs. 6.29 on the development set) but that the

breathing-based model generalizes much better to the test set (RMSE of 7.40 vs. 6.50). For all breathing-based models the difference in performance between development and test set is relatively small on the suggested models. This suggests that the models generalize well and do not overfit.

Feature Set	**Model Architecture**	**Performance (RMSE)**	
		Dev	*Test*
Peak / Zero-Crossing / Shape	LR	**6.29**	6.50
	RF	6.54	**6.18**
Peak-to-Peak / Slope / Inhale	LR	6.66	6.55
	RF	6.91	6.88

Table 6: Performance of best breathing models on dev and test set

For the random forest models, besides feature selection, hyperparameter tuning has also been performed to optimize model performance. A grid search was completed to find a combination that produces a model that reaches a (local) optimum. This grid search resulted in the following settings for the hyperparameters: max-depth = 20, max-features = SQRT, min-samples-leaf = 4, min-samples-split = 10, n-estimators = 50.For the linear regression, the only optimization that has been done is feature selection. In the table below, an overview is given of the most important features for both the linear regression model and the random forest model (both on the peak/zero-crossing/shape featureset).

Top #	**Linear Regression**	**Random Forest**
1	std_dev std_dev	mean mean
2	max max	max max
3	mean min	mean min
4	mean mean	std_dev std_dev
5	min mean	mean max

Table 7: Most important features for best performing breathing-based models according to SHAP analysis

From this table it becomes apparent that the most influential features all originate from the statistical functionals. For the features in the table the first word refers to the functional applied to every individual breathing prediction. The second word refers to the statistical function that is being applied to summarize the entire interview. For example: 'mean mean' refers to the mean of the means of the individual predictions, thus the mean breathing volume of the entire interview. In one model, three out of the five most important features come from the mean of the predicted breathing signals and in the other model, two out of the five most important features come from the mean of the breathing signal. In practice the mean score of a

predicted breathing signal refers to the average value that is predicted for the upper-belt value with which breathing is normally measured. In this specific case, the two features that are present in the list of most influential features of both models are 'mean min' and 'mean mean'. This implies that in the interviews, the breathing signal with the lowest mean is indicative of the presence of depression and that the mean of the breathing signal over the entire interview is also indicative of the presence of depression.

6.3.3 Correlation Analysis

6.4 Answer Clustering-Based Models

Having established the performance of breathing-based models that are trained on the entire training set, this section discusses the results of a different approach. One of the research questions that this thesis aims to answer revolves around the relationship between linguistic properties and breathing properties for the detection of depression. The experiments conducted to answer this question consist of automatically segmenting the answers provided by the patients based on three emotional dimensions: dominance, arousal and valence. For every dimension, multiple segmentation boundaries with respect to the score have been tested. Only the best performing answer segments will be brought forward here. In appendix A a more extensive set of graphs can be found that give insight into the results of all segments on which models were trained.

For the emotional dimension of arousal, the best performing model was the one that was built on the answer segment of the data consisting of answers that scored between 0.75 and 1. This linear regression model reached a performance of 5.95 (RMSE) on the development set and 6.36 on the test set. Thereby beating the performance of the breathing model that was built on all answers provided in the interviews. This implies that looking at the breathing patterns that are present when a high level of arousal is detected in the language used by the patient allows for a more accurate prediction of depression with respect to the breathing model that was trained on all answers. In the table below an overview is given of the performance of the models trained on all four segments.

Next, when looking at the emotional dimension of dominance, the results of the experiments done suggest that the best performing model is built on answers that score between 0 and 0.33. This model reaches a performance of 5.89 on the development set and 7.13 on the test set. Out of all conducted experiments with segmentation based on emotional dimensions scores this model reached the highest performance on the development set (on which it was optimized). Remarkably, for the dominance dimension the best performing model is one which was built on breathing data that occurred during the use of language that indicates low dominance. A possible explanation for this is that

Lastly, for the emotional dimension of valence, the results of the experiment indicate that the model that is built on the breathing predictions of answers that score between 0.5 and 1 on the valence dimension scores the highest and reaches a performance of 6.08 (RMSE) on the development set and 6.67 on the test set. This performance indicates that depression prediction models that are built on predicted breathing data that occurs when language is being used that indicates high valence, outperform models that are trained on breathing predictions from the entire session.

Because the best performing models mentioned above are built on subsets of the entire dataset it is

Emotional Dimension	Score Boundaries	Performance (RMSE)	
		Dev	*Test*
Arousal	0.75-1	5.95	6.36
Dominance	0-0.33	5.89	7.13
Valence	0.5-1	6.08	6.67

Table 8: Best performing models with score boundaries for all three emotional dimensions

important to analyze the distribution of datapoints between the training, development and test set on which the models are built. In the table below an overview can be seen for the three models that are described above. For valence this means all answers that score between 0.5 and 1, for dominance this means all answers that score between 0 and 0.33 and lastly, for arousal this means all answers that score between 0.75 and 1.

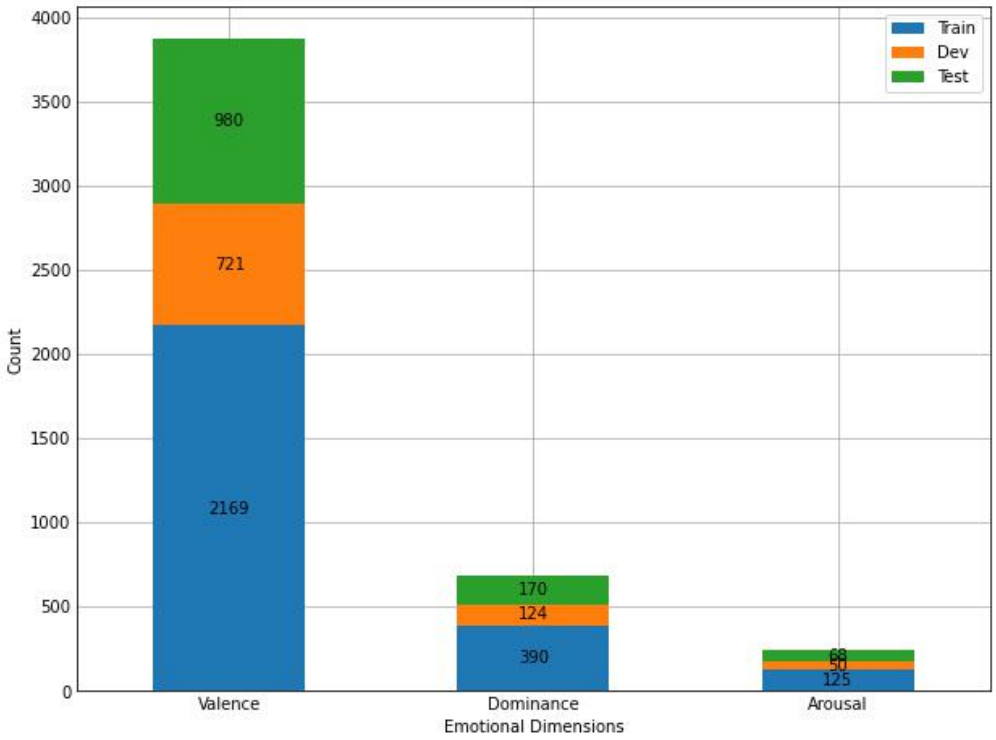

Figure 8: Distribution of number of datapoints between train / dev / test set for the three best performing breathing models built on answers that score between specific thresholds for the emotional dimensions. Valence (0.5-1), Dominance (0-0.33), Arousal (0.75-1)

When looking at the distribution of data points between the train, development and test set for the three models it becomes apparent that the data is distributed relatively evenly for all the models. For the model that was built on answers that score between 0.5 and 1 on the valence dimension, the total number of answers is 3870. Out of this total, 2169 answers make up the training set, the development set consists of 721 answers and the test set contains. 980 answers. This means that 56% of the answers are in the training set, 19% in

the development set and 25% in the test set. For the model that was built using answers that score between 0 and 0.33 on the dominance dimensions the total number of answers used is 684. 390 answers make up the training set, 124 make up the development set and the test set consists of 170 answers. This means that for this model, 57% of the answers make up the training set, 18% makes up the development set and 25% makes up the test set. Lastly, for the model that was built on answers that score between 0.75 and 1 on the arousal dimension, the total number of answers is 243, the training set consists of 125 answers, the development set consists of 50 answers and the test set is made up by 68 answers. Again, to be able to compare the distribution between the models, in terms of percentage this means that the training set makes up 51% of the answers, the development set makes up 21% and the test set makes up 28%. Taking all these numbers together it becomes clear that the number of answers on which each model was built is heavily dependent on the magnitude of the range, thus the models trained on answers that score in a smaller range on a specific emotion resulted to have less datapoints to be built on. However, when looking at the distribution it becomes clear that all three models have a relatively equal split between train, development and test set (in %).

	Modality	Feature Set	Model	Performance	
				Dev	*Test*
Challenge Baseline	Audio		RF	6.74	7.78
	Video		RF	7.13	**6.97**
	Audio + Video		RF	**6.62**	7.05
Gong et al.	Audio + Video + Text		SGD LR	3.54	4.99
Sun et al.	Text		RF	4.97	**4.98**
Yang et al.	Audio + Video + Text		DL	**3.09**	5.40
Yang et al.	Audio + Video + Text		DL	4.65	5.97
Proposed Baseline	Text	S-BERT	LR	5.50	5.75
Proposed Methods	Audio	Paralinguistic	LR	**6.16**	7.40
		Breathing	LR	6.29	**6.50**
	Selected Audio	Breathing (val.)	LR	6.08	6.67
		Breathing (dom.)	LR	**5.89**	7.13
		Breathing (aro.)	LR	5.95	**6.36**

Table 9: Performance of all proposed methods together with existing and suggested baselines

7 Conclusion

7.1 Discussion

7.2 Limitations

7.3 Further Research

Bibliography

[1] World Health Organization et al. *Depression and other common mental disorders: global health estimates*. Tech. rep. World Health Organization, 2017.

[2] Mary Jane Friedrich. "Depression is the leading cause of disability around the world". In: *Jama* 317.15 (2017), pp. 1517–1517.

[3] Catherine K Ettman et al. "Prevalence of depression symptoms in US adults before and during the COVID-19 pandemic". In: *JAMA network open* 3.9 (2020), e2019686–e2019686.

[4] Christina Sobin and Harold A Sackeim. "Psychomotor symptoms of depression". In: *American Journal of Psychiatry* 154.1 (1997), pp. 4–17.

[5] Eliseo J Pérez-Stable et al. "Depression in medical outpatients: underrecognition and misdiagnosis". In: *Archives of Internal Medicine* 150.5 (1990), pp. 1083–1088.

[6] Rosalind Picard. "Affective Computing". In: *MIT Media Laboratory Perceptual Computing Section Technical Report* 321 (1995).

[7] Soujanya Poria et al. "A review of affective computing: From unimodal analysis to multimodal fusion". In: *Information Fusion* 37 (2017), pp. 98–125.

[8] Max Hamilton. "A rating scale for depression". In: *Journal of neurology, neurosurgery, and psychiatry* 23.1 (1960), p. 56.

[9] Aaron T Beck, Robert A Steer, and Gregory K Brown. *Beck depression inventory (BDI-II)*. Vol. 10. Pearson London, UK, 1996.

[10] A John Rush et al. "The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression". In: *Biological psychiatry* 54.5 (2003), pp. 573–583.

[11] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. "The PHQ-9: validity of a brief depression severity measure". In: *Journal of general internal medicine* 16.9 (2001), pp. 606–613.

[12] Klaus R Scherer. "Vocal affect expression: a review and a model for future research." In: *Psychological bulletin* 99.2 (1986), p. 143.

[13] Gary Christopher and John MacDonald. "The impact of clinical depression on working memory". In: *Cognitive neuropsychiatry* 10.5 (2005), pp. 379–399.

[14] Alan Baddeley. "Working memory and language: An overview". In: *Journal of communication disorders* 36.3 (2003), pp. 189–208.

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

[16] Florian Hönig et al. “Automatic modelling of depressed speech: relevant features and relevance of gender”. In: (2014).

[17] Michael Cannizzaro et al. “Voice acoustical measurement of the severity of major depression”. In: *Brain and cognition* 56.1 (2004), pp. 30–35.

[18] James C Mundt et al. “Vocal acoustic biomarkers of depression severity and treatment response”. In: *Biological psychiatry* 72.7 (2012), pp. 580–587.

[19] Frans A Boiten, Nico H Frijda, and Cornelis JE Wientjes. “Emotions and respiratory patterns: review and critical analysis”. In: *International journal of psychophysiology* 17.2 (1994), pp. 103–128.

[20] Pat Ogden and Kekuni Minton. “Sensorimotor psychotherapy: One method for processing traumatic memory”. In: *Traumatology* 6.3 (2000), pp. 149–173.

[21] Yuri Masaoka and Ikuo Homma. “Anxiety and respiratory patterns: their relationship during mental stress and physical load”. In: *International Journal of Psychophysiology* 27.2 (1997), pp. 153–159.

[22] AT Beck. *Depression: Clinical, experimental and theoretical aspects London*. 1967.

[23] Tom Pyszczynski and Jeff Greenberg. “Self-regulatory perseveration and the depressive self-focusing style: a self-awareness theory of reactive depression.” In: *Psychological bulletin* 102.1 (1987), p. 122.

[24] Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. “Language use of depressed and depression-vulnerable college students”. In: *Cognition & Emotion* 18.8 (2004), pp. 1121–1133.

[25] Shannon Wiltsey Stirman and James W Pennebaker. “Word use in the poetry of suicidal and nonsuicidal poets”. In: *Psychosomatic medicine* 63.4 (2001), pp. 517–522.

[26] Nairan Ramirez-Esparza et al. “The psychology of word use in depression forums in English and in Spanish: Testing two text analytic approaches”. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 2. 1. 2008, pp. 102–108.

[27] Xinyu Wang et al. “A depression detection model based on sentiment analysis in micro-blog social network”. In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer. 2013, pp. 201–213.

[28] Amna Amanat et al. “Deep learning for depression detection from textual data”. In: *Electronics* 11.5 (2022), p. 676.

[29] Lang He and Cui Cao. “Automated depression analysis using convolutional neural networks from speech”. In: *Journal of biomedical informatics* 83 (2018), pp. 103–111.

[30] Karol Chlasta, Krzysztof Wołk, and Izabela Krejtz. "Automated speech-based screening of depression using deep convolutional neural networks". In: *Procedia Computer Science* 164 (2019), pp. 618–628.

[31] Tuka Al Hanai, Mohammad M Ghassemi, and James R Glass. "Detecting Depression with Audio/Text Sequence Modeling of Interviews." In: *Interspeech*. 2018, pp. 1716–1720.

[32] Jiayu Ye et al. "Multi-modal depression detection based on emotional audio and evaluation text". In: *Journal of Affective Disorders* 295 (2021), pp. 904–913.

[33] Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. "Early versus late fusion in semantic video analysis". In: *Proceedings of the 13th annual ACM international conference on Multimedia*. 2005, pp. 399–402.

[34] James W Pennebaker, Martha E Francis, and Roger J Booth. "Linguistic inquiry and word count: LIWC 2001". In: *Mahway: Lawrence Erlbaum Associates* 71.2001 (2001), p. 2001.

[35] James A Russell. "A circumplex model of affect." In: *Journal of personality and social psychology* 39.6 (1980), p. 1161.

[36] James A Russell. "Core affect and the psychological construction of emotion." In: *Psychological review* 110.1 (2003), p. 145.

[37] Nils Reimers and Iryna Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks". In: *arXiv preprint arXiv:1908.10084* (2019).

[38] Florian Eyben, Martin Wöllmer, and Björn Schuller. "Opensmile: the munich versatile and fast open-source audio feature extractor". In: *Proceedings of the 18th ACM international conference on Multimedia*. 2010, pp. 1459–1462.

[39] Karen Simonyan and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556* (2014).

[40] Maxim Markitantov et al. "Ensembling End-to-End Deep Models for Computational Paralinguistics Tasks: ComParE 2020 Mask and Breathing Sub-Challenges." In: *INTERSPEECH*. 2020, pp. 2072–2076.

[41] Scott M Lundberg and Su-In Lee. "A unified approach to interpreting model predictions". In: *Advances in neural information processing systems* 30 (2017).

[42] 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.

[43] 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.

[44] Gilles Degottex et al. “COVAREP—A collaborative voice analysis repository for speech technologies”. In: *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE. 2014, pp. 960–964.

[45] 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.

[46] Andrew Bailey and Mark D Plumbley. “Gender Bias in Depression Detection Using Audio Features”. In: *2021 29th European Signal Processing Conference (EUSIPCO)*. IEEE. 2021, pp. 596–600.

[47] 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*. 2017, pp. 3–9.

[48] Quoc Le and Tomas Mikolov. “Distributed representations of sentences and documents”. In: *International conference on machine learning*. PMLR. 2014, pp. 1188–1196.

[49] Tomas Mikolov et al. “Distributed representations of words and phrases and their compositionality”. In: *Advances in neural information processing systems* 26 (2013).

[50] Le Yang et al. “Multimodal measurement of depression using deep learning models”. In: *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. 2017, pp. 53–59.

[51] 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.

[52] Zafi Sherhan Syed, Kirill Sidorov, and David Marshall. “Depression severity prediction based on biomarkers of psychomotor retardation”. In: *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. 2017, pp. 37–43.

[53] 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.

[54] Brian McFee et al. “librosa: Audio and music signal analysis in python”. In: *Proceedings of the 14th python in science conference*. Vol. 8. 2015, pp. 18–25.

[55] James Lyons et al. “James lyons/python speech features: Release v0. 6.1”. In: *Zenodo. doi* 10 (2020).

[56] Björn W Schuller et al. “The interspeech 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks”. In: (2020).

[57] Francisca Pessanha et al. “Towards using Breathing Features for Multimodal Estimation of Depression Severity”. In: *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION*. 2022, pp. 128–138.

[58] Saif Mohammad. “Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words”. In: *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*. 2018, pp. 174–184.

[59] Le Yang, Dongmei Jiang, and Hichem Sahli. “Feature augmenting networks for improving depression severity estimation from speech signals”. In: *IEEE Access* 8 (2020), pp. 24033–24045.

[60] Martijn Goudbeek and Klaus Scherer. “Beyond arousal: Valence and potency/control cues in the vocal expression of emotion”. In: *The Journal of the Acoustical Society of America* 128.3 (2010), pp. 1322–1336.

[61] Constance Hammen. “Stress and depression”. In: *Annual Review of Clinical Psychology(2005)* 1.1 (2005), pp. 293–319.

Appendix A