Gender Fairness in Depression Severity Prediction using Multimodal Models

MSc Thesis

by
Martje (Mart) Jacoba Koek
0575208

Submitted to the Artificial Intelligence Graduate Program
in partial fulfillment of the requirements for the degree of Master of Science

Utrecht University,
July 3, 2024

First supervisor:	Dr. Heysem Kaya
Daily supervisor:	Gizem Soğancıoğlu
Second examiner:	Prof. dr. Albert A. Salah

Abstract

Gender Fairness in Depression Severity Prediction using Multimodal Models

Mental health illnesses cause significant suffering for individuals, their families, and society. Early, accurate, and responsible detection of mental health problems is crucial for effective intervention. This research aims to develop responsible methods to assist (not replace) medical experts in depression detection. The performance and gender fairness of uni- and bimodal audio-text models to predict depression severity are explored using the DAIC-WOZ dataset. Three key research questions are addressed: the comparative performance of uni- and bimodal models, the gender fairness of these models, and the effect of bias mitigation methods. The findings indicate that the unimodal text model outperforms state-of-the-art uni- and bimodal audio-text models. The best bimodal models could not improve the performance of our unimodal text model but outperform state-of-the-art bimodal audio-text models. Gender biases were found in all unimodal and bimodal models, with a general trend per modality: text models showed a bias favoring males and audio models favoring females. The bias mitigation methods showed mixed results, sometimes improving fairness but at the cost of overall performance.

Nederlands Abstract

Geestelijke gezondheidsproblemen veroorzaken leed voor individuen en hun omgeving. Vroegtijdige, accurate en verantwoorde detectie van geestelijke gezondheidsproblemen is cruciaal voor effectieve interventie. Deze studie onderzoekt methoden om medische experts te helpen (niet vervangen) bij het detecteren van depressie. Uni- en bimodale audio-tekstmodellen worden getest met behulp van de DAIC-WOZ dataset in het voorspellen van de ernst van depressie. Drie onderwerpen worden behandeld in deze studie: de vergelijking van uni- en bimodale modellen, de gender-eerlijkheid (*fairness*) van deze modellen en het effect van methoden om oneerlijkheid (*bias*) te verminderen. De resulaten laten zien dat het unimodale tekstmodel beter presteert dan bestaande uni- en bimodale audio-tekstmodellen. De beste bimodale modellen presteren niet beter dan ons unimodale tekstmodel, maar wel dan bestaande bimodale audio-tekstmodellen uit andere studies. Alle unimodale en bimodale modellen lieten genderverschillen zien, met een algemene trend per modaliteit: tekstmodellen presteerden beter voor mannen en audiomodellen voor vrouwen. De methoden om *bias* te verminderen lieten verschillende resultaten zien, waarbij de *fairness* soms verbeterde, meestal ten koste van de algehele prestaties.

Acknowledgements

I would like to thank Heysem for his patience, numerous ideas throughout this project, and overall warm supervision style. His quick responses to questions, consistent hosting of our weekly meetings with the essential refreshments, as well as his understanding of, and interest in my other activities, made the project very rewarding. I would also like to thank Gizem, for bringing humor into our weekly meetings, keeping us on track with time management during these meetings, and providing the necessary encouragement during stressful periods.

Furthermore, I am thankful to Stan, Lara, and Sytse for their companionship during this research project. Our shared experience and Wednesday working sessions made the project enjoyable, and the discussions about our progress and next steps helped shape my research. I am grateful to Prof. Salah for being my second examiner and dedicating time to provide feedback on my research proposal. Lastly, I would like to thank Wouter, for supporting me during this project and all my other pursuits.

Contents

List of Abbreviations

Acc	Accuracy
A(EA)	Audio modality (with emotional attention)
AEF	Audio modality with emotional filter
AI	Artificial Intelligence
AVEC	Audio-Visual Emotion Challenge
CNN	Convolutional Neural Network
CV	Cross-validation
DAIC	Distress Analysis Interview Corpus
DF	Decision Fusion
DSM	Diagnostic and Statistical Manual of Mental Disorders
EA	Emotional Attention
EASE	Engagement Arousal Self-Efficacy (dataset)
EF	Emotional Filter
EqAcc	Equal Accuracy
EqOpp	Equal Opportunity
FF	Feature Fusion
GPT	Generative Pre-trained Transformers
(G)WDF	(Gender) Weighted Decision Fusion
(K)ELM	(Kernel) Extreme Learning Machine
MAE	Mean Absolute Error
ML	Machine Learning
MDD	Major Depressive Disorder
NLP	Natural Language Processing
PDEM	Public Dimension Emotional Model
PHQ-8	Patient Health Questionnaire-8
PredEq	Predictive Equality
PTSD	Post-Traumatic Stress Disorder
RMSE	Root Mean Squared Error
R(S)Q	Research (Sub) Question
(S)BERT	(Sentence) Bidirectional Encoder Representations from Transformers
SMILE	Speech and Music Interpretation by Large-space Extraction
SP	Statistical Parity
T(EA)	Text modality (with emotional attention)
TN/FN(R)	True/False Negative (Rate)
TP/FP(R)	True/False Positive (Rate)
WOZ	Wizard-of-Oz (condition in DAIC)

List of Symbols

c	Index representing a class
C	Regularization coefficient of KELM
d_k	Mental disorder k
D	Set of all mental disorders, $D = \{d_1, d_2, \ldots, d_n\}$
$\mathcal{D}$	Training dataset with n instances and m features
f	Number of functionals
F_i	Feature vector for participant i
G	Sensitive attribute, $G \in \{m, f\}$
I	Identity matrix of size $n \times n$
i	Index representing a person
j	Index representing a symptom
k	Index representing a mental disorder
l	Index representing a value in an embedding
$\mathcal{K}$	Kernel matrix of size $n \times n$
m	Number of features in dataset
m_{PDEM}	Number of features in SBERT embedding
m_{SBERT}	Number of features in SBERT embedding
M_{EqAcc}	Fairness metric for equal accuracy (Gender-ratio in RMSE)
M_{EqOpp}	Fairness metric for equal opportunity (Gender-ratio in TPR)
M_{PredEq}	Fairness metric for predictive equality (Gender-ratio in FPR)
M_{SP}	Fairness metric for statistical parity
n	Number of instances in dataset
n_c	Number of instances in class c
n_i	Number of embeddings/instances for participant i
$\mathcal{N}$	Number of participants
o	Index representing a sentence/turn embedding
p_i	Health condition of person i
p_i^k	Indicator variable for person i labeled with disorder d_k
s_j	Symptom j
S	Set of all symptoms
S_k	Set of binary symptoms linked to mental disorder d_k
S^i	Set of binary symptoms that person i suffers from
S_i	True depression severity score for participant i
$\hat{\mathrm{S}}_i$	Predicted depression severity score for participant i
$\mathcal{T}$	Target variable/matrix

V_i	Set of sentence embeddings for participant i
v_o^A	Audio embedding o
v_{io}	Sentence embedding o for participant i
v_o^T	Audio embedding o
w_A^{f}	Female audio weight in decision fusion
w_T^{f}	Female text weight in decision fusion
w_A^{m}	Male audio weight in decision fusion
w_T^{m}	Male text weight in decision fusion
x_{ij}	Indicator variable for person i having symptom s_j
$\mathbf{x}$	Test instance vector
y_{ij}	Symptom value $\in \{0,1,2,3\}$ for the j'th symptom of the PHQ8
y_i	Ground truth symptom vector for participant i
$\hat{\mathrm{y}}_i$	Prediction for the test instance of participant i
$\hat{y}_i$	Constrained prediction for the test instance of participant i
Y	Binary depression outcome, $Y \in \{0,1\}$
$\hat{Y}$	Binary depression prediction, $\hat{Y} \in \{0,1\}$

β	Coefficient vector/matrix obtained from the regression KELM
θ_k	Threshold number of symptoms for disorder d_k
λ	Sample from a beta distribution
$\phi(\mathcal{D}, \mathbf{x})$	Kernel function evaluated on the training data $\mathcal{D}$ and test instance $\mathbf{x}$

List of Figures

List of Tables

1 Introduction

Mental health illnesses can cause a lot of suffering for individuals, their relatives, and society. People with severe mental disorders suffer not only from the burden of common symptoms of mental diseases such as emotional distress, sleep problems, and low mood, but they also represent a vulnerable and socially excluded part of society. They will likely be affected by lower social and educational opportunities, social alienation, increased morbidity (likelihood of having another medical condition), and increased mortality rates (Doran & Kinchin, 2017). Early and accurate detection is crucial for timely intervention and effective treatment, thereby decreasing the collective burden of mental health illnesses. This research aims to explore methods that can assist – rather than replace – medical experts in the diagnostic process in a responsible manner, to ensure the best possible outcomes for patients.

1.1 Problem Statement

Currently, in the Netherlands, there could be a six-month gap between the first consultation and the diagnosis of a mental disorder (Patientenfederatie Nederland, 2023). The average additional waiting time before the treatment begins is also long: for depression 19.2 weeks; for eating disorders 21 weeks; for traumas 21 weeks (Volksgezondheid en Zorg, 2023). This prolongs the stress of mental health illnesses on the people directly and indirectly affected by these disorders. Therefore, it is important to look into ways to improve the process of detecting mental illnesses.

The first version of the Diagnostic and Statistical Manual of Mental Disorders (DSM) was published in 1952 (American Psychiatric Association, 1952). This DSM enabled a systematic approach to diagnosing people with mental illnesses because it contained descriptions of mental disorders, their symptoms, and differential diagnoses (methods to distinguish a particular mental disorder from other disorders with similar symptoms). Since the first version, five editions have been published, of which the latest was released in 2013 (American Psychiatric Association, 2013b). Each new version of the DSM is a revision of the previous edition, with disorders being added, removed, or modified to reflect the latest research and clinical insights. The DSM is used by medical experts, making the diagnostic process more systematic, but a comprehensive diagnosis still takes considerable time.

Disorders should be identified correctly because receiving a diagnosis can have legal implications (Stein et al., 2021), consequences for insurance coverage (Tkacz & Brady, 2021), but also personal implications (Sims et al., 2021). Additionally, misdiagnosing a psychiatric disorder as the wrong disorder, a physical illness, or as drug/alcohol abuse are reasons for psychiatric nonresponse, incomplete response, and relapse (Shen et al., 2018; Estroff & Gold, 2018).

Machine learning (ML) models have been suggested to improve the detection of mental illnesses (Shatte et al., 2019; Khoo et al., 2024; Liu et al., 2021). ML models can be trained on large amounts of various data sources – such as physiological measures, motor activity, or behavioral signals – and predict relatively accurately and quickly whether someone suffers from a mental illness (Wright-Berryman et al., 2023; Mallol-Ragolta et al., 2019; Adarsh et al., 2023). Therefore, these models can be used as decision-support tools to improve the diagnostic process. To reiterate, the goal of using ML is not to replace a diagnosis by a medical professional but to assist medical experts during the diagnosis process responsibly.

In recent studies, clinical notes (Sogancioglu et al., 2023), behavioral data with audio information (Bailey & Plumbley, 2021), and video information from social media (Yoon et al., 2022) have been used to detect mental illnesses with ML models. These models, however, sometimes show gender inequalities in their accuracy (Cheong, Kuzucu, et al., 2023). While there are gender differences for some mental disorders reported in the mental health literature (Afifi, 2007), gender fairness in detecting these disorders is necessary because inaccurate predictions for certain groups can lead to unequal access to mental health resources and support, thereby possibly increasing unfairness.

Gender bias was reported by Bailey & Plumbley (2021) in a multimodal dataset that is often used to train and test ML models in detecting mental illnesses in a clinical setting: the Distress Analysis Interview Corpus (DAIC; Gratch et al., 2014). The biased outcomes could be mitigated, but not all available data modalities were used in this study to detect depression (Bailey & Plumbley, 2021). Only audio (acoustic) information was used, although the text modality (with semantic information) has been shown to be the most informative for depression prediction on this corpus, by the winners of the Audio-Visual Emotion Challenge (AVEC)'2019: Ray et al. (2019), but also by Van Steijn et al. (2022) and Oureshi et al. (2021).

1.2 Research Objectives

The current study will concern gender fairness in depression detection using clinical interviews with multimodal (audio and textual) information, thereby building on earlier work (Bailey & Plumbley, 2021; Wei et al., 2023). The combination of audio information and textual information is expected to reveal important gender patterns (Oureshi et al., 2021) and might improve the (fairness in) detection of mental disorders. If the gender bias remains, it is important to understand the source of the bias. To investigate the bias, different bias mitigation methods will be explored.

The contributions of this study will include an evaluation of the fairness of state-of-the-art uni- and multimodal models in detecting mental illnesses, using symptoms from the DSM. Overall, this study aspires also to contribute to a more effective and responsible detection of mental illnesses. In future practices, clinicians may use ML systems as a support tool in diagnosing different health conditions, therefore it is crucial to adhere to ethical standards, such as fairness, in the development phase of these systems.

1.3 Outline

This paper will be structured as follows. First, important background information and related research will be presented in Chapter 2. Then, computational approaches will be discussed in Chapter 3. Chapter 4 will present the research questions and explain the approach to answering the research questions. The results will be presented in Chapter 5, followed by a discussion of the results, the limitations, and a conclusion in Chapter 6.

2 Domain Background and Related Work

This section covers closely related literature on using ML to detect mental health illnesses. First, an introduction to mental disorders will be provided, along with various topics regarding their study. Next, relevant research on fairness regarding mental disorders will be discussed, as well as research concerning fair ML practices. The dataset used in this study is then described. Finally, the chapter concludes by identifying the research gap.

2.1 Mental Disorders

References to mental disorders – characterized by abnormal behavior with invisible causes – can be found throughout history (Hallowell, 1934; Stengel, 1959; Kessler et al., 2005). Different explanations have been popular, including supernatural notions proposing possession by demonic spirits, curses, or sin; somatogenic theories assigning disturbances to genetic inheritance or brain damage; and psychogenic theories focusing on traumatic or stressful experiences, or maladaptive learned associations and thought patterns (Farreras, 2019).

The classification of behavior as normal or abnormal has always been context-dependent, and currently the psychological discipline uses a biopsychosocial model to explain human behavior (Farreras, 2019). While individuals may be born with a genetic predisposition for some psychological disorder, certain psychological stressors need to be present to develop the disorder. As earlier mentioned, the first DSM provided a standardized diagnostic classification system, thereby creating a shared language that would aid clinical research. The fundamental idea of the DSM is that mental disorders are associated with symptoms of distress or impairment (Bolton, 2013).

2.1.1 Mental Disorders as Collections of Symptoms

The DSM assumes that people suffering from a mental disorder can be classified according to their symptoms, and therefore that a set of symptoms characterizes a particular mental disorder. These symptoms are considered binary: a symptom is either present or absent. Medical professionals use this symptom-based classification system to diagnose mental illnesses. The binary system has faced criticism, as it does not reflect the complexity of real-life experiences, where symptoms can vary in intensity and frequency (Lahey et al., 2022). However, due to health insurance requirements and the legal implications of a diagnosis, medical experts must justify their diagnoses based on the presence of specific symptoms.

Summarizing, a person has a mental disorder if the corresponding subset of symptoms is sufficiently present. This can be mathematically expressed as follows: let p_i represent the health condition of a person i. Let S be the set of all symptoms, $S = \{s_1, s_2, \ldots s_j, \ldots\}$. Let D be the set of all mental disorders in the DSM, $D = \{d_1, d_2, \ldots, d_k, \ldots\}$. Let S_k be the set of symptoms linked to a mental disorder d_k, and S^i be the set of symptoms that a person i suffers from. It is then possible to express that a person i has a symptom s_j with x_{ij}:

$$x_{ij} = \begin{cases} 1 & \text{if } s_j \in S^i \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

It is also possible to express that at least θ_k of the set of symptoms (S_k) from disorder d_k need to be present in a person i for them to be labeled with that mental disorder. So person i can be labeled with disorder d_k through variable p_i^k:

$$p_i^k = \begin{cases} 1 & \text{if } \displaystyle\sum_{s_j \in S_k} x_{ij} \geq \theta_k \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

An abstract representation of two mental disorders with overlapping symptoms and a person suffering from a subset of symptoms is illustrated in Figure 2.1.

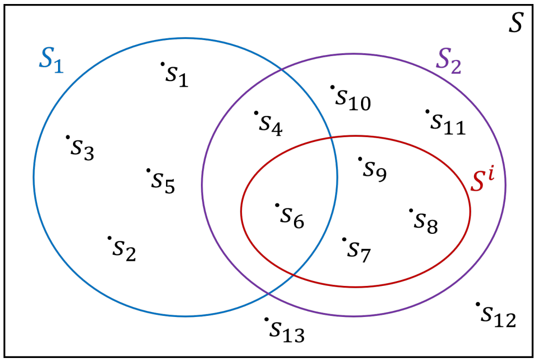


Figure 2.1: Hypothetical representation of symptoms and mental disorders in DSM. S_1 and S_2 are two sets of symptoms from hypothetical disorders d_1 and d_2, S^i is the set of symptoms that person i suffers from.

The representation of overlapping symptoms in Figure 2.1 can be illustrated through the symptom similarities between a major depressive disorder (MDD) and other mental disorders. As the symptoms of MDD are important for this study, the symptoms of this disorder according to DSM-V are presented in Table 2.1 (American Psychiatric Association, 2013a). Five (or more) should have been present during the same 2-week period and represent a change from previous functioning. And at least one of the symptoms is either (1) depressed mood or (2) loss of interest or pleasure.

1. Depressed mood most of the day, nearly every day, as indicated by either subjective report (e.g., feels sad, empty, hopeless) or observation made by others (e.g., appears tearful).
2. Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day (as indicated by either subjective account or observation).
3. Significant weight loss when not dieting or weight gain (e.g., a change of more than 5% of body weight in a month), or decrease or increase in appetite nearly every day.
4. Insomnia or hypersomnia nearly every day.
5. Psychomotor agitation or retardation nearly every day (observable by others, not merely subjective feelings of restlessness or being slowed down).
6. Fatigue or loss of energy nearly every day.
7. Feelings of worthlessness or excessive or inappropriate guilt (which may be delusional) nearly every day (not merely self-reproach or guilt about being sick).
8. Diminished ability to think or concentrate, or indecisiveness, nearly every day (either by subjective account or as observed by others).
9. Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide.

Table 2.1: Symptoms of Major Depressive Disorder in DSM-V.

In addition to this list, it is necessary for a diagnosis MDD that 1) these symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning, 2) the episode is not attributable to the physiological effects of a substance or another medical condition, 3) the occurrence of the major depressive episode is not better explained by schizoaffective disorder, schizophrenia, schizophreniform disorder, delusional disorder, or other specified and unspecified schizophrenia spectrum and other psychotic disorders, and 4) there has never been a manic episode or a hypomanic episode.

This can be expressed mathematically: let S_{MDD} be the set of symptoms for disorder MDD: $S_{\text{MDD}} = \{s_1, ...s_9\}$ from the list above. Then $\theta_{\text{MDD}} = 5$ because at least 5 symptoms need to be present for an individual to have this mental disorder. Person Anne can suffer from the set of symptoms $S^{\text{Anne}} = \{s_2, s_3, s_5, s_6, s_7, s_8\}$ and this could point to MDD, but an extensive analysis is needed to exclude other disorders. The DSM-V includes a differential diagnosis for each disorder to account for overlapping symptoms. For instance, symptom s_8 could be indicative of attention-deficit/hyperactivity disorder, while symptoms s_1 and s_2 could be linked to manic episodes with irritable mood or major depressive episodes with irritable mood. This shows the importance of experts taking time to completely understand an individual and their symptoms, to rule out some disorders, and to diagnose accurately.

Mental illnesses and their symptoms have been studied in various ways. The next section will dive into the different methodologies that can be used to collect and analyze data in the field of mental health research, as this can influence the interpretation and application of the results.

2.1.2 Studying Mental Disorders

This section explores two distinct approaches to gathering behavioral data for mental health assessment: natural/in-the-wild environments and clinical settings. Some studies use a natural setting (Bailey & Plumbley, 2021; Chen et al., 2014; Yoon et al., 2022; T. Zhang et al., 2022; Zogan et al., 2022) , and others choose a clinical or laboratory setting (Sogancioglu et al., 2023; Muzammel et al., 2021; Corcoran et al., 2018; Cohen et al., 2023; Cheong, Spitale, & Gunes, 2023; Adarsh et al., 2023).

Natural Research Setting

Supporters of using data from natural or in-the-wild environments, such as data from social media platforms or smartphones, argue that these platforms provide a realistic context for understanding natural behavior. Data can be relatively easily retrieved and online research may provide a diverse range of participants, while clinical settings may attract a specific demographic. Another advantage of a natural setting is it might show more ‘hidden’ factors that contribute to mental health, such as lifestyle and social factors. Additionally, it has been reported that people might feel more comfortable sharing feelings online than face-to-face (S. Zhao, 2005). It is worth noting, however, that not everyone uses social media to post about their thoughts or feelings.

There is a collection of work that uses behavioral data from social media to predict mental health: for example, an explainable depression detection framework was developed using deep learning of textual, behavioral, temporal, and semantic aspect features from social media (Zogan et al., 2022). With this framework, tweets were analyzed to explain why a user was depressed (which was clinically diagnosed). Similarly, videos from YouTube were used to test audio-video models to detect depression online (Yoon et al., 2022). The depressed state of these individuals was not clinically diagnosed. Text-based detection of mental health symptoms via Reddit has also been proven to be successful (Z. Zhang et al., 2022).

Clinical Research Setting

As this study aims to contribute to improving the process of diagnosis in mental health care, research from a clinical perspective is more relevant than from an in-the-wild perspective. The clinical setting resembles a situation where individuals are intentionally observed for a mental health assessment. Clinical notes or interviews are overseen by trained professionals, who can guide the interpretation, direction, and assessment of conversations about mental health. These settings often are standardized and validated, which can enhance the reliability and comparability of data. Another advantage is that the clinical setting allows for in-depth interviews and medically grounded questions, instead of only observing someone behave naturally.

Some examples of clinical multimodal datasets are the Distress Analysis Interview Corpus (DAIC), a standardized clinical dataset that is widely used for the detection of post-traumatic stress disorder (PTSD) and MDD (Gratch et al., 2014). This dataset will also be used in the current study. Other datasets are the Androids corpus, a recently published Italian dataset for speech-based depression detection (Tao et al., 2023); the Turkish Audio-Visual Bipolar Disorder Corpus with video recordings of patients with bipolar disorder (Çiftçi et al., 2018); the Engagement Arousal Self-Efficacy (EASE) dataset to detect PTSD (Dhamija & Boult, 2017). Overall, increasingly complex neural network approaches demonstrate efficient use of multimodal clinical data for mental health assessment (Khoo et al., 2024).

The natural and clinical research settings may vary in point of view, but both settings seek information about signs of abnormal behavior that can indicate mental disorders.

2.1.3 Behavioral Indicators of Mental Disorders

Existing research on detecting mental disorders is generally categorized into different data sources. One line of work focuses on monitoring physiological data such as physical strength, stress markers, or brain activity (Lever-van Milligen et al., 2020; Su et al., 2014). Another line of work studies motor activity, such as the duration, intensity, or timing of movement (Jakobsen et al., 2020; Difrancesco et al., 2021). A third line of work is grounded in findings that behavioral data differentiates between healthy individuals and individuals who suffer from a mental disorder. The current study falls into this third line of work.

In a clinical setting, information about a patient's speech, verbal communication, visual appearance, and body language may aid clinicians' diagnosis (Yamamoto et al., 2020; Ma et al., 2016; Corcoran et al., 2018). Therefore, this multimodal information can also be used in automatically detecting depression (Cohen et al., 2023). Three modalities will now be discussed that are commonly used in multimodal mental health prediction: the visual, audio, and textual modality.

External visual features reflect information about the internal state of individuals (Adarsh et al., 2023). Although the current study will not look into this topic in this study, some relevant work has been done here (Williamson et al., 2016; Muzammel et al., 2021; Mulay et al., 2020). Facial representations detected by ML models can encode expression information (Li & Deng, 2022), head pose and eye gaze have been used in detecting depression (Alghowinem et al., 2018). The visual modality has often been combined with audio information (Yoon et al., 2022; Booth et al., 2021).

Textual Indicators of Mental Disorders

Language is the basis by which others can infer our thought processes, and disorganized language is therefore considered to reflect disorganized thought. Natural Language Processing

(NLP) entails using computational techniques to analyze natural language and speech (Chowdhary, 2020; Nadkarni et al., 2011). In the context of mental disorder detection, NLP tools are useful for analyzing text data and extracting insights about linguistic patterns, sentiment, and specific words that suggest mental health conditions (Le Glaz et al., 2021; T. Zhang et al., 2022).

Research has shown that automated NLP methods can be used to indicate disturbances in semantics and syntax across different stages of psychotic disorders (Corcoran et al., 2018). NLP models have been used to identify depression, anxiety, and suicide risk in short (5-10 minutes) virtual screenings (Wright-Berryman et al., 2023). Transcribed clinical interviews were successfully used to detect depression (Mallol-Ragolta et al., 2019). The processing of textual information to detect mental health abnormalities can be found in Section 3.1.1.

Audio Indicators of Mental Disorders

In addition to language, vocal behavior (acoustic properties of a speaker's voice; paralinguistics) can also reflect information about the internal state of an individual. According to Schuller (2011), the field of paralinguistics can be divided into speaker states – dealing with changes over time such as affection, emotion, and health states – speaker traits – identifying permanent characteristics such as gender, height, or personality – and vocal behavior. In the context of this study, speaker states are the most relevant.

Traditional analyses of sounds used pitch, loudness, duration, and timbre. For example, the mean, variance, and autocorrelations of loudness, pitch, brightness, and bandwidth were used to classify sound recordings as (vocal behavior) laughter (Wold et al., 1996). There is evidence that produced audio data differs between healthy people and people with a mental disorder (Cummins et al., 2015; Ma et al., 2016; Bailey & Plumbley, 2021; Srimadhur & Lalitha, 2020). Speech markers such as the duration of speech, pitch, and speech tone can signal distress (Adarsh et al., 2023) and depressed participants show longer pause time, longer response time, and slower speech rate than healthy participants (Yamamoto et al., 2020).

Details about processing audio/speech data will be discussed in Section 3.1.2. Additionally, an elaboration on methods to combine data from different modalities – which might reveal even more information about an the internal state – can be found in Section 3.1.3. But first, fairness in terms of mental health and more general, in the field of ML will be discussed.

2.2 Fairness

Fairness[1] is the quality or state of being fair, where fair treatment is defined as the lack of favoritism toward one side or another. Fairness is an important aspect in understanding and treating mental health, as well as in developing and applying ML models. This section addresses

[1] Fairness: the quality or state of being fair; especially: fair or impartial treatment: lack of favoritism toward one side or another. Retrieved from `https://www.merriam-webster.com/dictionary/fairness` (accessed: 11-12-2023)

these two areas by examining fairness in the mental health domain and the strategies for achieving fair ML.

2.2.1 Gender Differences in the Mental Health Domain

An analysis of gender differences in mental health, and specifically depression, showed that there is not always a lack of favoritism towards one side. Doctors are more likely to diagnose depression in women compared to men, even when they have similar scores on standardized measures of depression (Afifi, 2007) and it seems more difficult to diagnose depression in females compared to males (Floyd, 1997).

Differences in detection rates might be caused by differences in latent variables (that are not directly observable) between the genders, or by differences in mental health expression between males and females. The DSM-V reports that although the most reproducible finding in the epidemiology of MDD is a higher prevalence in females, there is no clear difference between genders in symptoms, course, treatment response, or functional consequences (American Psychiatric Association, 2013a). However, early stages of depression in men differ from early stages of depression in women (Ogrodniczuk & Oliffe, 2011), and women generally report more bodily distress and more numerous, more intense, and more frequent somatic symptoms than men (Barsky et al., 2001).

Additionally, during adolescence, girls have a higher prevalence of depression and eating disorders and engage more in suicidal ideation and suicide attempts than boys (Hawton et al., 2002), who are more prone to engage in high-risk behaviors and commit suicide successfully more frequently (Parker & Roy, 2001). Men and women differ in written suicide notes (Lester & Heim, 1992) and suicidal distress (Straw & Callison-Burch, 2020). Other significant differences were found in the personality trait neuroticism (tendency to experience negative emotions) between males and females, where males score typically lower on neuroticism (less negative emotions) than females (Djudiyah et al., 2016). It was also reported that women have a heightened expression of positive emotion and internalizing of negative emotions, and men have an increased expression of anger (Chaplin, 2015).

Despite these differences, fair treatment of all genders is necessary to ensure fair mental health care. As ML has been proposed in various fields for diverse prediction tasks, with sometimes significant consequences for the subjects involved, the topic of fairness has become increasingly important.

2.2.2 Fair Machine Learning

Well-intentioned ML applications in (high-stake) decision domains can sometimes lead to outcomes that raise objections. Barocas et al. (2017) delved into fairness and ML, addressing arbitrary, inconsistent, or faulty decision-making concerns. A recent survey on bias and fairness in ML (Mehrabi et al., 2021) outlined definitions of fairness in this field. The concept of fairness can be categorized into individual and group fairness.

Individual fairness entails that two similar individuals with respect to a detection task should be treated similarly. Group fairness entails that the predictions of a model do not systematically differ between different groups (e.g., gender or age groups). This study will focus on the fairness between groups of people (gender), specific group fairness metrics are therefore discussed in Section 3.3.

When fairness measures are not satisfied, the model or system may show biases against certain groups. These biases can arise from various sources, such as the data on which the model was trained or choices related to the model's design (Mehrabi et al., 2021). In terms of data, ML algorithms typically assume a balanced distribution of samples across different groups (Krawczyk, 2016). When groups are imbalanced, learning algorithms tend to favor the majority group because the chance of an instance being in this group is a priori higher. Additionally, historical biases may be embedded in the data, or variables may not accurately represent groups or may be correlated with protected features. Existing approaches to mitigate biases in ML systems often focus on large datasets (Cheong, Spitale, & Gunes, 2023). However, most datasets in the mental health domain are small. Although there is much emphasis on the need for gender fairness in automatic prediction tasks for (mental) health, there is not much research evaluating the fairness of these models.

Bailey & Plumbley (2021) concluded that the demographic parity of depression was not satisfied in the DAIC-WOZ training and development data, because the number of depressed instances was not nearly equal for males and females. Wei et al. (2023) tested the gender fairness of different models on this dataset with the audio, visual, and text modalities. Their proposed models showed high overall performance but also gender biases.

2.3 Distress Analysis Interview Corpus and Related Work

The Distress Analysis Interview Corpus (DAIC) is a publicly available set of semi-structured clinical interviews, collected by Gratch et al. (2014). A subset of this dataset is the Wizard-of-Oz (WOZ) condition, which contains interviews conducted by an animated virtual interviewer called Ellie, controlled by a human interviewer in another room. Three modalities are included in the dataset: visual features collected with OpenFace (Baltrušaitis et al., 2016), which uses Conditional Local Neural Fields for facial landmark detection and tracking (Baltrusaitis et al., 2013); audio recordings with identifiable utterances scrubbed; and a text transcript. In this study, only the audio and text modality will be used.

The participants (189 in total) complete a series of questionnaires before the interview, including basic demographic questions and measures of psychological distress and current mood.

The Patient Health Questionnaire 8 (PHQ8; Kroenke et al., 2009) is used to measure psychological distress, which is a well-validated marker of depression disorder using the first eight MDD symptoms from DSM-V (American Psychiatric Association, 2013b). This questionnaire (included in Appendix A) contains eight questions asking the experienced severity (0 to 3) of a symptom. Unlike the binary perspective on symptom presence, this questionnaire provides a more nuanced view, with experiences symptom severity. The question about the 9th symptom (recurrent thoughts of death) was included in the PHQ9 (Kroenke et al., 2001) but omitted in the PHQ8 because responses to this item did not accurately reflect whether or not suicide risk is present, and PHQ8 and PHQ9 total scores were similar (Wu et al., 2020). When the total score on the PHQ8 (ranging from 0 to 24) is equal to or above 10, the participant is classified as depressed and otherwise as not depressed.

This dataset was used in the AVEC'2017 challenge for multimodal depression and affect recognition (Ringeval et al., 2017) and consists of three subsets: a training, development, and test set. The distributions of gender and depression classes are shown in Figure 2.2. In general, there are more nondepressed instances than depressed samples, and the gender balance in the training set is most disproportionate. There are more nondepressed males than nondepressed females, and more depressed females than depressed males.

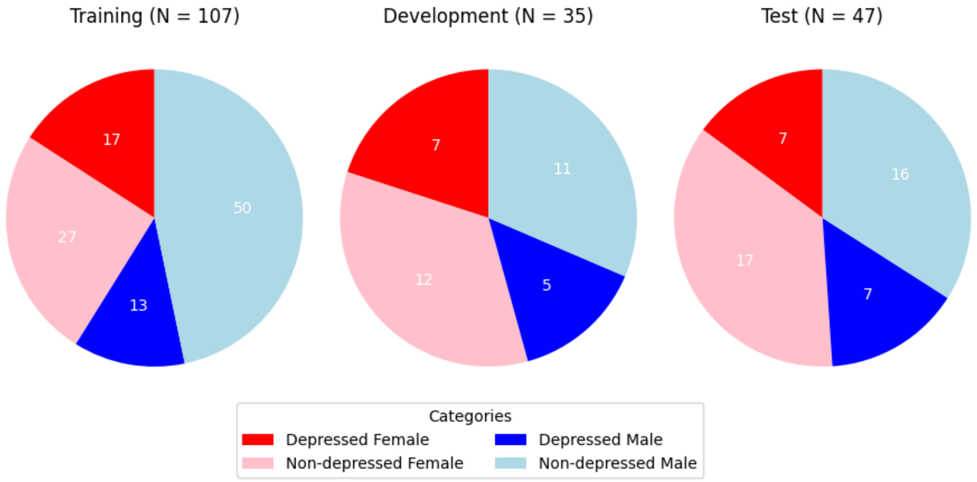


Figure 2.2: The gender-depression class imbalances in the DAIC-WOZ dataset. The sizes of the subclasses are shown in the charts.

The training and development sets are combined in this study for robustness. This combined set (142 participants) will be referred to as the validation set and is used to optimize models with N-fold Cross-Validation (CV; more details about this method in Section 4.2.2).

The gender-specific proportions of symptom severities in the validation set are shown in Figure 2.3. This illustrates that the distributions of symptom severities for males and females are relatively consistent over the eight symptoms of MDD. Generally, more males than females score lower (0 or 1) and more females than males score higher (2 or 3). Only PHQ8 Moving – a question referring to moving or speaking noticeably slow – shows a deviating distribution and is experienced by few.

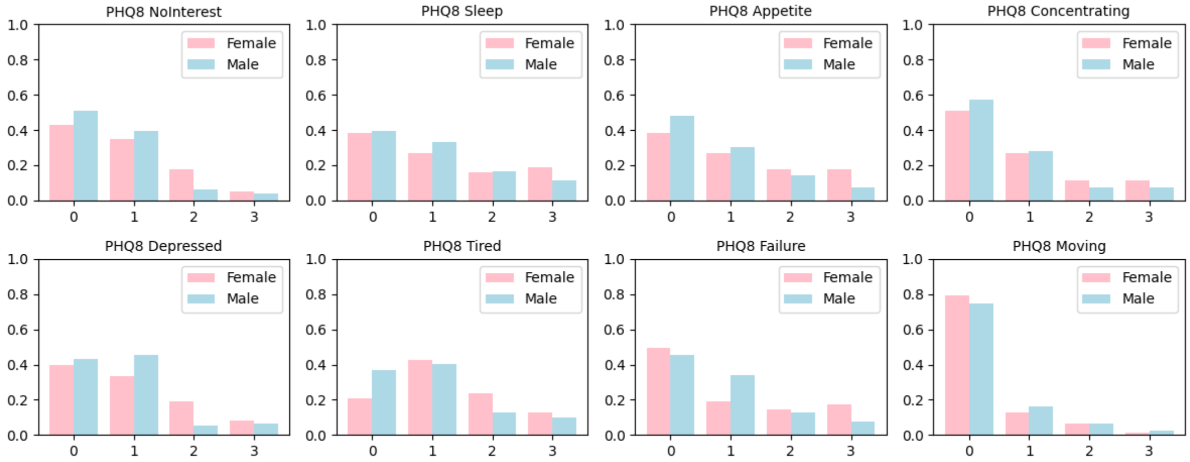


Figure 2.3: Symptom distributions (proportion) per gender in the DAIC-WOZ validation set.

The authors of the DAIC-WOZ dataset took some preprocessing steps: removing identifiable utterances from the audio recordings and converting transcriptions to lowercase sentences, ensuring ease of use for researchers. Therefore, this dataset has become an often-used benchmark for evaluating uni and multimodal models for predicting psychological distress. Some limitations are that the number of participants is relatively small compared to other datasets used to train ML models, and as mentioned before, the size imbalance in both gender and depression classes has led to biases in the models in the past (Bailey & Plumbley, 2021; Wei et al., 2023).

Several studies can highlight the significance of this dataset in multimodal depression recognition research. These studies represent just a fraction of the research conducted using this dataset. Muzammel et al. (2021) and Niu et al. (2021) developed a high-performing hierarchical context-aware graph attention model using the audio and text modality. Oureshi et al. (2021) discovered that gender information in the DAIC-WOZ can improve the performance of depression severity estimation and that the text modality is the best marker of depression. Guo et al. (2022) tested a late-fusion strategy to combine audio and text information, and they proved to diminish the overfitting issue caused by the small dataset. Despite the research that has been conducted in this area, a gap in the literature remains.

2.4 Research Gap

Summarizing the content above, there is a body of research in the domain of multimodal mental health detection in various research settings, but the exploration of fairness in this context has been limited, especially regarding gender. Therefore, there is an opportunity to enhance this research domain by integrating both audio and text modalities and evaluating gender fairness in a clinical setting, using the symptoms from the DSM. The DAIC-WOZ is a suitable clinical dataset with multimodal data, therefore this dataset will be in this study. Table 2.2 shows a selection of papers similar to this research, illustrating a gap in the research literature.

	Data		Modality					
	Clinical	DAIC-WOZ	Audio	Text	Fusion	Fairness	Gender	Symptoms
Oureshi et al. (2021)	X	X	X	X	X		X	
Bailey & Plumbley (2021)	X	X	X			X	X	
Booth et al. (2021)			X	X	X	X	X	
Muzammel et al. (2021)	X	X	X	X	X			
Guo et al. (2022)	X	X	X	X	X			
Niu et al. (2021)	X	X	X	X	X			
Van Steijn et al. (2022)	X	X		X			X	X
Yoon et al. (2022)*		X	X		X	X	X	
Cheong, Kuzucu, et al. (2023)	X		X		X	X	X	
Cheong, Spitale, & Gunes (2023)	X		X	X	X	X		
Sogancioglu et al. (2023)	X			X		X		X
Ma et al. (2016)	X	X	X			X		
Cohen et al. (2023)	X		X	X	X			
Milintsevich et al. (2023)	X	X		X				X
Current	X	X	X	X	X	X	X	X

* Primary dataset is non-clinical, authors perform cross-corpus validation with DAIC-WOZ

Table 2.2: An overview of related work. The columns indicate research directions: *fusion* refers to the study exploring different fusion methods for the multimodal data, *fairness* indicates an exploration of the impact of biases. *Gender* signifies a specific focus on gender differences, and *symptoms* indicates an analysis focused on mental health symptoms.

3 Background on Computational Approaches

This section introduces the technical details of some relevant computational methods for this study. Text processing, speech processing, and ways to combine multimodal data will be discussed. After this, details about the predictive models used in this study will be provided, as well as ways to measure fairness and mitigate biases.

3.1 Multimodal Data Processing

The background section illustrated that certain behaviors can indicate a healthy or unhealthy mental state. By analyzing textual and audio data, we can extract valuable insights and patterns that help to understand these indicators. This section discusses natural language processing, speech processing, and the methods to combine data from different modalities.

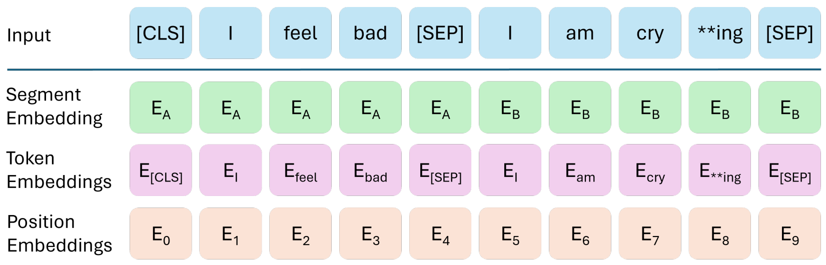
3.1.1 Natural Language Processing

Language can be represented computationally in different manners, words and sentences can for example be embedded with or without context.

Word and Sentence Embeddings

Non-contextual word embeddings like Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) capture semantic relationships between words by representing them in a continuous vector space (Miaschi & Dell'Orletta, 2020). Each word is a high-dimensional vector and a low distance between two vectors indicates two words with similar meanings. The same direction between two pairs of word embeddings indicates a similar relationship.

On the other hand, contextual word embeddings, take into account the surrounding context of a word, which enables capturing nuances in meaning that may vary depending on the context in which a word is used. Figure 3.1 illustrates the embedding of the words 'I feel bad, I am crying' in a sentence. The sentence is split into segments (A and B), each word is divided into tokens (the base form and pre- or suffixes), and the input embedding is the sum of the segmentation, token, and position embeddings.



Input	[CLS]	I	feel	bad	[SEP]	I	am	cry	**ing	[SEP]
Segment Embedding	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
Token Embeddings	$E_{[CLS]}$	E_I	E_{feel}	E_{bad}	$E_{[SEP]}$	E_I	E_{am}	E_{cry}	E_{**ing}	$E_{[SEP]}$
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9

Figure 3.1: Illustration of a sentence embedding.

These contextual word embeddings (or sentence embeddings) can be fed to transformer models: neural networks that can capture longer-range relationships.

Transformer Models

Transformer models can process sentence embeddings, producing outputs that vary based on the specific task to which they are applied. Figure 3.2 shows a transformer module for sentence classification. These modules are used by architectures such as Bidirectional Encoder Representations from Transformers (BERT; Devlin et al., 2019) and Generative Pre-trained Transformers (GPT; Radford et al., 2018). Sentence BERT (SBERT) computes sentence embeddings with sentence transformer models, giving as output numerical representations of sentences with relative meaning. The sentence 'I feel bad', will be more similar to 'I did something wrong' than 'Yesterday I ate pie'.

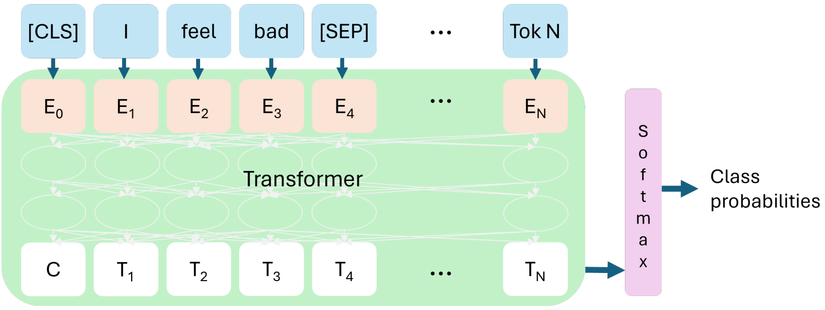


Figure 3.2: Illustration of a transformer module for text classification.

SBERT has a pre-training and fine-tuning stage. The pre-training stage involves two self-supervised tasks. The initial pre-training step entails the random masking of a percentage of input tokens, and the model is trained to predict these masked tokens. This step facilitates the learning of relatively short-range relationships between words. In the second pre-training step, BERT uses next-sentence prediction (NSP), enabling the model to learn sentence relationships. For each pre-training example, sentences A and B are selected, with a 50% chance that B is the actual next sentence following A and a 50% that it is a random sentence from the corpus. This strategy enables the model to grasp longer-range relationships.

During fine-tuning, task-specific token representations generated by BERT can be plugged into the transformer module, and parameters are fine-tuned. Using this process, mental health-related data can for example be analyzed with BERT (Xiao et al., 2021; Fan et al., 2019). In addition to textual data, audio (speech) data can be used to detect depression.

3.1.2 Speech Processing

The processing of audio information entails a few steps: first, an audio recording is divided into shorter chunks. Depending on the task, the file is split into words, 'turns' (based on speech onset until the offset of one speaker in a conversation), time slices with a constant length, or proportions of longer units.

The most important step in automated recognition of speaker states, traits, and vocal behavior is extracting relevant features. In the past, prosodic features (e.g., pitch, duration, and intensity) were more frequently used than voice quality features (e.g., harmonics-to-noise ratio, jitter, shimmer) (Schuller, 2011). Segmental, spectral features modeling formants, or cepstral features are often found in the literature (Ma et al., 2016; Bailey & Plumbley, 2021; Yamamoto et al., 2020).

State-of-the-art networks use audio feature extraction methods such as wav2vec2.0 (Baevski et al., 2020) and openSMILE (open Speech and Music Interpretation by Large-space Extraction) (Eyben et al., 2010). Wav2vec2.0 is a framework that is trained to predict the correct speech unit for masked parts of the audio file. The pre-trained model can then be fine-tuned on a specific dataset and for a specific task. This has been useful for speech recognition of various languages and dialects (Sharma, 2022).

Selecting a subset of features can improve reliability and (speed and memory) performance. Additionally, feature reduction reduces the complexity and number of free parameters to be learned by ML algorithms.

Public Dimension Emotional Model

The Public Dimension Emotional Model (PDEM) was developed by Wagner et al. (2023). This audio model is intended for speech emotion recognition and predicts values for emotional dimensions such as arousal, dominance, and valence. Arousal measures the intensity of an emotion, dominance measures the level of control or assertiveness in the emotion and valence measures the positivity or negativity of the emotion.

PDEM uses two pre-trained transformer-based models, the earlier mentioned wav2vec2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021). Both are designed for processing raw waveform audio input. The architecture of the PDEM is shown in Figure 3.3.

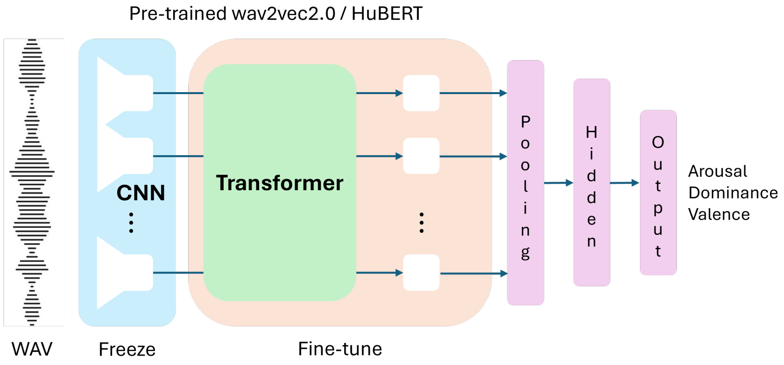


Figure 3.3: Illustration of the PDEM architecture, adapted from Wagner et al. (2023).

Not only the three output scores for arousal, dominance, and valence can be used, but the 1024 hidden units can serve as an embedding for an audio chunk, thereby embodying the emotional information. This model has been shown to be fair with respect to gender groups (Wagner et al., 2023). Because of the expected importance of emotional information in mental health detection, the PDEM will be used to extract audio embeddings in this study.

After preprocessing the data and determining the relevant text and audio features, the input for classification or regression is prepared. These input features can be used directly in sentence-level audio models or the feature vector can be weighted with an attention vector, emphasizing certain parts of the feature.

Attention Mechanism

Attention mechanisms focus on specific parts of the input and emphasize those that are more important for the task at hand. A systematic overview (Galassi et al., 2021) on attention models describes that the attention mechanism is part of a neural architecture that can be applied to raw input or higher-level representation, thereby computing a weight distribution on the input with higher values to more relevant elements. The attention mechanism can be applied to audio and text data, giving more weight to relevant information.

Whether using attention or not, the embeddings from audio or text can be used separately to predict depression (unimodal), or alternatively, the audio and text information can be combined in a multimodal model.

3.1.3 Combining Multimodal Data

As discussed in the introduction, multiple modalities (e.g., audio, visual, and textual) can predict an individual's mental well-being. Modalities can be combined or fused according to different strategies. State-of-the-art fusion approaches for depression recognition can be categorized as model-agnostic or model-based (Muzammel et al., 2021).

Model-Agnostic Fusion Strategies

Model-agnostic fusion strategies can be used with any model. Three methods can be identified, differing in when they are applied in the process: early, late, and hybrid fusion.

Early fusion, also known as feature-level fusion is a commonly used fusion strategy (Fang et al., 2023; Booth et al., 2021; Cohen et al., 2023). Features in ML context are informative, predictive characteristics of input data or signals that can be extracted from different modalities. Early fusion combines two or more feature vectors (where each vector may represent a single modality) into a single high-dimensional feature vector, this high-dimensional feature vector can then be used as input for any model.

Late fusion is also called decision-level fusion. First, multiple base models (usually unimodal) are used to make predictions, and then an algebraic combination rule (e.g., majority voting, averaging, or stacking) is used to integrate these predictions. Majority voting uses the most frequently predicted label of the base models. Alternatively, the predictions can be averaged in a weighted or unweighted manner, and then predict the average. In stacking, the final decision is made by another classifier that uses as input either the predicted labels or the predicted class probabilities of each base model (Wolpert, 1992).

Hybrid fusion uses both early and late fusion to combine individual classification scores of uni or multimodal models. For example, Yang, Sahli, et al. (2017) fused the predictions of audiovisual and text-based models: two audiovisual models with early fusion were trained separately to predict depression, and these predictions were then combined with the predictions from two text-based models.

Model-Level Fusion Strategies

The three fusion strategies above can be applied to any model. Model-level fusion strategies are model-dependent and try to learn a joint representation of different modalities after unimodal feature extraction and concatenation, or directly with the raw signal. Recent related work has for example used multiple kernel learning to find an optimal combination of input modalities' features for emotion recognition with visual and audio features (Chen et al., 2014), or neural networks-based approaches to detect mental health illnesses (Al Hanai et al., 2018; Yang, Jiang, et al., 2017; S. Zhang et al., 2017).

Now we have discussed the input for models to predict mental health. The next topic will be the specific architecture used in this study to map input (audio or textual information) to output (depression severity).

3.2 Kernel Extreme Learning Machine

Kernel Extreme Learning Machine (KELM) is an advanced version of an Extreme Learning Machine (ELM), known for fast and efficient learning. An ELM is a single hidden layer feed-forward neural network where the input weights and biases are first randomly assigned, and the output weights are optimized (Huang et al., 2004). KELM enhances ELM by using kernel methods: mapping input features into high-dimensional features allowing it to handle non-linear data effectively.

Kernel methods operate by computing the inner products between all pairs of data points in the feature space, represented by the kernel matrix. This approach enables KELM to predict both classification and regression labels efficiently (Huang et al., 2011). Commonly used kernel functions are linear, polynomial, Gaussian (RBF), and sigmoid kernels.

KELM's efficacy has been demonstrated in paralinguistic and affective computing tasks (Kaya et al., 2017; Gurpinar et al., 2016; Kaya et al., 2019). This has shown that KELM is useful in contexts where data is sparse and high-dimensional. KELM offers a balance of speed and accuracy, unlike Support Vector Machines (SVMs), which can be slow, and other neural networks that may require extensive training on smaller datasets.

Given a training dataset $\mathcal{D} \in \mathbb{R}^{n \times m}$ with n instances and m features, KELM solves a regularized least squares regression problem. The relationship between the kernel matrix $\mathcal{K} \in \mathbb{R}^{n \times n}$ and the target variable/matrix $\mathcal{T}$ is expressed as:

$$\beta = \left(\frac{I}{C} + \mathcal{K} \right)^{-1} \mathcal{T}$$

where C is the regularization coefficient, and I is an $n \times n$ identity matrix. For a test instance $\mathbf{x}$, the prediction $\hat{y}$ is obtained using:

$$\hat{y} = \phi(\mathcal{D}, \mathbf{x})\beta$$

where ϕ denotes the kernel function. $\mathcal{T}$ can be extended for multi-task regression or classification, where each column represents a different task (such as predicting multiple symptom severities). The multi-task regression will be used in the current study.

In classification tasks, categories are often one-hot encoded, and an ordinal encoding can be used for ordinal classification (Shi et al., 2019). To handle class imbalance in classification tasks – which is common in mental health datasets – weights can be included in the model. By integrating weights, KELM can mitigate the impact of class imbalance, ensuring the model does not favor the majority class excessively (Zong et al., 2013). Although the weights will not be used in this study, the efficiency and flexibility of KELM in terms of regression/classification, weights, and parameters makes it suitable for complex mental health detection tasks. With the input data and the model, the predictions of the model can be evaluated on overall performance and on fairness.

3.3 Measuring Fairness in Predictive Modeling

Various metrics are used to quantify fairness and evaluate biases in ML models. This section discusses one fairness metric based on a continuous prediction (e.g., depression severity), and several fairness metrics based on binary predictions.

Continuous metric: Equal Accuracy / RMSE Ratio

As AVEC'2017 (Ringeval et al., 2017) used the root mean squared error (RMSE) as a baseline accuracy measure in depression severity prediction, we will take the accuracy performance ratio of different classes (in this case gender) as a fairness measure. Let $\hat{\mathrm{S}}_i$ be the predicted depression severity score and S_i be the true depression severity score for participant i in a subset of the data with size n. The overall RMSE of the predictions in this subset can be calculated as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mathbf{S}}_i - \mathbf{S}_i)^2} \tag{3}$$

Let $G \in \{m, f\}$ be the sensitive attribute (gender), RMSE_f be the RMSE for females, and RMSE_m be the RMSE for males. For a classifier to be deemed perfectly fair according to equal accuracy (EqAcc) $\text{RMSE}_f = \text{RMSE}_m$ and therefore:

$$M_{\text{EqAcc}} = \frac{\text{RMSE}_f}{\text{RMSE}_m} \tag{4}$$

If $M_{\text{EqAcc}} > 1$, then $\text{RMSE}_f > \text{RMSE}_m$. This indicates a bias favored towards males, as a lower RMSE signifies better predictive accuracy. If $M_{\text{EqAcc}} < 1$, the predictive system is biased in favor of females.

Basic terms that lie at the root of binary classification metrics are presented in Table 3.1. These terms can be used to calculate Equations 5 - 9. Ratios of some of these performance measures across groups can again be used to evaluate fairness.

		Actual label	
		1	0
Predicted label	1 (positive)	True Positive (TP)	False Positive (FP)
	0 (negative)	False Negative (FN)	True Negative (TN)

Table 3.1: Confusion matrix for binary classification with performance metrics.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{5}$$

$$\text{Sensitivity / True Positive Rate (TPR) / Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{6}$$

$$\text{Precision / Positive Predictive Value} = \frac{\text{TP}}{\text{TP+FP}} \tag{7}$$

$$\text{False Positive Rate (FPR)} = \frac{\text{FP}}{\text{TN+FP}} \tag{8}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

The most common binary group fairness measures will now be discussed through an example: a depression detection system predicting the presence of depression (prediction $\hat{Y} = 1$) or the absence of depression (prediction $\hat{Y} = 0$) for patients that either have depression (outcome $Y = 1$) or not (outcome $Y = 0$).

Binary metric: Statistical / Demographic Parity

Demographic parity means that the likelihood of a positive outcome is the same, regardless of group. If the depression detection system adheres to the demographic parity measure for gender, an equal proportion of men and women would be predicted to have depression. For a classifier to be deemed fair according to this measure $P[\hat{Y} = 1|G = f] = P[\hat{Y} = 1|G = m]$ and therefore:

$$M_{\text{SP}} = \frac{P[\hat{Y} = 1|G = f]}{P[\hat{Y} = 1|G = m]} \tag{10}$$

A limitation of this fairness measure is that it does not consider the actual outcome (whether someone has depression, or what the exact prevalence of depression is in different groups). The classifier can make random predictions and still adhere to this fairness measure, therefore this measure will not be used in this study.

Binary metric: Equal Opportunity / True Positive Rate Ratio

Equal opportunity (EqOpp) ensures that people with the same (positive) outcome – depressed people – are treated the same. This is similar to the recall or TPR (Equation 6) being equal for each group of a given sensitive attribute. If the predictive system adheres to the equal opportunity measure, everyone who has depression has the same likelihood of being predicted as depressed (regardless of gender). For a classifier to be deemed fair according to equal opportunity $P[\hat{Y} = 1|Y = 1, G = f] = P[\hat{Y} = 1|Y = 1, G = m]$ and therefore:

$$M_{\text{EqOpp}} = \frac{P[\hat{Y} = 1|Y = 1, G = f]}{P[\hat{Y} = 1|Y = 1, G = m]} \tag{11}$$

If $M_{\text{EqOpp}} > 1$, the predictive system is biased in favor of females, if $M_{\text{EOpp}} < 1$, the predictive system is biased in favor of males.

Binary metric: Predictive Equality / False Positive Rate Ratio

Predictive Equality (PredEq) also ensures that people with the same outcome are treated the same. Regardless of gender, nondepressed people should have the same likelihood of receiving an incorrect label (preferably low). This is similar to the FPR (Equation 8) being equal for each group of a given sensitive attribute. For a classifier to be deemed fair according to predictive equality $P[\hat{Y} = 1|Y = 0, G = f] = P[\hat{Y} = 1|Y = 0, G = m]$ and therefore:

$$M_{\text{PredEq}} = \frac{P[\hat{Y} = 1|Y = 0, G = f]}{P[\hat{Y} = 1|Y = 0, G = m]} \tag{12}$$

If $M_{\text{PredEq}} > 1$, the predictive system is biased in favor of males, if $M_{\text{PredEq}} < 1$, the predictive system is biased in favor of females. Instead of insisting on equal scores for all the above fairness metrics, it is possible to relax this constraint with the principle of disparate impact (Feldman et al., 2015). The ideal score of 1 indicates a perfectly fair system, but for practical experimental purposes, this study will follow existing literature that considers 0.8 and 1.2 as acceptable lower and upper fairness bounds respectively (Cheong, Kuzucu, et al., 2023; Zanna et al., 2022; Park et al., 2022).

In the mental healthcare literature, different measures have been used to assess the fairness of binary systems: F1-scores per gender (and depression) subgroup (Bailey & Plumbley, 2021); statistical parity, equal opportunity, equal accuracy, and equalized odds (Cheong, Kuzucu, et al., 2023; Park et al., 2022); equal accuracy and disparate impact (ratio of the probability of a favorable outcome for the unprivileged group to that of the privileged group) (Cheong, Spitale, & Gunes, 2023); macro-averaged F1-score, equal opportunity, predictive equality, and mismatch ratio (Sogancioglu et al., 2023). In this study, the most often used fairness measures will be reported: equal accuracy, equal opportunity, and predictive equality. When a fairness metric is above or below the fairness bounds, the classifier shows a bias in favor of or against a group.

3.4 Bias Mitigation Methods

Mitigating biases can make predictive systems more fair. There is, however, often a potential trade-off between maximizing overall accuracy and ensuring fairness (Wang et al., 2021; Dutta et al., 2020). Specific algorithms, (hyper)parameters, loss functions, and thresholds can all contribute to this trade-off. Nevertheless, methods to mitigate biases can be divided into pre-processing, in-processing, and post-processing methods.

Pre-processing methods

Pre-processing methods focus on the first stage in the training process, changing or adjusting the data to remove bias to ensure fairer data to obtain a fairer model (Cheong, Spitale, & Gunes, 2023). Two examples of pre-processing to ensure fair data are data augmentation and reweighing. Reweighing can sometimes be applied directly when the classifier works with weights, otherwise the data can be resampled to mimic weights.

Balancing the data is a data augmentation method where the number of samples from minority groups is increased to match the number of samples from majority groups. This can be done by re-using samples from the minority groups (same as reweighting), or by generating synthetic data. The MixFeat approach (Cheong, Spitale, & Gunes, 2023), based on the MixUp method (H. Zhang et al., 2017) was used to balance data from the DAIC-WOZ using interpolation to generate data points. Given a dataset of size N, where v^A represents an audio embedding and v^T represents a text embedding, the new training sample $(\mathrm{v}_{o*}, \mathrm{v}_{o*})$ for a minority group can be generated as follows:

$$\begin{aligned} \mathrm{v}^A_{o*} &= \lambda_A \cdot \mathrm{v}^A_o + (1 - \lambda_A) \cdot \mathrm{v}^A_{o^+} \\ \mathrm{v}^T_{o*} &= \lambda_T \cdot \mathrm{v}^T_o + (1 - \lambda_T) \cdot \mathrm{v}^T_{o^+} \end{aligned} \tag{13}$$

where $o, o^+ \in \{1, ..., N\}, o \neq o^+$ and $\lambda_A, \lambda_T \sim$ Beta(0,1).

Another way to augment the data, e.g., when there is an imbalance in the number of instances per class, is to split the interview into subsets of sentences and create multiple (sub)session-level embeddings of a single subject from a minority class. Let V_i represent the set of sentence embeddings for participant i in a minority class. To balance the subclasses, we might require three times more data samples in this class. The set $\mathrm{V}_{i1} = \{\mathrm{v}_{io} \mid o \in \{1, \ldots, n_i/3\}\}$ can be used as the first part of the interview, $\mathrm{V}_{i2} = \{\mathrm{v}_{io} \mid o \in \{n_i/3 + 1, \ldots, 2n_i/3\}\}$ as the second part, and $\mathrm{V}_{i3} = \{\mathrm{v}_{io} \mid o \in \{2n_i/3 + 1, \ldots, n_i\}\}$ as the last section of the interview. With these sets of embeddings and the approach in Section 4.2.1, three new session-level feature vectors can be computed and the classes can be balanced.

In-processing methods

In-processing methods mitigate biases during the model training process for example via regularization and constraints, or adversarial learning. A (weighted) regularization term can be added to make the classifier fairer, imposing a cost on the optimization function (Kamishima et al., 2012). Adversarial debiasing is used when there are proxy features for the sensitive feature (e.g., due to historical biases or systemic discrimination), which might result in the model learning these biases. Generative adversarial networks have two parallel goals: predict the outcome (goal of the predictor module) and predict the protected attribute from the output (goal of the adversary module). The adversarial module penalizes biased or unfair predictions and the system will learn to maximize the predictive ability of the predictor while minimizing the predicting ability of the adversarial (B. H. Zhang et al., 2018). This method will not be explored in the current study.

Post-processing methods

Post-processing methods are applied after model training, acting on the model outcomes (Hort et al., 2024). These methods change the predictions of the models, thereby directly controlling the outcomes. Post-processing methods in a classification task can use different decision thresholds for different groups, to ensure for example equal opportunity (Hardt et al., 2016), or equalized odds (Awasthi et al., 2020). This method will also not be explored in the current study.

4 Methodology

This section describes the approach to answering the research questions. First, the research questions will be presented, then the general procedure will be outlined with the data, models, and evaluation of the models.

4.1 Research Questions

The background section identified a gap in the literature (see Section 2.4 and Table 2.2). To summarize: gender bias in the literature on clinical depression detection with both audio and text data has not yet been sufficiently analyzed. The main research questions (RQ) and research subquestions (RSQ) in this study are:

RQ 1: *What is the performance of unimodal audio models, unimodal text models, and bimodal audio-text models when predicting clinically diagnosed depression severity using the DAIC-WOZ corpus – as measured by RMSE and MAE?*

> **RSQ 1A:** What is the effect of adding an emotional attention component to the unimodal models on the depression severity prediction using the DAIC-WOZ corpus – as measured by RMSE and MAE?
>
> **RSQ 1B:** What is the effect of adding an emotional filtering component to the unimodal audio model on the depression severity prediction using the DAIC-WOZ corpus – as measured by RMSE and MAE?

RQ 2: *What is the gender fairness of the uni and bimodal models from RQ 1 in depression severity prediction and binary depression detection using the DAIC-WOZ corpus – as measured by EqAcc, EqOpp, and PredEq?*

RQ 3: *What is the effect of three pre-processing bias mitigation methods (resampling, interview split, and MixFeat) on gender fairness using the best models from RQ 1 for depression severity prediction and binary depression detection using the DAIC-WOZ corpus– as measured by EqAcc, EqOpp, and PredEq?*

4.2 Proposed Approach

This research will be conducted in a `python` environment. To answer the research questions, the data will be preprocessed, and the models will be optimized and evaluated based on performance and fairness metrics. All the code needed to replicate this study can be found in a GitHub repository[2]. A pipeline of the proposed approach to answer the research questions can be found in Figure 4.1.

[2] `https://github.com/MartKoek/master-thesis`

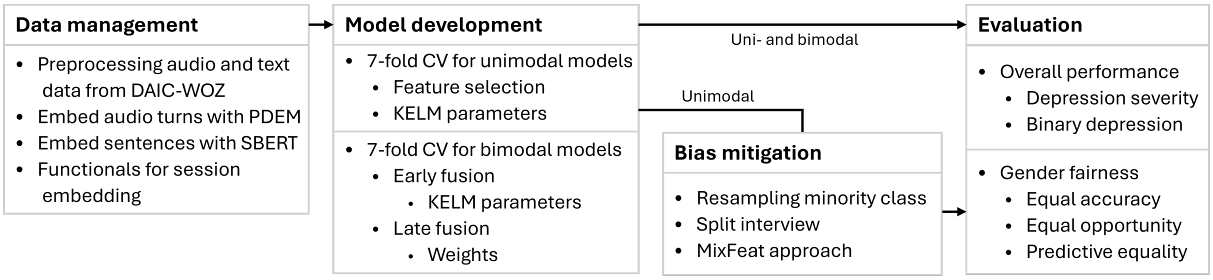


Figure 4.1: Pipeline of the proposed approach.

First, the data management steps will be discussed, including preprocessing and the generation of session-level embeddings from sentence-level embeddings with functionals. Then the specific uni- and bimodal models using KELM will be outlined. Additionally, the process of optimizing and evaluating these models will be explained.

4.2.1 Data Management

Each interview contains subject turns and Ellie turns (the avatar interviewer). As Ellie's turns are relatively constant over all the interviews, they are not informative of the depressed state of the participants, therefore they are filtered out. Additionally, only turns with a length over 500 ms are retained, as shorter turns might predominantly add noise. After preprocessing, the average number of turns per subject in the validation set is 150.49. Gender-specific information is presented in Table 4.1. This shows that the average number of turns in a conversation and the average number of words per turn are slightly higher for females.

Gender (size)	Avg. No. turns in conversation	Avg. No. words per turn	Avg. PHQ8 severity
Female (63)	153.52	9.53	7.51
Male (79)	148.06	9.09	6.00

Table 4.1: Data statistics per gender (averages) in the validation set.

Using SBERT, each sentence from the interview transcription is encoded into a text embedding. Each session-level .wav file with audio data is split into chunks with the start and stop times of each turn (included in the transcript file). Using the PDEM architecture, each sentence-level audio chuck is converted into an audio embedding and scored on the emotional dimensions valence, arousal, and dominance (VAD). Therefore, each turn is translated into an SBERT embedding, PDEM embedding, and VAD vector.

An often-used method to summarize sentence-level embeddings into session-level data is to calculate functionals over all sentences in an interview for each feature value. Let $n = 142$ be the number of instances in the validation dataset, let m be the length of an embedding vector, with $m_{\text{SBERT}} = 384$ as the length of each SBERT vector, and $m_{\text{PDEM}} = 1024$ as the length of each

PDEM embedding. Let n_i be the number of turns (embeddings) for participant i. Let f be the number of functionals applied to the embeddings of each participant, namely: mean, median, standard deviation (sd), var, max, q75, thus $f = 6$.

Let V_i be the set of embeddings for participant i, with $\mathrm{V}_i = \{\mathrm{v}_{i1}, \dots \mathrm{v}_{io}, \dots, \mathrm{v}_{in_i}\}$ with $o \in \{1, ..., n_i\}$. For each embedding vector v_{io}: Let $\mathrm{v}_{io}[l]$ be the l-th value of the o-th vector for participant i with $l \in \{1, 2, \dots, m\}$. The functionals over the n_i embeddings for each embedding value l can then be calculated:

$$\mathrm{mean}_l = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathrm{v}_{io}[l] \tag{14}$$

$$\mathrm{median}_l = \mathrm{median}(\mathrm{v}_{i1}[l], \mathrm{v}_{i2}[l], \dots, \mathrm{v}_{in_i}[l]) \tag{15}$$

$$\mathrm{var}_l = \frac{1}{n_i}\sum_{j=1}^{n_i} (\mathrm{v}_{io}[l] - \mathrm{mean}_l)^2 \tag{16}$$

$$\mathrm{sd}_l = \sqrt{\mathrm{var}_l} \tag{17}$$

$$\mathrm{max}_l = \mathrm{max}(\mathrm{v}_{i1}[l], \mathrm{v}_{i2}[l], \dots, \mathrm{v}_{in_i}[l]) \tag{18}$$

$$\mathrm{q75}_l = \text{75th percentile of } (\mathrm{v}_{i1}[l], \mathrm{v}_{i2}[l], \dots, \mathrm{v}_{in_i}[l]) \tag{19}$$

Thus, for each participant i, a session-level feature vector $\mathrm{F}_i \in \mathbb{R}^{f \times m}$, is constructed, where:

$\mathrm{F}_i = (\mathrm{mean}_1, \mathrm{median}_1, \dots, \mathrm{q75}_1, \dots \mathrm{mean}_m, \mathrm{median}_m \dots \mathrm{q75}_m)$

This results in a feature vector $\mathrm{F}_{\mathrm{SBERT}_i}$ for each participant with a total length of $f \times n = 6 \times 384 = 2304$ values and a $\mathrm{F}_{\mathrm{PDEM}_i}$ with a total length of $f \times n = 6 \times 1024 = 6144$ values. The most informative functionals were selected during the 7-fold CV stage. All possible combinations of functionals were tested within the feature vectors, and the functionals from the best-performing model were used in subsequent experiments on the test set.

4.2.2 Depression Severity Models

This section discusses the prediction of depression severity scores and binary depression scores.

The regression KELM (explained in Section 3.2) was trained on the session-level PDEM embeddings $\mathrm{F}_{\mathrm{PDEM}}$ and SBERT embeddings $\mathrm{F}_{\mathrm{SBERT}}$ in the validation set to predict the symptom vector of a participant. Let F_i be the session-level embedding for participant i and let $\mathrm{y}_i = (y_{i1}, y_{i2}, \dots, y_{i8})$ be the symptom vector for participant i, corresponding to the eight symptom scores on the PHQ-8 questionnaire with $y_{ij} \in \{0, 1, 2, 3\}$. As we predict the depression severity and sum the symptom severities, we chose to use regression of individual symptoms and not ordinal classification of symptom severities, because the exact severity of each symptom is less important. The KELM model is trained to predict $\hat{\mathrm{y}}_i$ from F_i:

$$\hat{\mathrm{y}}_i = \phi(\mathcal{D}, \mathrm{F}_i)\beta \tag{20}$$

The predicted values are then constrained within the range of 0 to 3, ensuring that any values below 0 are set to 0 and above 3 are set to 3: $\hat{y}_{ij} = \max(0, \min(3, \hat{y}_{ij}))$ for $j \in \{1, 2, \ldots, 8\}$. Then, the individual symptom scores are aggregated and the resulting sum is rounded to the nearest integer to obtain the overall depression severity score $\hat{\mathrm{S}}_i$ for participant i:

$$\hat{\mathrm{S}}_i = \left\lfloor 0.5 + \sum_{j=1}^{8} \hat{y}_{ij} \right\rfloor \tag{21}$$

where $\hat{y}_i = (\hat{y}_{i1}, \hat{y}_{i2}, \ldots, \hat{y}_{i8})$ are the constrained predicted scores for participant i and $\lfloor \cdot \rfloor$ denotes the floor function. This approach is interpretable, allowing for the identification of individual symptom severities that the subject is likely experiencing.

Uni- and Bimodal Models

We construct four unimodal and five bimodal models to predict depression severity from audio or text embeddings. With the architecture in Figure 4.2, four unimodal models and one bimodal model can be constructed. The optional VAD component indicates the use of emotional attention or emotional filtering. Without this component, we have two plain unimodal models: an audio model with PDEM embeddings (model 1: A) and a text model with SBERT embeddings (model 2: T).

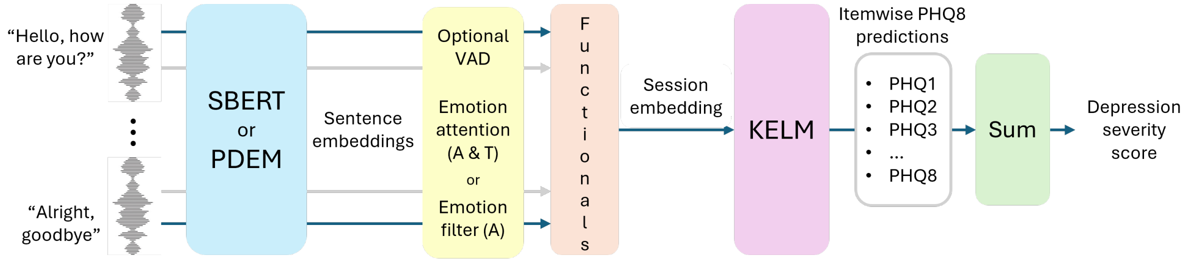


Figure 4.2: Unimodal and bimodal architecture to predict depression severity. The optional VAD component indicates an optional emotional attention or filtering layer using the VAD values from PDEM. Attention is applied to both the audio and text modality, filtering is only applied to the audio modality.

Emotional filtering (EF) entails using the top X% emotional turns in an interview. The distance between the VAD value of a sentence and the mean VAD value in that interview is calculated. All sentences are then ranked based on this distance and the top emotional part of the interview is used to compute an emotional session-level embedding. The percentage and emotional dimension (valence, arousal, or dominance) are optimized using N-fold CV. Emotional filtering is applied exclusively to the PDEM embeddings, resulting in an unimodal audio model with emotional filtering (model 3: AEF).

Emotional attention (EA) is applied by multiplying the sentence embeddings by the raw original or absolute VAD values, thereby increasing the weight of more emotional sentences before computing the session-level embeddings with functionals. As the emotional attention stems from the audio recordings, applying this to the PDEM embeddings results in an unimodal audio model with emotional attention (model 4: AEA). Applying emotional attention to the SBERT embeddings results in a bimodal audio-text model with model-level fusion (model 5: TEA).

In addition to the model-level fusion method mentioned above, two model-agnostic fusion methods are explored. Feature-level fusion entails concatenating the session-level feature vectors of PDEM and SBERT into a single fused embedding (see Figure 4.3). This audio-text embedding is then used as input for the KELM, resulting in a bimodal audio-text model with feature fusion (model 6: FF).

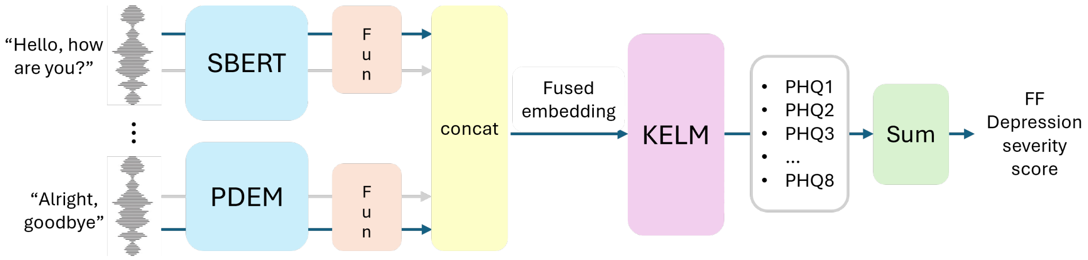


Figure 4.3: Bimodal architecture using feature fusion to predict depression severity.

The architecture to construct three bimodal decision fusion models is shown in Figure 4.4. Unweighted decision fusion (DF) combines the predictions of unimodal models from SBERT and PDEM embeddings (model 7: DF). Weighted DF (WDF) involves computing a weighted average of the unimodal predictions, where the optimal weights for the audio and text predictions are decided with N-fold cross-validation (model 8: WDF). Gender WDF (GWDF) uses gender-specific audio and text weights to weigh the predictions of the unimodal models (model 9: GWDF).

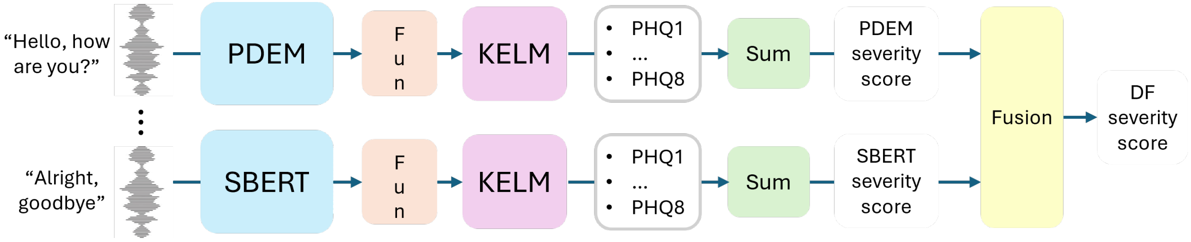


Figure 4.4: Bimodal architecture using decision fusion to predict depression severity.

Evaluating Performance

This study measures performance to predict depression severity with two key metrics, RMSE and mean absolute error (MAE), calculated as the discrepancy between predicted and reported PHQ-8 scores averaged over all sequences of a subset of data. These metrics were also used in the AVEC'2017 challenge (Ringeval et al., 2017), which established the baseline performance for this dataset. With $\hat{\mathrm{S}}_i$ being the predicted depression severity score and S_i being the true depression severity score for participant i for a subset of the data with size n MAE can be calculated:

$$\mathrm{MAE} = \frac{1}{n} \sum_{i=1}^{n} |\hat{\mathrm{S}}_i - \mathrm{S}_i| \tag{22}$$

Using these the RMSE and MAE, we can optimize the parameters of the models for depression severity prediction. From the predicted depression severity score $\hat{\mathrm{S}}_i$, we can derive the predicted binary depression score $\hat{Y}_i$. According to the PHQ8 method, instances with a severity score of 10 or higher are classified as depressed (label = 1) and otherwise as not depressed (label = 0):

$$\hat{Y}_i = \begin{cases} 1 & \text{if } S_i \geq 10 \\ 0 & \text{if } S_i < 10 \end{cases} \tag{23}$$

We do not train KELM directly to predict the binary depression score to maintain the model's interpretability. By predicting individual symptom severities first, we can understand why the model classifies an instance as depressed or not based on the predicted symptoms. But with the binary depression prediction, we can still evaluate the model on binary fairness measures (discussed in Section 3.3).

Stratified 7-fold Cross Validation

N-fold cross validation (CV) is a statistical technique used to assess a model's performance. In this study, the validation set was split semi-randomly into 7 folds with 20-21 instances. Details on the exact split (for replicating the results) can be found in the GitHub repository. Given that the validation set is small and the data is imbalanced, it is important to keep the gender and depression class balance roughly equal over the folds, this is called stratification. Stratified N-fold CV ensures that each fold represents each class proportionally, making the evaluation metrics more reliable.

In each fold, one part is used once as a temporary test set while the other parts are used as the temporary training set. After the 7 iterations, the 7 temporary test set predictions are concatenated and the overall performance (RMSE) is calculated. This performance is compared to the overall performance of the model with different (hyper)parameters. For each of the 9 models, the settings of the best-performing model are then used to retrain the model on the entire validation set and the model is tested on the original test set.

5 Experimental Results

In this section, the findings of this study on multimodal depression prediction and gender fairness will be presented. First, the performance of the optimized models on the validation set will be given. Then, the performance of models on depression severity prediction and binary depression detection on the test set will be presented and compared to state-of-the-art research. A fairness analysis will follow this and lastly, the effect of bias mitigation methods will be discussed.

5.1 Experimental Results on the Validation Set

With 7-fold CV, (hyper)parameters for KELM were optimized. Three different SBERT embedders were explored: `all-MiniLM-L12-v2`, `bert-base-nli-mean-tokens`, and `all-mpnet-base-v2`[3]. All best-performing models used the embeddings from `all-MiniLM-L12-v2`. Using KELM, four kernels were explored: RBF, linear, polynomial, and sigmoid. All best-performing models used the linear kernel. Different values for the regularization coefficient C were explored in the range $[1, 20]$ and all possible combinations of functionals were tested for each model.

Table 5.1 shows the best-performing models on the validation set, optimized with model-specific parameters. Regarding the emotional attention and emotional filter models, the best-performing parameters for each emotional setting are shown. The RMSE column shows the performance of the models when predicting depression severity for the combined instances in the 7 temporary test parts (142 values in total). The results indicate that the plain unimodal text model (T) performs better than the unimodal audio model (A). The addition of an emotional filter to the audio modality (AEF) improves the performance of the plain audio model (A) the most. The top 35% emotional sentences with respect to valence seem to contain the most relevant information for a depression severity prediction.

The use of emotional attention with the audio modality (AEA) improves the performance of the plain audio model (A) more than the emotional filtering (AEF). Attention with raw arousal values shows the best performance for this model, performing similarly to the plain text model (T). The use of emotional attention with the text modality (TEA) improves the performance of the plain text model (T), this model shows the best performance on the validation set.

The feature fusion model (FF) does not improve the predictions of the plain unimodal models (A and T). The models with decision fusion (DF), however, do outperform the plain unimodal models. The best fusion method is the model-level fusion (TEA). The best model-agnostic fusion strategy is gender-weighted decision fusion (GWDF).

[3] Sentence Transformers by HuggingFace

Model	**Modality**	**Functionals**	*C*	**RMSE**
A	A	Mean, var, median, max	2	**5.13**
T	T	Median	3	**4.93**

Model	**Modality**	**VAD**	**Top%**	**Functionals**	*C*	**RMSE**
AEF	A	V	35	Mean	4	**5.01**
		A	25	Mean, var, sd, q75, max	9	5.03
		D	80	Var, sd, median, max	6	5.08

Model	**Modality**	**VAD**	**Abs**	**Functionals**	*C*	**RMSE**
AEA	A	V		Mean, var, sd, max	2	5.07
		V	abs	Mean, var, sd, q75, max	2	5.08
		A		Mean, var, sd, median, max	9	**4.93**
		A	abs	Var, sd, median, max	9	4.99
		D		Mean, var, sd, q75, max	6	5.06
		D	abs	Mean, var, q75, max	10	5.05
TEA	A+T	V		Median, q75, max	5	4.82
		V	abs	Median, q75, max	5	4.82
		A		Q75	10	**4.62**
		A	abs	Q75	10	4.63
		D		Sd, q75	8	4.70
		D	abs	Sd, q75	8	4.70

Model	**Modality**	*C*	**RMSE**
FF	A+T	2	**5.07**

Model	**Modality**	w_A^f	w_T^f	w_A^m	w_T^m	**RMSE**
DF	A+T	0.5	0.5	0.5	0.5	4.85
WDF	A+T	0.2	0.8	0.2	0.8	4.82
GWDF	A+T	0.2	0.8	0.51	0.49	**4.78**

Table 5.1: Performance overview of the 9 (4 unimodal and 5 bimodal) models on the validation set with their respective parameters. The setting with the best performance per model is highlighted in bold. The FF model and the DF models use the predictions of the plain A and T models. The column names for the three DF models indicate (optimal) weights for the audio modality, and text modality, for females and males.

5.2 Comparison with State-Of-The-Art Models

A comparison of our unimodal models with state-of-the-art unimodal models in predicting depression severity on the test set of the DAIC-WOZ can be found in Table 5.2. The corresponding RMSE performance on the 7-fold CV using the validation set is also shown, to check generalizability of the optimized models to the test set.

The unimodal text model (highlighted in bold) outperforms the current best predictive uni- and bimodal model using the audio and/or text modality, both in RMSE and MAE. The other unimodal models surpass the AVEC'2017 baseline (Ringeval et al., 2017). The addition of emotional filtering improves the performance of the plain audio model (similar to the results on the validation set), but adding emotional attention to the plain audio model to improve performance does not generalize to the test set.

Source	**Modality**	**RMSE**	**MAE**	CV RMSE
AVEC'2017 baseline	A	7.78	5.72	
Z. Zhao et al. (2020)	A	5.66	4.28	
Oureshi et al. (2021)	A	-	5.11	
Oureshi et al. (2021)	T	-	3.78	
Rohanian et al. (2019)	T	6.05	4.98	
Ours	A	5.44	4.26	5.13
	A (EF)	5.41	4.12	5.01
	A (EA)	5.82	4.68	**4.93**
	T	**4.23**	**3.62**	**4.93**

Table 5.2: Unimodal depression severity model performances on the DAIC-WOZ test set. The CV column contains the performance of this setting using 7-fold CV on the validation set (also shown in Table 5.1). The best-performing model per column is highlighted in bold.

The performance of the bimodal models on the test set (and validation set) can be found in Table 5.3. This shows that adding emotional attention to the plain text model does not improve the performance on the test set. Additionally, TEA is not the best-performing bimodal model on the test set, but WDF outperforms all the other bimodal models. This bimodal model also performs better than state-of-the-art unimodal and bimodal audio-text models. WDF does not perform better than our unimodal text model or the model by Rohanian et al. (2019) in terms of MAE on the test set. However, RMSE was the main challenge performance measure in the AVEC'2017 instead of MAE.

Source	**Fusion method**	**RMSE**	**MAE**	CV RMSE
Al Hanai et al. (2018)	Fusion scoring	6.27	4.97	
Muzammel et al. (2021)	Model-level	4.47	-	
Rohanian et al. (2019)	Word-level	5.14	**3.66**	
Ours	Model-level (TEA)	5.06	4.04	**4.62**
	FF	5.30	4.17	5.07
	DF	5.31	4.00	4.85
	WDF	**4.35**	3.72	4.82
	GWDF	4.62	3.96	4.78

Table 5.3: Bimodal audio-text depression severity model performances on the DAIC-WOZ test set. The CV column contains the performance of this setting using 7-fold CV on the validation set. The best-performing model per column is highlighted in bold.

Many studies do not predict depression severity, but binary depression score, therefore our severity predictions were converted to binary scores (using Equation 23) and evaluated on binary performance metrics using the test set (see Table 5.4). The results of DF are not included as this model did not predict any severity score above or equal to 10, therefore no depression labels were predicted.

Source	**Modality**	**Acc**	**Recall**	**Precision**	**F1-score**
Milintsevich et al. (2023)	T	-	-	-	0.74
Rohanian et al. (2019)	T	-	-	0.68	0.69
Rohanian et al. (2019)	A + T	-	-	**0.78**	**0.80**
Guo et al. (2022)	A + T	0.74	0.79	0.55	0.65
Al Hanai et al. (2018)	A + T	**-**	-	0.72	0.75
Ours	A (A)	0.74	0.50	0.58	0.54
	A (AEF)	0.77	0.50	0.64	0.56
	A (AEA)	0.72	0.36	0.56	0.43
	T (T)	**0.83**	**0.79**	0.69	0.73
	A + T (TEA)	0.66	0.57	0.44	0.50
	A + T (FF)	0.77	0.50	0.64	0.56
	A + T (WDF)	0.79	0.57	0.67	0.62
	A + T (GWDF)	0.77	0.50	0.64	0.56

Table 5.4: Binary depression detection model performances on the DAIC-WOZ test set.

The results above show that our models perform on par with the state-of-the-art models. The text model outperforms the existing models on depression severity prediction in terms of accuracy and recall, although the bimodal model by Al Hanai et al. (2018) and Rohanian et al. (2019) are slightly better on the binary depression classification in terms of precision and F1. The models will now be evaluated on gender fairness.

5.3 Gender Fairness of the Models

As mentioned earlier, gender fairness will be measured by the gender ratio of performance metrics on the test set: M_{EqAcc}, M_{EqOpp}, M_{PredEq} (see Equations 4, 11, 12). Table 5.5 presents the gender fairness of the uni and bimodal models on the test set with performance values outside the range of fair ratios (below 0.8 or above 1.2) highlighted in bold. Again, the results of DF are not included as this model did not predict any severity score above or equal to 10, therefore no depression labels were predicted.

Model	M_{EqAcc}	M_{EqOpp}	M_{PredEq}
A	0.93	(f) **1.33**	(m) **1.41**
AEA	1.10	(f) **1.50**	0.94
AEF	0.89	(f) **1.33**	0.94
T	(m) **1.21**	0.83	(m) **1.41**
TEA	1.16	1	(m) **2.20**
FF	0.93	(f) **1.33**	0.94
DF	1.07	-	-
WDF	1.12	1.00	0.94
GWDF	0.99	(f) **1.33**	0.94

Table 5.5: Gender fairness of our uni and bimodal models on DAIC-WOZ test set. Unfair performances are highlighted in bold. The letter of the sex (f = female, m = male) indicates a bias favoring this group.

The unimodal audio model (A) does not show a bias in terms of the equal accuracy, but it does show a (contradicting) bias in terms of equal opportunity and predictive equality: M_{EqOpp} signifies a bias in favor of females and M_{PredEq} signifies a bias in favor of males. The audio model with attention (AEA) and emotional filter (AEF) both show a bias in favor of females ($M_{\text{PredEq}} > 1.2$), but the bias in terms of predictive equality disappeared. The unimodal text model (T) is biased towards males ($M_{\text{EqAcc}} > 1.2$ and $M_{\text{PredEq}} > 1.2$). The bimodal text model with attention (TEA) is fair in terms of equal accuracy and equal opportunity, but not in terms of predictive equality ($M_{\text{PredEq}} > 1.2$), indicating a large bias in favor of males. The other bimodal models (FF and GWDF) show a bias in favor of females ($M_{\text{EqOpp}} > 1.2$). The WDF is solely deemed fair according to these fairness metrics.

Adding an emotional filter or emotional attention to the audio model (A) had the same effect regarding gender fairness: it removed the bias favoring males in terms of predictive equality, but the bias favoring females in terms of equal opportunity remained. Adding emotional attention to the text model (T) did remove the bias favoring males in terms of equal accuracy, but in increased the bias favoring males in terms of predictive equality.

The distributions of true depression scores and predictions of the overall best-performing model (T; bias in favor of males) on the test set are shown in Figure 5.1. This figure shows that the prediction value range is narrower than the ground truth value range in the test set. This phenomenon is referred to as *regression to the mean problem*, implying that predictions tend to be more conservative, clustering around the mean value of the training set to minimize RMSE loss.

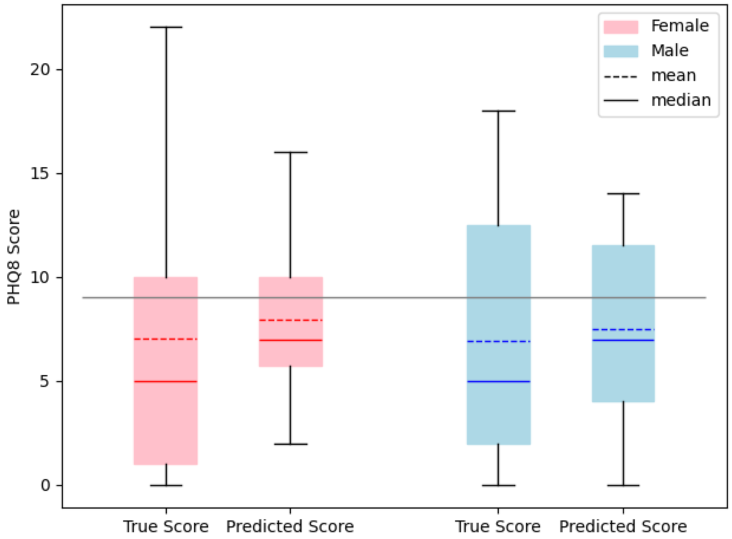


Figure 5.1: Boxplot of true and predicted PHQ8 scores per gender in the DAIC-WOZ test set by the best unimodal model (T).

Figure 5.1 also indicates that there is not enough evidence that a post-processing bias mitigation method (such as gender-specific thresholding for the depression binary score) would make the model more fair for females. The continuous fairness metric M_{EqAcc} would not change – as thresholding for binary classification has no effect on depression severity prediction – and it is unclear whether the binary fairness metrics would improve.

This section showed that not all uni and bimodal models are equally fair. The two plain unimodal models perform good overall, but are the least fair, therefore three pre-processing bias mitigation methods will now be explored with the aim of improving the fairness of these models.

5.4 Effect of Bias Mitigation

Table 5.6 illustrates the data imbalance in the validation set, which serves as a training set for models evaluated on the test set. The pre-processing bias mitigation methods address this imbalance by increasing the number of samples in the minority classes to match the sample count of the majority class (nondepressed males). Specifically, the number of samples for nondepressed females will be increased by 32, for depressed females by 38, and for depressed males by 43.

No. samples	**Females**	**Males**
Nondepressed	29	61
Depressed	23	18

Table 5.6: Number of samples per class in the validation set of the DAIC-WOZ.

Table 5.7 shows the effect of bias mitigation methods for the plain text model (T) and audio model (A). Resampling entails re-using samples from the minority classes. For the text model (T) – which is biased in favor of males – resampling does not improve the fairness of the audio model significantly in terms of equal accuracy. It increases the existing bias in favor of males in terms of predictive equality.

Model	**Bias mitigation**	**RMSE**	M_{EqAcc}	M_{EqOpp}	M_{PredEq}
T	-	4.23	(m) **1.21**	0.83	(m) **1.41**
	Resampling	4.64	1.20	0.83	(m) **1.88**
	Split interview	4.90	(m) **1.36**	0.80	0.94
	MixFeat	4.38	(m) **1.21**	0.83	(m) **1.41**
A	-	5.45	0.93	(f) **1.33**	(m) **1.41**
	Resampling	5.43	0.86	(f) **2.00**	0.94
	Split interview	6.20	1.01	1.00	0.94
	MixFeat	5.49	0.92	(f) **1.67**	(m) **1.41**

Table 5.7: Effect of bias mitigation methods for unimodal models on DAIC-WOZ test set. Unfair performances are highlighted in bold. The top row for SBERT and PDEM shows the original model performance without bias mitigation.

Resampling has a positive effect for females regarding the audio model (A). Both the equal opportunity and predictive equality measures become more biased towards females, thereby increasing the existing bias in equal opportunity and removing the bias in predictive equality.

The split interview technique uses subsets of interviews to make multiple session-level embeddings. For the depressed males, each interview was split into three subsets, creating three new samples for each participant (54 new samples in total). For the depressed and nondepressed females, each interview was split in two (creating 46 and 58 new samples respectively). A subset of new samples is selected for each class and added to the original samples of that class, ensuring that the combined total matches the sample size of the majority class.

For the text model (T), splitting each interview into subsets increases the existing bias favoring males in equal accuracy, but removes the bias favoring males in predictive equality. For the audio model, the contradicting biases are mitigated and the split interview technique results in a fair model. However, splitting the interview has a negative impact on overall performance.

The MixFeat approach generates synthetic samples per minority class to match the size of the majority class (Cheong, Spitale, & Gunes, 2023). This technique does not improve the fairness of the text model, and it increases the bias favoring females in the audio model.

Overall, some of the bias mitigation methods succeed in improving fairness, but only for some models or for specific fairness measures. Almost all methods have a negative effect on the overall performance.

6 Discussion

This section addresses the research questions and discusses the implications of the findings in this study on multimodal depression prediction and gender fairness. Additionally, the limitations of the research will be presented. This will be followed by future directions and the conclusion of this study.

6.1 Answering the Research Questions

This research studied three research questions. In the following sections, each of them will be discussed.

6.1.1 Performance of Uni- and Bimodal Models (RQ 1)

The performance of various models in predicting depression severity using the DAIC-WOZ dataset showed that our unimodal text model, using SBERT embeddings, outperforms current state-of-the-art unimodal and bimodal audio-text models on the test set.

Based on the results in the validation set, it was expected that the model-level fusion method (adding emotional attention from speech to the text model) would outperform the plain text model on the test set, but this was not found. The unimodal audio models surpassed the baseline of AVEC'2017 but generally did not perform as well as the plain text model on the test set. This was similar to the results on the validation set.

Among the model-agnostic fusion models, the gender-weighted decision fusion model showed the best performance on the validation set, but the weighted decision fusion model outperformed the other bimodal models on the test set. This model also outperformed the state-of-the-art audio-text models.

Regarding binary depression detection, most of our models perform on par with the state-of-the-art models. Even though our models were not trained to predict binary depression scores, the unimodal text model outperformed existing models in accuracy and recall.

Overall, the models in this study perform on a state-of-the-art level. The findings are similar to earlier research as the text modality is the most informative for depression prediction (Ray et al., 2019; Van Steijn et al., 2022; Oureshi et al., 2021). Combining the audio modality with the text modality did not improve the depression severity predictions.

Effect of adding emotional attention to unimodal models (RSQ 1A)

Adding an emotional attention component to the unimodal audio model improved the performance on the validation set. However, this improvement did not generalize well to the test set, where adding emotional attention did not enhance performance. These results are similar to the text model results: adding emotional attention improved performance on the validation set, making it the best-performing model, but this improvement was not found when the model was tested on the test set.

Effect of adding emotional filtering to unimodal audio model (RSQ 1B)

The emotional filtering component improved the performance of the plain audio model on the validation set. This improvement was consistent with the results observed on the test set, where emotional filtering also improved predictions by the plain audio model. This shows that the emotional part of a conversation (based on audio information) is more informative for depression severity prediction than the conversation as a whole.

6.1.2 Gender Fairness of Unimodal and Bimodal Models (RQ 2)

The result section revealed that the unimodal and bimodal models demonstrate gender biases, although the direction and magnitude of these biases varied. The plain unimodal text model showed a bias in favor of males, as well as the bimodal text model with emotional attention.

The plain unimodal audio model showed conflicting biases, with equal opportunity favoring females and predictive equality favoring males. This can be explained by a higher number of depressed predictions (severity prediction equal to or above 10) for the female class than for the male class, possibly caused by the imbalance in depression samples across genders in the validation set. The audio models with emotional attention and emotional filtering showed a bias in favor of females in equal opportunity.

For the model-agnostic bimodal models, feature fusion and gender-weighted decision fusion showed a bias favoring females. The decision fusion model was fair according to our gender fairness metrics. Overall, it seems like the text data is more informative for males than for females in depression severity prediction, and the audio data is more informative for females than for males in depression severity prediction.

6.1.3 Effect of Bias Mitigation Methods on Gender Fairness (RQ 3)

Several pre-processing bias mitigation methods were explored to balance the data, which could have caused gender biases in the unimodal models: resampling, the split interview technique, and the MixFeat approach (Cheong, Spitale, & Gunes, 2023).

Resampling was effective for the audio model in terms of reducing bias in predictive equality but less effective for the text model, where it increased existing biases. Splitting the interviews was effective in achieving fairness for the audio model but at the cost of overall performance. For the text model, this method partially improved fairness by removing the bias in predictive equality. The MixFeat approach was not effective in improving the fairness of the text model and it increased the gender bias in the audio model.

Overall, the bias mitigation methods show mixed results, improving fairness in some models and measures, while often negatively impacting overall performance. This highlights the trade-off between fairness and accuracy (Dutta et al., 2020).

6.2 Discussion on the DAIC-WOZ and Performance Metrics

Given the regression to the mean phenomenon and the predominance of non-depressive samples in the DAIC-WOZ, all the models tend to predict lower severity scores than the true severity scores. A recent expert user study (Sogancioglu et al., 2024) on the importance of different fairness measures in mental healthcare revealed that for major depression disorder, a lower prediction is more harmful than a higher prediction compared to the actual severity score when there is a suicidal risk. However, for mild depression cases, incorrectly predicting someone to be more depressed than they are can have negative consequences due to unnecessary use of medication.

Because of the size of the dataset, models were optimized on the validation set with CV. The optimal parameters showed a good fit in the CV stage, however, in certain cases the 7-fold CV performance was not indicative of the test set performance. Nevertheless, the best uni- and bimodal models still outperformed state-of-the-art models.

The DAIC-WOZ contains data of participants talking to a virtual avatar. This avatar (controlled by a human in another room) conducts a semi-structured interview, asking each participant a similar set of questions. It has been reported that virtual humans can increase willingness to disclose (Lucas et al., 2014), but it is unclear whether the audio and/or text data would be similar in a situation where the interviewer was human. Earlier research shows that participants perceive rapport[4] more often and enclose mental symptoms more often in a conversation with a real expert than with a virtual agent (Yokotani et al., 2018). More research should be conducted to learn about the best setting for collecting multimodal data in order to detect depression or predict depression severity.

6.2.1 Discussion on Fairness Metrics

The results show the importance of motivating the chosen fairness metrics. Metrics such as equal accuracy, equal opportunity, and predictive equality provide different perspectives on model fairness. Equal accuracy was based on the ratio in RMSE between different classes of our sensitive attribute. While RMSE provides an average error per gender class, it does not indicate the direction of the error. This limitation can lead to significant issues. For example, if predictive accuracy is equal between genders but the model consistently overestimates depression severity

[4] Rapport: A relationship characterized by agreement, mutual understanding, or empathy that makes communication possible. Retrieved from `https://www.merriam-webster.com/dictionary/rapport` (accessed: 13-07-2024)

for females and underestimates depression severity for males, a bias is present. This scenario could result in females being taken more seriously and treated with greater caution, whereas depressed males might be perceived as less depressive. Such biases can have serious and unfair consequences.

Evaluating gender fairness in binary depression detection raises questions about the utility of such metrics. Binary depression prediction simplifies the task, but does not consider the range of depression severity within the binary depression classes. It might be more important to classify someone with depression if the severity is high, than classifying someone with depression if this person scores just above 10 on the PHQ8. The limitation of binary fairness metrics is linked to the approach of viewing mental illnesses and their symptoms as binary states – either present or absent. The DSM implies that individuals either have a mental illness or not, without considering the spectrum of symptom severity. In contrast, tools like the PHQ8 acknowledge nuances in symptom severity, making continuous fairness metrics more suitable for this context.

6.3 Ethical Considerations

This section contains ethical considerations in the research phase of this study, but also ethical considerations about the use of this study in future practices. The current study complies with the ethics and privacy regulations for research projects at Utrecht University. Before conducting the research, an ethics and privacy quick scan was completed with ethical considerations related to human participants, data protection and storage, potential harms, and conflicts of interest.

The research project involves personal data, defined as information about an identifiable living person: the DAIC-WOZ contains anonymized sensitive personal data about individuals' mental health. The data collection was conducted via informed consent and the participants in this study can not be identified using the available information, although it is theoretically possible to recognize someone from their voice recordings. Nonetheless, all data was stored securely and protected by passwords. No contractual conditions are attached to the use of this data.

In processing sensitive information, innovative technology[5] was used, of which the primary objective was to predict mental disorders from behavioral indicators. This has no impact on the participants in the dataset. Additionally, the likelihood of harm to any institution is extremely low and no conflict of interest could affect the research outcome, as the project is not funded and no specific outcome has beneficial consequences for any party involved.

Other ethical considerations of this study are related to interpretability and fairness. The models – such as those employed in this study – can be used in practice to support the decision of medical experts to treat someone in a particular way. In the development of the models, we

[5] As defined by Utrecht University: machine learning (including deep learning).

chose to predict the severity of 8 symptoms rather than directly predicting a total depression severity score or binary depression score. This approach allows for an analysis of individual symptom contributions to the total severity score, which can explain the total severity score. The current study has not explored the interpretability of this method, however, this is an important consideration if such predictive tools are to be used in future practices. The other ethical aspect of the current research is fairness. As mentioned earlier, unfairness in mental health detection models can have dangerous consequences for patients. Different bias mitigation methods have been explored to combat unintentional biases, with varying success.

It is important to note that depression detection models are not intended for personal use and should not be made publicly available. Public access to these models could lead individuals to mistakenly accept the model predictions as definitive, without knowing the risk of wrong predictions.

6.4 Future Directions

In this section, specific research directions relevant to the proposed approach in this research will be discussed. Then more general future research directions will be explored.

The current approach used multitask regression to predict symptom severities, summing the regression scores to a total severity score. Future research could explore ordinal classification for symptom severity instead of regression. Ordinal classification could provide more insights into the severity of each symptom, providing the set of symptoms that are particularly burdensome for a patient. This approach may help medical experts to develop effective treatment strategies.

Regarding the features in the used models. Speech rate has not been included in the prediction of depresion severity, although this feature has been shown to differ between depressed and nondepressed people (Yamamoto et al., 2020). Additionally, the emotional dimensions valence, arousal, and dominance have now been extracted from the audio modality and applied to both modalities in terms of attention and filtering. Sentiment analysis from text data could be included in future research to confirm the emotional load of sentences, or to develop a bimodal model by combining audio data with textual emotion.

As mentioned in the introduction, the presence of certain symptoms from the DSM does not necessarily indicate Major Depressive Disorder. Future mental health predictive systems should be capable of distinguishing between multiple mental disorders based on the symptom criteria outlined in the DSM. This would enhance the precision of mental health assessment tools and ensure that patients receive appropriate and targeted treatment.

The beginning of this report stressed that medical experts should be able to understand fully why a patient is suffering. With access to more data and more depressed samples for both females and males, we can develop models that are not only more accurate but also fairer. As with any ML approach, the availability of enough data is crucial. Collecting more data could enable the development of models that generalize better across different populations and subsets of data, thereby enhancing the robustness and applicability of the models in real-world scenarios.

Additionally, to ensure the generalizability of the current models, the models in the current study can be evaluated on other (clinical) datasets to detect depression (severity). This validation would help verify whether the research findings extend to other contexts.

Although the current models demonstrate better performance in predicting depression severity than existing models, they need to be further improved. Future research should address challenges such as improving model performance, fairness, and validating the models in diverse real-world settings. This is essential for applying these models as decision-support tools in clinical settings to improve the process of diagnosing mental illnesses.

6.5 Conclusion

This research aimed to explore methods that can assist medical experts in the diagnostic process in a responsible manner, to ensure the best possible outcomes for patients with and without a mental disorder. The performance and fairness of uni- and multimodal models in predicting depression severity in a clinical setting have been evaluated, using the DAIC-WOZ dataset. The unimodal text model outperformed state-of-the-art models, while the unimodal audio and bimodal audio-text models exceeded the AVEC'2017 baseline but not the text model. The combination of audio and textual information did not reveal important gender patterns that could improve over the unimodal models. Gender biases were observed in the unimodal and bimodal models, and data augmentation bias mitigation methods showed mixed results. These methods sometimes improved fairness, but mostly at the cost of overall performance.

Ethical considerations such as fairness and interpretability have been taken into account during model development. Future research directions include evaluating the proposed models on other clinical datasets, exploring more features, using ordinal classification for symptom severity, and developing models capable of distinguishing multiple mental disorders based on DSM criteria. More research is needed before these models can be applied in clinical practice to enhance mental health diagnosis and possibly reduce the time that patients, their relatives, and society suffer from the burden of mental health illnesses.

References

Adarsh, V., Arun Kumar, P., Lavanya, V., & Gangadharan, G. (2023). Fair and explainable depression detection in social media. *Information Processing & Management*, *60*(1), 103168. doi: 10.1016/j.ipm.2022.103168

Afifi, M. (2007). Gender differences in mental health. *Singapore medical journal*, *48*(5), 385.

Alghowinem, S., Goecke, R., Wagner, M., Epps, J., Hyett, M., Parker, G., & Breakspear, M. (2018). Multimodal depression detection: Fusion analysis of paralinguistic, head pose and eye gaze behaviors. *IEEE Transactions on Affective Computing*, *9*(4), 478-490. doi: 10.1109/TAFFC.2016.2634527

Al Hanai, T., Ghassemi, M. M., & Glass, J. R. (2018). Detecting depression with audio/text sequence modeling of interviews. In *Interspeech* (pp. 1716–1720). doi: 10.21437/Interspeech.2018-2522

American Psychiatric Association. (1952). *Diagnostic and statistical manual of mental disorders.* Washington, D.C.: American Psychiatric Association.

American Psychiatric Association. (2013a). Depressive disorders. In *Diagnostic and statistical manual of mental disorders* (pp. 160–168). doi: 10.1176/appi.books.9780890425787.x04_Depressive_Disorders

American Psychiatric Association. (2013b). *Diagnostic and statistical manual of mental disorders, fifth edition.* Arlington, VA: American Psychiatric Publishing.

Awasthi, P., Kleindessner, M., & Morgenstern, J. (2020). Equalized odds postprocessing under imperfect group information. In *International conference on artificial intelligence and statistics* (pp. 1770–1780).

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449–12460.

Bailey, A., & Plumbley, M. D. (2021). Gender bias in depression detection using audio features. In *2021 29th European Signal Processing Conference (EUSIPCO)* (pp. 596–600).

Baltrusaitis, T., Robinson, P., & Morency, L.-P. (2013). Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 354–361).

Baltrušaitis, T., Robinson, P., & Morency, L.-P. (2016). Openface: an open source facial behavior analysis toolkit. In *2016 IEEE winter conference on applications of computer vision (WACV)* (pp. 1–10).

Barocas, S., Hardt, M., & Narayanan, A. (2017). Fairness in machine learning. *Nips tutorial*, *1*, 2017.

Barsky, A. J., Peekna, H. M., & Borus, J. F. (2001). Somatic symptom reporting in women and men. *Journal of general internal medicine*, *16*(4), 266–275.

Bolton, D. (2013). What is mental illness. *The Oxford handbook of philosophy and psychiatry*, 434–450.

Booth, B. M., Hickman, L., Subburaj, S. K., Tay, L., Woo, S. E., & D'Mello, S. K. (2021). Bias and fairness in multimodal machine learning: A case study of automated video interviews. In *Proceedings of the 2021 international conference on multimodal interaction* (pp. 268–277).

Chaplin, T. M. (2015). Gender and emotion expression: A developmental contextual perspective. *Emotion Review*, *7*(1), 14–21.

Chen, J., Chen, Z., Chi, Z., & Fu, H. (2014). Emotion recognition in the wild with feature fusion and multiple kernel learning. In *Proceedings of the 16th international conference on multimodal interaction* (pp. 508–513).

Cheong, J., Kuzucu, S., Kalkan, S., & Gunes, H. (2023). Towards gender fairness for mental health prediction. In *International joint conferences on artificial intelligence organization.* doi: 10.17863/CAM.96646

Cheong, J., Spitale, M., & Gunes, H. (2023). "it's not fair!"; fairness for a small dataset of multi-modal dyadic mental well-being coaching.. doi: 10.17863/CAM.97210

Chowdhary, K. R. (2020). Natural language processing. In *Fundamentals of Artificial Intelligence* (pp. 603–649). New Delhi: Springer India. doi: 10.1007/978-81-322-3972-7_19

Çiftçi, E., Kaya, H., Güleç, H., & Salah, A. A. (2018). The turkish audio-visual bipolar disorder corpus. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)* (pp. 1–6).

Cohen, J., Richter, V., Neumann, M., Black, D., Haq, A., Wright-Berryman, J., & Ramanarayanan, V. (2023). A multimodal dialog approach to mental state characterization in clinically depressed, anxious, and suicidal populations. *Frontiers in psychology*, *14*.

Corcoran, C. M., Carrillo, F., Fernández-Slezak, D., Bedi, G., Klim, C., Javitt, D. C., ... Cecchi, G. A. (2018). Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, *17*(1), 67–75.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech communication*, *71*, 10–49.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. doi: 10.18653/v1/N19-1423

Dhamija, S., & Boult, T. E. (2017). Exploring contextual engagement for trauma recovery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 19–29).

Difrancesco, S., Riese, H., Merikangas, K. R., Shou, H., Zipunnikov, V., Antypa, N., ... Lamers, F. (2021, Feb 17). Sociodemographic, health and lifestyle, sampling, and mental health determinants of 24-hour motor activity patterns: Observational study. *J Med Internet Res*, *23*(2), e20700. doi: 10.2196/20700

Djudiyah, M. S., Harding, D., & Sumantri, S. (2016). Gender differences in neuroticism on college students. In *Proceedings of the 2nd ASEAN Conference on Psychology & Humanity.*

Doran, C. M., & Kinchin, I. (2017). A review of the economic impact of mental illness. *Australian Health Review*, *43*(1), 43–48.

Dutta, S., Wei, D., Yueksel, H., Chen, P.-Y., Liu, S., & Varshney, K. (2020, 13–18 Jul). Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing. In H. D. III & A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning* (Vol. 119, pp. 2803–2813). PMLR.

Estroff, T. W., & Gold, M. S. (2018). Psychiatric misdiagnosis. In *Advances in psychopharmacology* (pp. 33–66). CRC Press.

Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia* (pp. 1459–1462).

Fan, W., He, Z., Xing, X., Cai, B., & Lu, W. (2019). Multi-modality depression detection via multi-scale temporal dilated CNNs. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop* (pp. 73–80).

Fang, M., Peng, S., Liang, Y., Hung, C.-C., & Liu, S. (2023). A multimodal fusion model with multi-level attention mechanism for depression detection. *Biomedical Signal Processing and Control*, *82*, 104561.

Farreras, I. G. (2019). History of mental illness. *General psychology: required reading*, *244*.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 259–268).

Floyd, B. J. (1997). Problems in accurate medical diagnosis of depression in female patients. *Social science & medicine*, *44*(3), 403–412.

Galassi, A., Lippi, M., & Torroni, P. (2021). Attention in natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(10), 4291-4308. doi: 10.1109/TNNLS.2020.3019893

Gratch, J., Artstein, R., Lucas, G. M., Stratou, G., Scherer, S., Nazarian, A., … Marsella, S. (2014). The distress analysis interview corpus of human and computer interviews. In *Lrec* (pp. 3123–3128).

Guo, Y., Zhu, C., Hao, S., & Hong, R. (2022). A topic-attentive transformer-based model for multimodal depression detection. *arXiv preprint arXiv:2206.13256*. Retrieved from `https://doi.org/10.48550/arXiv.2206.13256` doi: 10.48550/arXiv.2206.13256

Gurpinar, F., Kaya, H., & Salah, A. A. (2016). Kernel elm and cnn based facial age estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 80–86).

Hallowell, A. I. (1934). Culture and mental disorder. *The Journal of Abnormal and Social Psychology*, *29*(1), 1.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, *29*.

Hawton, K., Rodham, K., Evans, E., & Weatherall, R. (2002). Deliberate self harm in adolescents: self report survey in schools in england. *Bmj*, *325*(7374), 1207–1211.

Hort, M., Chen, Z., Zhang, J. M., Harman, M., & Sarro, F. (2024). Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. In (Vol. 1). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3631326

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3451-3460. doi: 10.1109/TASLP.2021.3122291

Huang, G.-B., Zhou, H., Ding, X., & Zhang, R. (2011). Extreme learning machine for regression and multiclass classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *42*(2), 513–529.

Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. (2004, Jul. 25–29). Extreme learning machine: A new learning scheme of feedforward neural networks. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN)* (Vol. 2, pp. 985–990). Budapest, Hungary.

Jakobsen, P., Garcia-Ceja, E., Riegler, M., Stabell, L. A., Nordgreen, T., Torresen, J., … Oedegaard, K. J. (2020, 08). Applying machine learning in motor activity time series of depressed bipolar and unipolar patients compared to healthy controls. *PLOS ONE*, *15*, 1-16. doi: 10.1371/journal.pone.0231995

Kamishima, T., Akaho, S., Asoh, H., & Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23* (pp. 35–50).

Kaya, H., Gürpınar, F., & Salah, A. A. (2017). Video-based emotion recognition in the wild using deep transfer learning and score fusion. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction* (pp. 188–195).

Kaya, H., Hantke, S., Schuller, B. W., Valstar, M., Scherer, K., & Pantic, M. (2019). Predicting depression and emotions in the cross-roads of cultures, para-linguistics, and non-linguistics. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop* (pp. 27–35).

Kessler, R. C., Demler, O., Frank, R. G., Olfson, M., Pincus, H. A., Walters, E. E., … Zaslavsky, A. M. (2005). Prevalence and treatment of mental disorders, 1990 to 2003. *New England Journal of Medicine*, *352*(24), 2515–2523.

Khoo, L. S., Lim, M. K., Chong, C. Y., & McNaney, R. (2024). Machine learning for multimodal mental health detection: A systematic review of passive sensing approaches. *Sensors*, *24*(2), 348.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232.

Kroenke, K., Spitzer, R. L., & Williams, J. B. (2001). The phq-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613.

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B., Berry, J. T., & Mokdad, A. H. (2009). The phq-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, *114*(1), 163-173. doi: https://doi.org/10.1016/j.jad.2008.06.026

Lahey, B. B., Tiemeier, H., & Krueger, R. F. (2022). Seven reasons why binary diagnostic categories should be replaced with empirically sounder and less stigmatizing dimensions. *JCPP advances*, *2*(4), e12108.

Le Glaz, A., Haralambous, Y., Kim-Dufor, D.-H., Lenca, P., Billot, R., Ryan, T. C., ... Berrouiguet, S. (2021). Machine learning and natural language processing in mental health: systematic review. *Journal of Medical Internet Research*, *23*(5), e15708.

Lester, D., & Heim, N. (1992). Sex differences in suicide notes. *Perceptual and Motor Skills*, *75*(2), 582–582.

Lever-van Milligen, B. A., Lamers, F., Smit, J. H., & Penninx, B. W. (2020). Physiological stress markers, mental health and objective physical function. *Journal of Psychosomatic Research*, *133*, 109996. doi: https://doi.org/10.1016/j.jpsychores.2020.109996

Li, S., & Deng, W. (2022). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, *13*(3), 1195-1215. doi: 10.1109/TAFFC.2020.2981446

Liu, Y., Hankey, J., Cao, B., & Chokka, P. (2021). Screening for major depressive disorder in a tertiary mental health centre using EarlyDetect: A machine learning-based pilot study. *Journal of affective disorders reports*, *3*, 100062.

Lucas, G. M., Gratch, J., King, A., & Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, *37*, 94-100. doi: https://doi.org/10.1016/j.chb.2014.04.043

Ma, X., Yang, H., Chen, Q., Huang, D., & Wang, Y. (2016). Depaudionet: An efficient deep model for audio based depression classification. In *Proceedings of the 6th international workshop on Audio/Visual Emotion Challenge* (pp. 35–42).

Mallol-Ragolta, A., Zhao, Z., Stappen, L., Cummins, N., & Schuller, B. W. (2019). A Hierarchical Attention Network-Based Approach for Depression Detection from Transcribed Clinical Interviews. In *Proc. interspeech 2019* (pp. 221–225). doi: 10.21437/Interspeech.2019-2036

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, *54*(6), 1–35.

Miaschi, A., & Dell'Orletta, F. (2020). Contextual and non-contextual word embeddings: an in-depth linguistic investigation. In S. Gella et al. (Eds.), *Proceedings of the 5th Workshop on Representation Learning for NLP* (pp. 110–119). Association for Computational Linguistics. doi: 10.18653/v1/2020.repl4nlp-1.15

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, *26*.

Milintsevich, K., Sirts, K., & Dias, G. (2023). Towards automatic text-based estimation of depression through symptom prediction. *Brain Informatics*, *10*(1), 1–14.

Mulay, A., Dhekne, A., Wani, R., Kadam, S., Deshpande, P., & Deshpande, P. (2020). Automatic depression level detection through visual input. In *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)* (pp. 19–22).

Muzammel, M., Salam, H., & Othmani, A. (2021). End-to-end multimodal clinical depression recognition using deep neural networks: A comparative analysis. *Computer Methods and Programs in Biomedicine*, *211*, 106433.

Nadkarni, P. M., Ohno-Machado, L., & Chapman, W. W. (2011, 09). Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, *18*(5), 544-551. doi: 10.1136/amiajnl-2011-000464

Niu, M., Chen, K., Chen, Q., & Yang, L. (2021). Hcag: A hierarchical context-aware graph attention model for depression detection. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4235–4239).

Ogrodniczuk, J. S., & Oliffe, J. L. (2011). Men and depression. *Canadian Family Physician*, *57*(2), 153–155.

Oureshi, S. A., Dias, G., Saha, S., & Hasanuzzaman, M. (2021). Gender-aware estimation of depression severity level in a multimodal setting. In *2021 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8).

Park, J., Arunachalam, R., Silenzio, V., & Singh, V. K. (2022, Jun 14). Fairness in mobile phone–based mental health assessment algorithms: Exploratory study. *JMIR Form Res*, *6*(6), e34366.

Parker, G., & Roy, K. (2001). Adolescent depression: a review. *Australian & New Zealand Journal of Psychiatry*, *35*(5), 572–580.

Patientenfederatie Nederland. (2023, october). *Diagnose GGZ.* `https://kennisbank.patientenfederatie.nl/app/answers/detail/a_id/1759/~/diagnose-%28ggz%29#:~:text=Hij%20kan%20vervolgens%20aangeven%20waar,aandoening%20er%20precies%.` (Accessed: 2023-11-16)

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. OpenAI. Retrieved from `https://openai.com/index/language-unsupervised/`

Ray, A., Kumar, S., Reddy, R., Mukherjee, P., & Garg, R. (2019). Multi-level attention network using text, audio and video for depression prediction. In *Proceedings of the 9th international on Audio/Visual Emotion Challenge and workshop* (pp. 81–88).

Ringeval, F., Schuller, B., Valstar, M., Gratch, J., Cowie, R., Scherer, S., … Pantic, M. (2017). Avec 2017: Real-life depression, and affect recognition workshop and challenge. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge* (p. 3–9). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3133944.3133953

Rohanian, M., Hough, J., & Purver, M. (2019). Detecting depression with word-level multimodal fusion. In *Interspeech* (pp. 1443–1447).

Schuller, B. (2011). Voice and speech analysis in search of states and traits. In *Computer analysis of human behavior* (pp. 227–253). Springer.

Sharma, M. (2022). Multi-lingual multi-task speech emotion recognition using wav2vec 2.0. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (p. 6907-6911). doi: 10.1109/ICASSP43922.2022.9747417

Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: a scoping review of methods and applications. *Psychological Medicine*, *49*(9), 1426–1448. doi: 10.1017/S0033291719000151

Shen, H., Zhang, L., Xu, C., Zhu, J., Chen, M., & Fang, Y. (2018). Analysis of misdiagnosis of bipolar disorder in an outpatient setting. *Shanghai Archives of Psychiatry*, *30*(2), 93.

Shi, Y., Li, P., Yuan, H., Miao, J., & Niu, L. (2019). Fast kernel extreme learning machine for ordinal regression. *Knowledge-Based Systems*, *177*, 44-54. doi: 10.1016/j.knosys.2019.04.003

Sims, R., Michaleff, Z. A., Glasziou, P., & Thomas, R. (2021). Consequences of a diagnostic label: A systematic scoping review and thematic framework. *Frontiers in Public Health*, *9*.

Sogancioglu, G., Kaya, H., & Salah, A. A. (2023). The effects of gender bias in word embeddings on patient phenotyping in the mental health domain. In *11th International Conference on Affective Computing and Intelligent Interaction (ACII)* (p. 1-8). doi: 10.1109/ACII59096.2023.10388203

Sogancioglu, G., Mosteiro, P., Salah, A. A., Scheepers, F., & Kaya, H. (2024). *Fairness in AI-Based Mental Health: Clinician Perspectives and Bias Mitigation.* (Under review)

Srimadhur, N., & Lalitha, S. (2020). An end-to-end model for detection and assessment of depression levels using speech. *Procedia Computer Science*, *171*, 12–21.

Stein, D. J., Palk, A. C., & Kendler, K. S. (2021). What is a mental disorder? an exemplar-focused approach. *Psychological Medicine*, *51*(6), 894–901. doi: 10.1017/S0033291721001185

Stengel, E. (1959). Classification of mental disorders. *Bulletin of the World Health Organization*, *21*(4-5), 601.

Straw, I., & Callison-Burch, C. (2020, 12). Artificial intelligence in mental health and the biases of language based models. *PLOS ONE*, *15*(12), 1-19. doi: 10.1371/journal.pone.0240376

Su, Y., Hu, B., Xu, L., Cai, H., Moore, P., Zhang, X., & Chen, J. (2014). Emotiono+: Physiological signals knowledge representation and emotion reasoning model for mental health monitoring. In *2014 IEEE International Conference on Bioinformatics and Biomedicine*

(BIBM) (p. 529-535). doi: 10.1109/BIBM.2014.6999215

Tao, F., Esposito, A., & Vinciarelli, A. (2023). The androids corpus: A new publicly available benchmark for speech based depression detection. *Depression*, *47*, 11–9.

Tkacz, J., & Brady, B. L. (2021). Increasing rate of diagnosed childhood mental illness in the united states: Incidence, prevalence and costs. *Public Health in Practice*, *2*, 100204. doi: 10.1016/j.puhip.2021.100204

Van Steijn, F., Sogancioglu, G., & Kaya, H. (2022). Text-based interpretable depression severity modeling via symptom predictions. In *Proceedings of the 2022 International Conference on Multimodal Interaction* (pp. 139–147).

Volksgezondheid en Zorg. (2023, october). *Wachttijden GGZ.* `https://www.vzinfo.nl/wachttijden/geestelijke-gezondheidszorg`. (Accessed: 2023-11-16)

Wagner, J., Triantafyllopoulos, A., Wierstorf, H., Schmitt, M., Burkhardt, F., Eyben, F., & Schuller, B. W. (2023). Dawn of the transformer era in speech emotion recognition: Closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(9), 10745-10759. doi: 10.1109/TPAMI.2023.3263585

Wang, Y., Wang, X., Beutel, A., Prost, F., Chen, J., & Chi, E. H. (2021). Understanding and improving fairness-accuracy trade-offs in multi-task learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (p. 1748–1757). New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3447548.3467326

Wei, P.-C., Peng, K., Roitberg, A., Yang, K., Zhang, J., & Stiefelhagen, R. (2023). Multimodal depression estimation based on sub-attentional fusion. In L. Karlinsky, T. Michaeli, & K. Nishino (Eds.), *Computer Vision – ECCV 2022 Workshops* (pp. 623–639).

Williamson, J. R., Godoy, E., Cha, M., Schwarzentruber, A., Khorrami, P., Gwon, Y., ... Quatieri, T. F. (2016). Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge* (pp. 11–18).

Wold, E., Blum, T., Keislar, D., & Wheaten, J. (1996). Content-based classification, search, and retrieval of audio. *IEEE multimedia*, *3*(3), 27–36.

Wolpert, D. H. (1992). Stacked generalization. *Neural networks*, *5*(2), 241–259.

Wright-Berryman, J., Cohen, J., Haq, A., Black, D. P., & Pease, J. L. (2023). Virtually screening adults for depression, anxiety, and suicide risk using machine learning and language from an open-ended interview. *Frontiers in Psychiatry*, *14*, 1143175.

Wu, Y., Levis, B., Riehm, K. E., Saadat, N., Levis, A. W., Azar, M., ... et al. (2020). Equivalency of the diagnostic accuracy of the phq-8 and phq-9: a systematic review and individual participant data meta-analysis. *Psychological Medicine*, *50*(8), 1368–1380. doi: 10.1017/S0033291719001314

Xiao, J., Huang, Y., Zhang, G., & Liu, W. (2021). A deep learning method on audio and text sequences for automatic depression detection. In *2021 3rd International Conference on Applied Machine Learning (ICAML)* (pp. 388–392).

Yamamoto, M., Takamiya, A., Sawada, K., Yoshimura, M., Kitazawa, M., Liang, K.-c., ... Kishimoto, T. (2020). Using speech recognition technology to investigate the association between timing-related speech features and depression severity. *PloS one*, *15*(9), e0238726.

Yang, L., Jiang, D., Xia, X., Pei, E., Oveneke, M. C., & Sahli, H. (2017). Multimodal measurement of depression using deep learning models. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge* (pp. 53–59).

Yang, L., Sahli, H., Xia, X., Pei, E., Oveneke, M. C., & Jiang, D. (2017). Hybrid depression classification and estimation from audio video and text information. In *Proceedings of the 7th annual workshop on Audio/Visual Emotion Challenge* (pp. 45–51).

Yokotani, K., Takagi, G., & Wakashima, K. (2018). Advantages of virtual agents over clinical psychologists during comprehensive mental health interviews using a mixed methods design. *Computers in Human Behavior*, *85*, 135-145. doi: https://doi.org/10.1016/j.chb.2018.03.045

Yoon, J., Kang, C., Kim, S., & Han, J. (2022). D-vlog: Multimodal vlog dataset for depression detection. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, pp. 12226–12234).

Zanna, K., Sridhar, K., Yu, H., & Sano, A. (2022). Bias reducing multitask learning on mental health prediction. In *2022 10th International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 1–8).

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (p. 335–340). Association for Computing Machinery. doi: 10.1145/3278721.3278779

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2017). mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, S., Zhang, S., Huang, T., Gao, W., & Tian, Q. (2017). Learning affective features with a hybrid deep model for audio–visual emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, *28*(10), 3030–3043.

Zhang, T., Schoene, A. M., Ji, S., & Ananiadou, S. (2022). Natural language processing applied to mental illness detection: a narrative review. *NPJ digital medicine*, *5*(1), 46.

Zhang, Z., Chen, S., Wu, M., & Zhu, K. (2022). Symptom identification for interpretable detection of multiple mental disorders on social media. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 9970–9985).

Zhao, S. (2005). The digital self: Through the looking glass of telecopresent others. *Symbolic interaction*, *28*(3), 387–405.

Zhao, Z., Bao, Z., Zhang, Z., Cummins, N., Wang, H., & Schuller, B. (2020). Hierarchical attention transfer networks for depression assessment from speech. In *2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 7159–7163).

Zogan, H., Razzak, I., Wang, X., Jameel, S., & Xu, G. (2022). Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web*, *25*(1), 281–304.

Zong, W., Huang, G.-B., & Chen, Y.-S. (2013). Weighted extreme learning machine for imbalance learning. *Neurocomputing*, *101*, 229–242.

Appendix A. Patient Health Questionnaire 8

The PHQ8 by Kroenke et al. (2009):

Using these options:

0	Not at all
1	Several days
2	More than half the days
3	Nearly everyday

Over the last 2 weeks, how often have you been bothered by any of the following problems?

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

The individual scores can be summed into a total depression severity score [0, 24].