UTRECHT UNIVERSITY
Faculty of Science
Department of Information and Computing Sciences
MSc Artificial Intelligence

A MULTIMODAL DEEP LEARNING APPROACH FOR AUTOMATED ASSESSMENT OF DEPRESSIVE SYMPTOMS IN CHILDREN

A THESIS BY
Areti Psaropoulou
2222922

Project supervisor Dr. Itir Önal Ertuğrul
Daily supervisor Mang Ning
Second examiner Prof. dr. Albert Salah

Utrecht University

Abstract

Depression is a global health issue affecting individuals coming from all age groups, with the prevalence in children rising. The early detection of depression in young ages is crucial, yet several obstacles hinder the accurate assessment of depressive symptoms in young populations. This study proposes a multimodal deep learning framework for the automated assessment of depressive symptoms in children. We extracted audio and text features from videos of parent-child interaction, to train and evaluate deep learning models. Our methodology involved the use of advanced feature extraction techniques, including Wav2Vec2.0 and CLAP for audio features, and RobBERT and SBERT for text features. In addition to examining each modality independently, we explored how multimodal fusion could enhance the accuracy of detecting depressive symptoms in children. This study indicates that multimodal deep representations can effectively identify depressive symptoms in children, particularly in contexts involving cooperative tasks. The combination of SBERT and CLAP feature representations yielded an AUC score of 0.810 in the cooperative scenario. This result serves as a strong foundation for exploring the complex process of assessing depressive symptoms as more data becomes available.

Table of Contents

1. Introduction

Depression has emerged a global health problem, affecting a great percentage of the global population, estimated at 280 million people[1]. This demographic includes people of all ages, from young children to old adults. However, there is a concerning trend of an increasing prevalence of depression in children, especially young girls [109].

Prevalence of depression in children presents a wide range of difficulties, significantly impacting their lives at an essential point in the development of their mental health. During this time, the human brain is developing and becoming more complex, and it is also the time when social, emotional, and cognitive skills are developing. These changes may create a basis for later in adulthood the emergence of probable psychopathologies [59]. Mental health and psychosocial development is of high importance especially for young people and neglecting it is not only limited to discomfort during childhood, but may also lead to limited opportunities during adulthood.

WHO estimates that 10% of children and adolescents globally suffer from mental problems, with many of them not seeking help or receiving proper care, despite the need for early detection and intervention. It is also reported that suicide ranks as the fourth leading cause of death for adolescents aged from 15 to 19 years old. Thus, assessing child depression is an emerging issue and may prevent many problems both for individuals and society.

Despite the pressing need for assessment, numerous obstacles block successful detection of depressive symptoms in children. These barriers include a financial burden, unreliable parental reporting, a shortage of qualified staff and screening tools, among many others [63]. In addition, detecting discomfort and irritated mood in children is challenging, even though it is reported that almost 1 out of 5 children experience a disorder during childhood. Hence, it is crucial to develop screening tools that can detect depressive symptoms and lessen the burden faced by public health providers, at least regarding the initial assessment.

This task has become popular in the machine learning community; many researchers have already explored adult depression assessment based on visual, audio and linguistic cues. However, there remains a gap in research concerning child depression and the detection of its symptoms, presenting an opportunity for further advancement.

[1] World Health Organization, `https://www.who.int/news-room/fact-sheets/detail/depression`

This study aims to address this gap by leveraging multimodal features extracted from parent-child interaction videos and deep learning models to detect depressive symptoms in children. The main research question this study poses and attempts to answer is the following: "To what extent can multimodal embeddings obtained from large audio and language models detect depressive symptoms in children from parent-child interaction videos?"

2. Research questions

The automated evaluation of depressed symptoms in children is essential in several domains, including the psychological, social and financial. As a result, this study's research and sub-research questions, as outlined below, will attempt to assess this need. More information on the subjects of depression, depression in children, and the evaluation of its symptoms using both conventional and machine learning techniques can be found in Section 3. Section 5 of this proposal will cover the strategies and tactics for answering the questions.

Research question: To what extent can multimodal embeddings obtained from large audio and language models detect depressive symptoms in children from parent-child interaction videos?

The main research question outlined above provides a broad foundation for this study and will be further expanded upon in the experimental phase. The primary objective of this study is to leverage valuable embeddings extracted from large audio and language models to evaluate depressive symptoms in children. Specifically, these foundation models will be supplied with audio and language data taken from the parent-child videos dataset.

Sub-research question 1: To what extent can depressive symptoms in children be assessed separately in both cooperative and conflicting scenarios?

This sub-research question aims to investigate the degree to which depressive symptoms can be identified during interactions between a parent and a child in cooperative and conflicting situations separately. In this case, the assessment will rely on embeddings extracted from data related to each scenario independently, trying to determine whether depressive symptoms differentiate depending on the nature of the task.

Sub-research question 2: To what extent can depressive symptoms in children be accurately assessed using solely the audio and text embeddings?

While the main research interest of this study is based on multimodal embeddings, it is interesting to examine the individual efficacy of audio embeddings and text embeddings in the assessment of depressive symptoms. To address the aforementioned question, predictions regarding depressive symptoms will be made based on unimodal embeddings only, entailing that predictions will be made by exclusively audio embedding and separately from text ones.

Sub-research question 3: How does examining only the parent's segment impact the assessment of depressive symptoms in children?

Another intriguing topic that the sub-research question above examines is the impact of parent segments. Orvaschel et al. [75] was one of the first studies associating the depressive symptoms in children with depressive parents. The children of depressives are at a higher risk of developing psychopathology, while getting exposed to a vulnerable environment. Therefore, embeddings extracted from the parent segments will be compared with those extracted from child's segment exclusively.

3. Literature review

3.1 Depression

Commonly referred to as major depressive disorder, depression is a common mental disorder. It is estimated that this mental illness affects 5% of people worldwide [98]. According to Kessler et al. [51], this condition is popular throughout all nations where epidemiological surveys have been conducted, with a higher prevalence in countries with higher incomes. Thus, treating depression is extremely difficult, especially in low- and middle-income nations where 75% of people with mental diseases do not receive treatment, according to the World Health Organization (WHO)[1] [98].

Depression has a profound effect on people's lives that goes beyond mental health. These aspects include relationships with relatives and friends, community involvement, academic performance, and productivity at work. Furthermore, there is a connection between physical health issues like cardiovascular disease and depression. According to estimates, depression costs the US economy $43 billion annually as it causes higher medical expenses, early deaths, and lower productivity at work [37].

The depression construct can be approached from different perspectives. Since it offers a framework for the diagnosis of mental disorders, including depression, the Diagnostic and Statistical Manual of Mental Disorders [6], Fifth Edition, also known as DSM-5, is regarded as a major foundational work in the field of mental health. According to the DSM-5, depression is a unidimensional construct, with the presence of a depressed episode being determined by the combination of symptoms. According to DSM-5, one of the indicators for a Major Depressive Episode (MDE) expects the presence of at least five from the subsequent symptoms for a timeframe of minimum of two weeks. Among the symptoms should be a depressed mood, which is characterised by a pervasive feeling of sadness, emptiness, or hopelessness that is noted not only by the patient but also by others. Anhedonia, another potential symptom, is characterised by a decreased interest or pleasure in nearly every aspect of life. Both disruptions in the sleep-wake cycle, which may lead to insomnia or hypersomnia and changes in one's appetite and weight are potentially symptoms of the depressive disorder, as well. Additional symptoms include fatigue or low energy, psychomotor agitation or retardation, and these may significantly impact day-to-day functioning. Moreover, individuals may experience poor concentration, difficulty

[1] World Health Organization, `https://www.who.int/news-room/fact-sheets/detail/depression`

making decisions and feelings of worthlessness, or excessive guilt. Last but not least, suicidal ideation—whether planned or not—as well as actual suicide attempts may serve as early warning signs of a depressive episode. According to this manual, the symptoms that have been recorded earlier indicate a major depressive episode, if they cannot be attributed to any substance or medical condition.

While alternative perspectives question the unidimensional model, one such approach claims that depression is better characterized by a two-factor model [28], where symptoms can be classified as somatic and non-somatic. In this framework, somatic symptoms encompass issues like sleep disturbances, changes in appetite, and fatigue, among others. Conversely, non-somatic symptoms, as noted by Elhai et al. [28], include depressed mood, feelings of worthlessness, and suicidal ideation.

In conclusion, depression is characterised by enduring melancholy and a loss of interest in or enjoyment from previously enjoyed activities. In addition, it can cause fatigue and interfere with eating and sleep cycles. The World Health Organization [98] lists excessive guilt, hopelessness, poor focus, and suicidal thoughts as additional symptoms. Numerous social, psychological, and biological factors, such as loss experiences, unemployment, or childhood adversity, might contribute to these symptoms.

In addition to the symptoms previously discussed, individuals with depression often present observable symptoms. According to Sobin et al. [105], psychomotor symptoms, though significant indicators of depression, can be challenging to identify. Depressed subjects usually exhibit lower activity than healthy groups of people, with the level of activity varying depending on the subtype of depression. Interviews have revealed specific patterns in behaviours such as direct eye contact, self-touching levels, and eyebrow movement among depressed subjects. These groups tend to make less direct eye contact with the interviewer, exhibit reduced eye and eyebrow movement in comparison to healthy ones. When it comes to self-touching, it was observed more in depressed patients, while their motor speed when they had to respond in a simple task was slow. Caligiuri et al. [15] underline the significance of motor retardation as a depression symptom. An interesting finding presented in the study was the inability of depressed subjects to increase movement velocity in a specific task even when the target distance was decreased. To conclude, these observable symptoms highlight the complexity of the condition and the significance of thorough diagnostic and treatment approaches, offering insightful information about how depression emerges itself.

3.2 Depression assessment

Given the increasing prevalence and varied impact of depression on individuals and society, there arises a critical need for assessment methods to effectively address this widespread mental health condition. Several methods are used to evaluate depression, particularly in primary care, where general practitioners and primary care providers deliver most of the treatment for depression [115; 66]. A standardized evaluation tool for mood, anxiety, somatoform, alcohol-related, and eating disorders is the Patient Health Questionnaire-9 (PHQ-9). It consists of nine questions that are based on the previously described DSM-5 symptom criteria. The scoring method used to determine the findings ranges from 5 to 20; 5, 10, 15, and 20, with each score denoting a different level of severity—"mild," "moderate," "moderately severe," or "severe." The Hospital Anxiety and Depression Scale for Depression (HADS-D) is an additional evaluation tool that is used only when the subjects are medically unwell patients; as a result, somatic aspects are not measured. The updated Beck Depression Inventory (BDI-II) is an alternative assessment instrument that is more appropriate based on the most recent definition of major depressive disorder. Subjects respond to BDI-II based on their last-two-weeks' symptoms.

Although questionnaires and self-assessment tools are the main techniques used to identify depression, they do not take into account observable symptoms. These techniques produce subjective results that could be misinterpreted because they only take into account the symptoms that patients, family members, and carers describe. Despite its importance in detecting depression in people, measures including movement of the body, gaze, and facial expression are not included in self-assessment tools [60]. A multimodal approach to depression assessment is becoming more and more necessary, as the majority of current techniques concentrate only on clinical symptoms. Including a behavioural viewpoint in the evaluation of depression might improve overall diagnosis accuracy and provide a new understanding of the condition.

3.3 Depression in children and adolescents

Depression is a common disease among children and adolescents as well. In children between the ages of 10 and 19, mental disorders affect about 1 in 7 of them. Twenty percent of teenagers may suffer from a mental health issue, according to Mental Health Foundation [64]; depression and anxiety account for half of these cases, according to UNICEF [2]. According to Moreau [67], it was acknowledged as a clinical and diagnostic entity as early as the 1990s.

[2]United Nations Children's Fund (UNICEF), `https://www.unicef.org/eu/press-releases/worsening-mental-health-situation-europes-children`

The DSM-5 [6] criteria for major depressive disorder state that symptoms can differ in children and adolescents. The DSM-5 lists sadness as the initial symptom in case of adult depression, which in case of children may show as irritable mood. It is sometimes regarded as an indication of children depression when a child fails to gain the anticipated amount of weight for their age and developmental stage. In addition to irritation, children may also be more prone to physical symptoms and social disengagement, according to the study by Korczak et al. [56]. This demonstrates the advancements made in the last many years in the study of childhood depression. A few decades ago, experts [67] said that children were evaluated for depression using the same standards as adults. There are still some who support this view. Bernaras et al. [10], for example, believe that the construction of childhood depression is the same as that of adult depression. Conversely, Neto et al. [92] assert that depression in childhood is a separate illness from that of adults.

According to a study by Chen et al. [16], parent-child and family relationships are crucial in the development of depressive episodes. Living in an unfavourable environment can cause family members to experience melancholy, particularly when it comes to the family, which is seen to be one of a child's most significant influencers [104; 107]. Living and developing inside a system, particularly the family, can cause people to experience problems that mirror those that are already present in the family system. Given that many children view their parents as their significant others [62], parenting styles and the messages parents pass to their kids have a big impact on how they develop now and in the future. The relationship between parenting styles—those of both mothers and fathers—and how they affect children's symptoms of depression is the main focus of the study conducted by Chen et al. An intriguing discovery is that girls are more impacted by messages from their mothers and males by words from their fathers.

After reviewing over the various viewpoints on childhood depression, it is clear that treating this mental condition is of vital importance given the possible effects it may have on kids' growth. Childhood depression is extremely important since it is a condition that can harm a child's development as they grow older [92]. It may interfere with their long-term, persistent symptoms, which can impair their social, emotional, and cognitive development. Deep alterations in the child's brain area are part of cognitive growth, so early trauma may cause insufficient brain development. Furthermore, according to WHO [3], childhood and adolescent development of cognitive and social-emotional skills is crucial for eventual maturity and societal duties. Korczak et al. [56] point out that depression is a common adolescent disorder in Canada, which has a long-term negative impact on life quality. Additionally, they note that the majority of depressed teenagers do not receive treatment,

[3]World Health Organization, https://www.who.int/activities/improving-the-mental-and-brain-health-of-children-and-adolescents

which highlights how difficult it is to identify depression in these vulnerable age groups.

3.4 Assessment of depression in children and adolescents

Following the recognition of childhood depression as a separate illness and the increase in the prevalence of depressive symptoms in youth, concerted efforts have been made to create a variety of evaluation tools that are geared towards this population. A common measuring instrument for evaluating depression in children is the Children's Depression Inventory (CDI) [29]. It was developed using the BDI, a measure for evaluating adult depression. In addition, a number of other instruments have been developed with only the intent of assessing childhood depression. These include the Rating Scale for Children (DSRS) [12] and the Children's Depression Scale (CDS) [57], the latter of which is also capable of measuring the disorder's severity. However, these child depression assessment questionnaires frequently ignore visible cues such changes in body language or facial expression, much like adult assessment tools.

4. Literature Review: Automated depression detection

Several works on automatic depression recognition and assessment have been presented in the literature. This field has evolved significantly since 2011, particularly with the introduction and successive editions of the Audio/Visual Emotion Challenge AVEC [96].

Typically, depressed individuals tend to change their expressions at a very slow rate and deliver flat sentences with stretched pauses. Similarly, their language usage gets affected, characterized by specific linguistic patterns. Therefore, for detecting depression, audio and linguistic features are frequently leveraged due to their ability and consistency to reveal signs of depression. Numerous proposed approaches follow a similar workflow, including four main processing steps: preprocessing, feature extraction, dimension reduction, and classification.

Sections 4.1 and 4.2 review papers that use acoustic and linguistic features for assessing depression. Combining more than one modality and its significance for resolving the task of automatic depression assessment are discussed along with the corresponding research in section 4.3. The works under review cover a wide range of methods and models, including hand-crafted feature extraction and the latest advancements in deep learning technology, with foundation models included.

4.1 Acoustic modality

Speech communication is widely considered as the cornerstone of human interactions [55]. Robust speech representations are useful for addressing a range of paralinguistic tasks, which are connected to non-linguistic occurrences [91]. Therefore, acoustic features are a great base for analyzing human behavior and earning valuable information from non-linguistic occurrences. Features extracted from audio can be important indicators of the absence or not of a mental health condition [52]. Since speech is linked to such conditions and the quality of speech can be affected by them. Research has explored the differences in speech patterns between individuals with depression and those mentally healthy, with those who have depression exhibiting monotonous, dull, and lifeless speech [34; 22]. Since these changes have been quantified, it is now possible to identify depressed speech by taking into account particular speech characteristics like loudness, pitch, and harshness [34; 22]. Therefore, in the topic of depression assessment, the extraction and analysis of audio features are essential.

The first part of the research explored acoustic modality using hand-crafted features. Hand-crafted feature extraction mainly involves extracting fundamental audio features such as frequency, jitter, shimmer, loudness often achieved with the use of tools such as openSMILE [32]. These features have also been extracted in the study by Smith et al. [103] from interviews, in which 65-year-old or older aged participants engaged in two specific tasks. Predictions regarding whether a participant was depressed or not were created using an SVM fitted with repeated 10-fold-cross-validation. This approach's accuracy showed that the machine learning models leveraged, were able to distinguish between depressed and non-depressed older adults.

Similarly, Pan et al. [76] have extracted several hand-crafted features from audio interviews. They used 26 physical features commonly used in emotion recognition; including loudness, F0, intensity, zero-crossing rate. They have shown that voice features are crucial indicators of depression prediction. However, it is important to note that their study was limited to female participants, making it difficult to generalize without further research.

Another intriguing study was conducted by Pessanha et al. [82], in which breathing patterns serve as the features. The features, extracted from audio sequences, are utilized to predict depression severity. The corpuses used were UCL Speech Breath Monitoring [97] – including spontaneous speech instances, along with the Distress Analysis Interview Corpus-Wizard-of-Oz (DAIC-WOZ) [38] - consisting of interviews aiding in the diagnostic procedure of mental disorders. Both linear and non-linear models have been trained trained in the AVEC 2017 challenge, with random forests demonstrated superior performance.

Seneviratne et al. [99] directed their focus towards articulation and how this may lead to conclusions for depression prediction. It is worth mentioning their use of Mel Frequency Cepstral Coefficients (MFCCs), which are low-level audio descriptors commonly used in various audio processing tasks, as they imitate how speech signal is perceived by human brain. A model was trained using coordination features extracted from MFCCs and vocal tract variables and then were classified using a Support Vector Machine classifier. This study has reported the highest accuracy, when glottal tract variable is also included along with articulatory coordination features.

While deep learning models are increasingly dominating machine learning tasks, many applications have been proposed to automatically assess depression. Although numerous features can be extracted from audio, not all are equally significant for depression assessment. An effort to reduce high dimensionality and avoid overfitting caused by extraction of many speech features was made by Sardari et al. [95]. They proposed an adaptation of a CNN AE (auto encoder), which reconstructs the input signal in a way that dimensionality is

effectively reduced. The full proposed model can be seen in 1. Then an SVM classifier - the best performing among four classifiers – was employed to distinguish between depressed and non-depressed subjects.

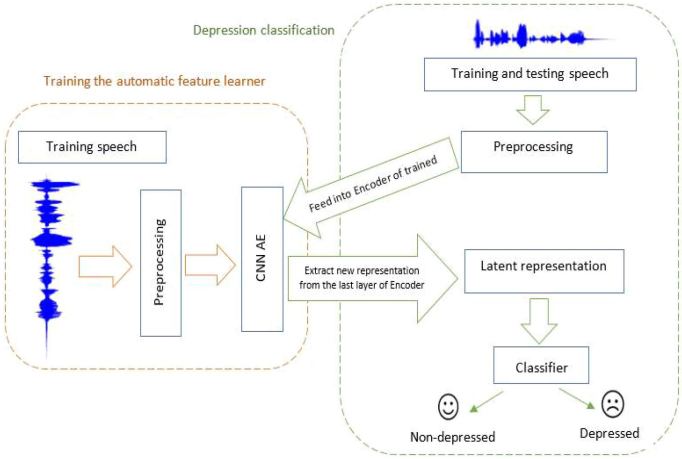

Figure 1. Framework for assessing depression with audio features by [95]

Rejaibi et al. [88] used LSTM [42] in their study, a recurrent neural network good at capturing long-term dependencies in sequential data, thus ideal for solving audio processing tasks. Following data preprocessing and extraction of MFCC from the audio interviews, they extracted high-level audio features by feeding the LSTM with depression data. Finally, features from the LSTM layers were passed through two fully connected layers and a sigmoid function, which predicted the depression severity. As an extension to their experiment and due to scarcity of depression-related data, they used transfer learning strategy. The network was initially trained in a similar task, that of emotion recognition, with the weights obtained after the emotion recognition task served as the initial point for fine-tuning the model on the depression recognition task. The accuracy achieved by the transfer learning strategy surpassed the accuracy of the model when not pretrained on the emotion recognition task.

Zhao et al. [124] have introduced a hybrid network, comprising two different deep learning mechanisms: self-attention networks and deep convolutional neural networks (DCNNs), both trained on related tasks. The feature representations obtained from these models are concatenated and average pooling is applied. The final layer consists of a support vector regression, which produces a score in the range of Beck Depression Inventory-II score. Remarkably, the combination of the two models performs better compared to each model individually. Evaluation and testing of the models were conducted in AVEC 2013 [112] and AVEC 2014 [111] datasets.

4.2 Linguistic modality

This section focuses on linguistic modality, as language is a significant factor in depression assessment. Language can serve as a valid marker for psychopathology and a useful tool for depression diagnosis; both language and verbalism are important for depression diagnosis [110]. Clinical psychologists have conducted research revealing specific linguistic patterns observed in the speech or writing of depressed individuals [93]. Examples of such patterns include the use of first-person singular pronouns and a lack of positive words, particularly when referring to depressed children [18]. As a result, studying linguistic features, such as words, parts of speech, n-grams, negation, as well as syntactic and grammatic choices would contribute to the automatic assessment of depression.

Natural Language Processing (NLP) is a machine learning technique which utilizes algorithms for analyzing textual data, thus it offers the ability to computers to understand natural language. Common NLP tasks are machine translation, language generation, classification and prediction based on specific tasks and concepts. In the case of depression assessment, the extraction of text features and evaluation of them could add value in automated assessment of depression. The studies addressed in the following section are using only text data for assessing the depression detection task combined with both baseline classifiers, such as SVM and deep learning models, like GPT-4 [72].

Recently, there has been an increase in the platforms where people can share their thoughts, chat with people, seek for emotional attention and express their feelings online offers an enormous amount of text data and a great opportunity to explore the assessment of depression through text. Hence, there have been already works on automatic depression assessment in the direction of text modality that use mainly social media extracted data. A very recent work by Tejaswini et al. [108] have used data extracted from Reddit and Twitter; sources extremely popular for depression assessment tasks. 6.164 records are preprocessed by tokenizing, lower casing, removing stop words, performing stemming and lemmatizing. After that fastText [49] embeddings are created and are fed to the model. This study proposes a model consisting of an LSTM [42] and a CNN that outperformed other works in the same dataset.

Another study by Singh et al. [2] utilized a similar dataset extracted from Twitter, employing preprocessing techniques such as removing the punctuation and data normalization. After that feature extraction techniques were applied to the preprocessed dataset, such as Bags-of-n-grams and glove embedding-based feature extraction. The features extracted in the previous step were considered for a weighted feature selection and were fed into the proposed CNN-Ensemble learning. This model consisted of a CNN that its last classifica-

tion layer was replaced by different classifiers, such as Support Vector Machine, Random Forest, Long Short Term Memory Network and Artificial Neural Network, achieved greater results in the three datasets tested when compared with existing models.

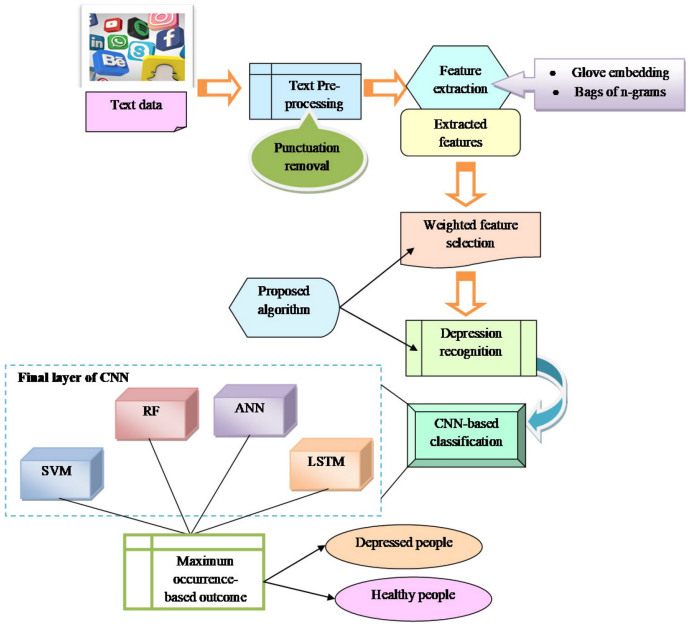

Figure 2. Framework for assessing depression with textual features. Figure by [2].

Additionally, there has been an increase in the use of large language models (LLMs), especially since the introduction of models such as ChatGPT. Agrawal's [102] work combines both conventional machine learning and large language models (LLMs) for depression detection and provides the users with personal explanations, as well. This study leveraged both clinical interviews and Reddit forums for feeding various conventional machine learning models including SVM and deep learning models including CNN, BERT-CNN, and LSTM to detect depression. The next step in the workflow included some preprocessing steps, such as tokenization, stopword removing and transforming the text to word embeddings. Two explainability methods, namely SHAP (SHapley Additive exPlanations) and LIME have been employed for presenting the result in a human-interpretable way. The second part of the study utilizes the integration of three advanced LLMs; specifically GPT-4 [72], Llama 2 [31] and Gemini [35]. The main goal of this approach is to provide users with similar to human explanations regarding the diagnosis produced by the AI tool. A prompt interface which asks questions, tries to understand the user, and detects depression symptoms in accordance with the DSM-5 [6] was used to interact with the user.

Using data from Twitter, Chiong et al. [19] trained several classifiers, such as Logistic Regression, Support Vector Machine, Multilayer Perceptron, Decision Tree, Random Forest, Adaptive Boosting, Bagging Predictors, and Gradient Boosting and evaluated their

performance using data retrieved from other online platforms. Preprocessing was carried out in the same manner as in previous studies, including Bag-of-Words feature extraction combined with count vectorization and n-gram words as input features for the machine learning models. The primary objective of this research was to identify depression in participants who might not be aware of or deny their depressive state, with the best results achieved by Logistic Regression classifier.

In addition to social media posts, content of speech from patient-therapist interactions has been used to assess depression. Milintsevich et al. [65] provided an alternative viewpoint on the research on depression detection by attempting to predict depression severity scores for each symptom in a way that offers a more thorough understanding of the mental health state. In this case, the dataset consists of the interviews between patients and Ellie [120];Ellie serves as the interviewer and was a virtual assistant functioning as a therapist, who asks questions in a non-judgmental manner to encourage subjects to respond. Therefore, the dataset used was DAIC-WOZ [38], containing clinical interviews between Ellie and the subjects. In order to predict the scores of individual symptoms, utterances were encoded both per dialogue turn and per dialogue level. These encoded transcripts were then fed into a multi-target regression model. For encoding, deep learning models like Bidirectional LSTM (BiLSTM) [47] and S-RoBERTa [86] were used. Although its limitations, the present study is consistent with the latest methodologies suggested by psychiatrists and could serve as a basis for future research along these lines.

In the study conducted by Havigerova et al. [40], the dataset is different than what has been discussed so far. It consists of fictive letters of 180-200 words that participants wrote using predefined scenarios, such as a cover letter, a letter from holidays, a letter of complaint, and a letter of apology. Various morphological features were detected through lemmatization and morphological tagging of the 688 texts. Following a filtering step and two steps of variable reduction, eight logistic regression models were created corresponding to different text genres and genders. The prediction primarily relied on the linguistic properties of the subjects' written text, focusing on quantitative linguistic syntactical and morphological variables rather than content and semantics, as commonly seen in similar studies. Several patterns in the usage of morpho-syntactic elements in the texts were observed among women and men, explaining the differences found between the two groups. Notably, informal letters discussing holidays were found to be of high value in depression detection.

4.3 Multimodality

The preceding sections discussed the use of unimodal models based on either speech or text for depression assessment and their performance in depression detection tasks. However,

there are advancements in the field, involving the combination of different modalities, so a multimodal model is formed combining all the different modalities. Integrating various modalities has shown promising results in research and especially in depression detection, as explored in the following section.

The world is multimodal as well, meaning that there are several ways to experience and perceive one's surroundings. For instance, anxiety can be expressed by a trembling voice, a high heartbeat, sweat or even difficulty in sleeping. All these and even more independent pieces construct only one feeling and capturing them offers a better understanding of how humans experience not only their environment, but also themselves. Similarly, depression is multi-faceted disease that has many different symptoms and studying different modalities would give great value to the depression assessment task.

4.3.1 Multimodal fusion

The procedure that two or more modalities are combined and return one unified prediction is called fusion. There are a number of ways to combine two modalities, the first one is called early fusion [118]. When features are concatenated immediately after they are extracted from each modality it is known as early fusion. It is considered an attempt to make the model learn in a multimodal way, as finally only one single model is trained, in which several modalities are embedded.

An alternative approach known as late fusion unifies the independent predictions made by each modality. The fusion of the separate decision values from the unimodal models can be achieved in multiple way, namely by averaging, applying weighting voting schemes or weighting based on channel noise. Late fusion provides flexibility, as each decision may come from a different predictor even in the absence of a modality.

Early and late fusion components are combined in a third strategy called the hybrid fusion method. This method tries to combine the advantages of both approaches by merging the output from early fusion with separate unimodal predictors, producing an enriched multimodal model.

Related work on Multimodal assessment of depression

As mentioned earlier, there has been some advancement in multimodal architectures for depression detection. Some recent research that achieves impressive results will be examined, along with details about the feature extraction, training of the deep learning models and the type of fusion performed. In a very recent study by Aloshban et al. [5], interviews of both depressed and non-depressed subjects are utilized for training and evaluating BLSTMs [47] both in a unimodal and a multimodal approach. In the multimodal approach, there are three different ways that the modalities are combined in

this study; namely late fusion, early fusion and feeding the unimodal representations to a gated multimodal unit. The multimodal approach performed significantly better than unimodal ones, achieving an accuracy equal to 84.7%. In addition, an interesting finding occurred, when compared acoustic and linguistic modalities individually, acoustic features are more valuable for depression specific information.

In their study, Shen et al. [100] developed EATD-Corpus dataset, which was the first one existing Chinese depression dataset, containing text and audio data extracted from interviews where participants respond to three questions. A GRU [20] model and a BiLSTM [47] with attention were trained on audio and text data, respectively. The last layer representations of the two models are horizontally concatenated, and the next step includes the calculation of the dot product between the concatenation and the specially trained weight vector on the different modalities. The final prediction was made using a fully connected layer, taking the dot product as input. Again, the multimodal approach achieves a better F1 score than single-modal methods. Also, the proposed fusion method outperformed the one proposed by Alhanai et al. [3], which utilized a multimodal LSTM model that took both modalities as an input to the two-branched model.

Taking advantage of the progress made in multimodal approaches to depression detection, Ya et al. [52] investigated the use of audio and text features to detect depression in a reading experiment with 160 participants. This experiment was called Segmental Emotional Speech Experiment (SESE) and examines audio features, text features and how these perform when fused. A big number of audio frames were produced during the DeepSpectrum audio feature extraction process, which got eliminated by averaging the features extracted per second; therefore there was a dimensionality reduction. In order to capture emotional information, DeepSpectrum features were fed into a Temporal Convolution Network with attention layers (Att-TCN) [44], after being passed through an AutoEncoder. Trying to extract information to the greatest extent, an "one-hot Transformer", able to catch long-term dependencies was utilized and model's input was the combination of text features and the corresponding location of the words. This model was proved to be by 0.7% more efficient than the original Transformer model. The fusion of the models occurs at decision level, as the features of the two models are concatenated and passed through dense layers, with softmax producing the final result, as this can be seen in Figure3. The accuracy of the fusion model is higher than the results obtained by employing each of the modalities independently.

Makiuchi et al. [91] have trained their model on data sourced from the AVEC 2019 DDS Challenge database, particularly the DAIC-WOZ dataset [38]. This research focuses on two modalities: audio and text, each leveraging different architectures. For the audio

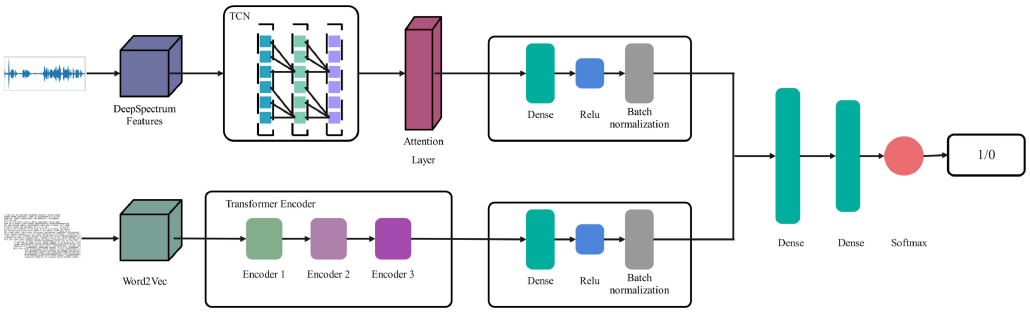

Figure 3. Fusion model of audio and text features. Figure by [121].

modality, raw audio was being preprocessed and was transformed into spectrogram images, serving as input for a pretrained VGG-16 network [101]. Since the raw audio files varied in duration, zero padding was applied to standardize the input features' lengths. As far as the linguistic modality is concerned, features are extracted from the last layer of a pretrained BERT-large model [25]. Each word token is represented by a feature array of size 1024, resulting in a transcript of K words being represented as K x 1024. These features served as the input into two different models: an LSTM-based model and a CNN-LSTM-based model. Visual features are taken from a ResNet-50 model [41] and used as input in a model like the GCNN-LSTM model in order to investigate the visual modality as well. Not only are all the fused models examined for the three distinct modalities, but also for pairs of modalities. Combining text and audio yields the best results, with a CCC score of 0.403 in the test set and 0.696 in the development set. It's interesting to note that combining all three modalities results in a lower score than combining text and audio, which illustrates the worth of the latter two combined. The modalities are fused at a feature level, meaning that extracted features from each modality independently are fed into the multimodal network, after being concatenated.

The study by Bilalpur et al. [11] examines how depression in mothers may affect the occurrence of depressive symptoms in adolescents. Except for manual data annotation of both verbal and nonverbal behavior, multimodal cues were extracted from the video-audio recordings. Multimodal cues, namely measures of facial expression, face and head dynamics, prosody, speech behavior and linguistics, were extracted from the adolescents in the audio-video recordings, where mothers had a fifteen-minute problem-solving session with their child. In addition, data was annotated manually by expert coders examining both verbal and nonverbal behavior. Mother-child dyads were chosen based on the mothers' depressive state and the dataset consisted of balanced depressive and nondepressive dyads, thus features are extracted only by the adolescents. For the classification, SVMs were leveraged and fed with either the concatenation of the unimodal features or combining the decision of the trained classifiers with the weighted sum criterion for early and late fusion, respectively. Given the plethora of features, SHAP was introduced for feature selection

and presented different results depending on the modality, e.g. a four-fold decrease in the number of features for prosody. Computational measures outperformed the manual annotations not only, in terms of accuracy, but also in interpretability by providing insight about the features. They concluded that maternal depression had an impact in adolescent behavior, with prosody and multimodal features presenting superior predictions.

One of the earliest and most significant works in the multimodal assessment of depression was conducted by Dibeklioğlu et al. [26]. Depressed participants were interviewed at four specific time intervals following their diagnosis, during which three separate cameras were employed: two recorded the participant's face and body, while the third captured the interviewer. This study examined visual and auditory features, particularly examining facial and head movements visually, as well as vocal fundamental frequency and pauses in voice. Facial and head movements were analyzed and encoded per frame, thus emerged the need of combining them; they were combined by training Stacked Denoising Autoencoders (SDAE) [113]. In terms of the audio analysis, preprocessing was carried out prior to feature extraction and the final features were selected using the Min-Redundancy Mac-Relevance (mRMR) [85] algorithm. For both the single modality and the multimodal cases, logistic regression classifiers were used to classify the severity of depression. The three modalities—voice, face, and head—had their feature sets fused together at the feature level, leading to a higher level of accuracy than when the modalities were used separately or when combined by two. This study moved one step further, as it did not just predict the presence of depression or not, but predicted depression severity, resulting in three depression classes; moderate to severe depression, mild depression and lower indicate remission for cases that recover from depression.

This study discusses and mainly focuses on the fusion of audio and text modalities and the outcomes that can be obtained by this approach. However, numerous works go beyond this fusion and combine various modalities such as image, text, audio, sentiment analysis and combinations among them.

4.3.2 Foundation models

Foundation models, with their large scale and extensive pretraining on a variety of datasets, signify a paradigm shift in artificial intelligence. These models, which include GPT [72], BERT [25], CLIP [83], and more, have proven to be quite good in creating, understanding, and adjusting to diverse kinds of content in a variety of domains. Their training process is based on scaling, which boosts their performance in several tasks, not limited to the ones they were trained on. Scaling and training in large datasets are used to train pre-existing deep neural networks to discover broad patterns and meanings. Furthermore, these models perform exceptionally well on activities that they were not explicitly pretrained for as well

as new tasks, demonstrating their adaptability and relevance in a variety of contexts.

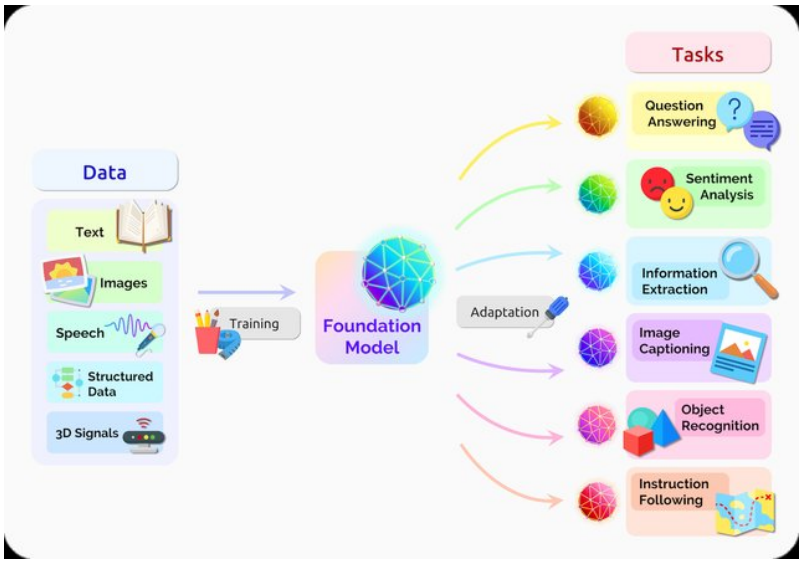

Figure 4. Multimodality in foundation models. Figure by [13].

The development of foundation models has greatly advanced a number of fields, including linguistics, vision, thinking, and human interaction. They possess remarkable abilities in language understanding, visual comprehension, reasoning, and even influencing physical environments. Foundation models have advanced to the forefront of AI research and application thanks to their excellent skills in creating, understanding, and modifying a wide range of content as well as creating new ones across a wide range of fields.

Related work on Foundation Models

Building upon the progress of foundation models, Kim et al. [54] propose an automated audio captioning (AAC) framework that combines EnCodec and CLAP for acoustic processing and utilizes the pretrained BART model [58] for language comprehension. Depending on the BART model version used, this model is available in two versions: enCLAP-large and enCLAP-base. It has been tested on both the AudioCaps and Clotho datasets. Significantly, enCLAP-large outperforms previous benchmarks in AudioCaps, setting a new state-of-the-art; on the Clotho dataset, its performance varies based on the training data sources, exhibiting better results in terms of AAC when trained on both Clotho and AudioCaps than when trained only on Clotho.

A multimodal study in progress for prompt-based speech generation and comprehension was proposed by Huang et al. [45]. A distinctive architecture is created by fusing audio processing models with Large Language Models (LLMs) to combine text and audio modalities. Using already existing models of the audio foundation, LLMs serve as the interface, allowing input and output to be seamlessly integrated. Figure5 illustrates the four steps of the AudioGPT pipeline, which provide a variety of transformations, such as

Audio-to-Text, Audio-to-Audio, and Audio-to-Video. Of special relevance to this study is the "Audio Caption" function from Audio-to-Text.

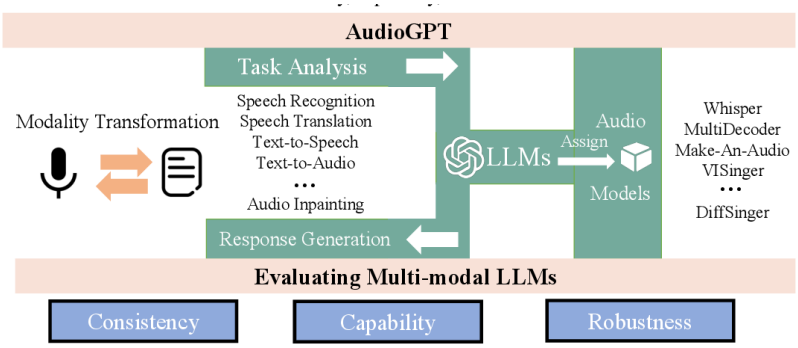

Figure 5. AudioGPT pipeline. Figure by [45].

Deshmukh et al. [24] created Pengi, which is a cutting-edge audio language model that approaches all audio issues as problems involving text. Pengi operates on text and audio instructions and outputs free-form text. Four separate steps make up its architecture, as shown in Figure 6. Pengi's training data is built using a predefined template, which makes use of numerous audio datasets. Examples of this template include Task: "speech emotion recognition," Input prompt: "this emotion is," and Output format: "emotion." This template-driven technique enhances the results obtained for both closed- and open-ended assignments. The audio encoder-audio transformer component from CLAP [30], which translates audio input into an audio embedding, is the first step in the model pipeline. The text prompt is simultaneously converted into an embedding by a text encoder; different text encoders are available for this purpose. The two input embeddings are then mapped, resulting in the construction of two mapping networks that translate each input into a series of k embeddings. The autoregressive causal language model receives these sequences as input when they are concatenated to create a fixed-length prefix. Both closed-ended tasks that produce specified values, classification results, or input retrieval, and open-ended tasks that generate text output are handled by the model during training. Pengi outperforms supervised learning methods in terms of CIDEr and SPICE when it comes to open-ended tasks like audio captioning, exhibiting state-of-the-art performance.

In their research, Xu et al. [119] proposed a framework for speech emotion captioning (SECap), in which they employed two significant models one for each modality, LLaMA [31] for text and HuBERT [43] for speech, respectively. HuBERT [43] served as the encoder in SECap, and was used for speech feature extraction. Speech features extracted from HuBERT [43] were then compressed by Querying Transformer(Q-Former) that was standing as a Bridge-Net. Bridge-Net aimed to eliminate the features extracted only to the ones that were related to speech emotion and excluded those relevant to content, the latter could be extracted by the transcript. At the same time, LLaMA [31] was introduced

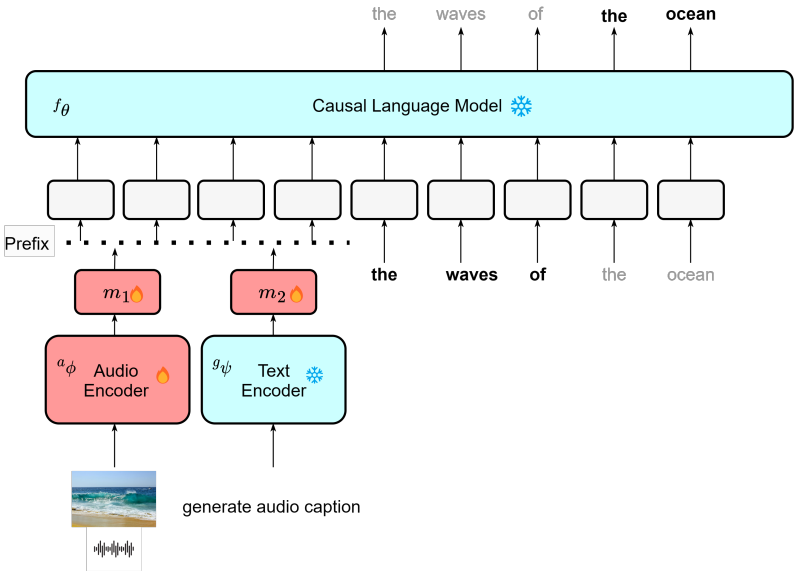

Figure 6. Pengi architecture. Figure by [24].

in the training process of Q-Former, trying to enhance emotion projection in terms of text. Q-Former aimed to extract the emotion-related features from the speech features extracted from HuBERT [43], by utilizing an attention mechanism resulting to a fixed-length representation regardless of the input length. The training process consisted of two stages, one included the compression of the HuBERT's features, so they obtained relevance to emotion, followed by their alignment with LLaMA's representation space. Results were evaluated to both objective and subjective metrics introduced in this work, as Speech Emotion Caption task was a novel task. SeCap not only outperformed earlier works when objectively compared with them, but it also produced emotion captions of comparable quality with those created by human annotators.

5. Methodology

5.1 Dataset

The data used in this research originated from a larger dataset consisting of recorded parent-child interaction videos. This dataset was part of YOUth study [71], an accelerated population-based longitudinal cohort study that aimed to predict and gain insight into the factors explaining individual variability in the development of self-regulation and social competence. In an accelerated longitudinal design, researchers follow several groups of people, each group starting at different ages. Among the other tasks of the study, there was also the parent-child interaction (PCI) session, which aimed to assess how the development of social competence and self-regulation of children was shaped in the context of interaction with the social environment, particularly with parents. The children and parents who participated in these interaction sessions gave their permission and accepted to be recorded for the study's purposes.

This study utilized only data from the YOUth Child & Adolescent wave, more specifically 'Around 9' wave – containing video recordings and questionnaire data from 100 Dutch participants in age around 9 years old (8 to 10 years old), of which 52% was female. During each PCI session, parent and child were asked to discuss based on a pleasant and a difficult topic, such as, make some plans for a break or holiday and talk about a conflict they had earlier in the month. Therefore, the dataset consisted of 200 videos in total, each lasting approximately 15 minutes. Every video began with the instructor explaining the task to both parent and child and the main part depicted the parent and child alone in a room, seated on chairs while discussing. The language spoken in this dataset is Dutch. Section 5.2.2 discusses and explains the formation of the final dataset used by this study.

Along with the video recordings of the sessions, corresponding data in which parents report information about their child is available, as well. Parents were requested to complete the Child Behavior Checklist questionnaire (CBCL) [1], which consists of questions that assess behavioral and emotional problems. Questions included in CBCL related to depressive symptoms can be found in the following Table 1. Even though the dataset includes extensive information about the subjects, this study will utilize only specific details such as subject ID, gender, and CBCL scores.

Each question item can be responded to according to the following 3-point scale.

Questions related to depressive symptoms	
Question nr.	Question
5.	There is very little he/she enjoys
42.	Would rather be alone than with others
65.	Refuses to talk
69.	Secretive, keeps things to self
75.	Too shy or timid
102.	Underactive, slow moving, or lacks energy
103.	Unhappy, sad, or depressed
111.	Withdrawn, doesn't get involved with others

Table 1. Questions of the CBCL Scale related to depressive symptoms

0 - *Not True (as far as you know)*
1 – *Somewhat or Sometimes True*
2 - *Very True or Often True*

Thus, the depressive symptom score in accordance to 3-point scale and the aforementioned questions ranged from a minimum of 0 to a maximum of 16. In the database utilized in this study, values ranged between 0 and 14, and the distribution of these values is depicted in Figure 7. These values were later used as labels in the study to facilitate the analysis and research of depressive symptoms from the PCI videos. It is essential to clarify that this score is intended to reflect only the severity of withdrawn/depressive symptoms and cannot be considered a clinical diagnosis.

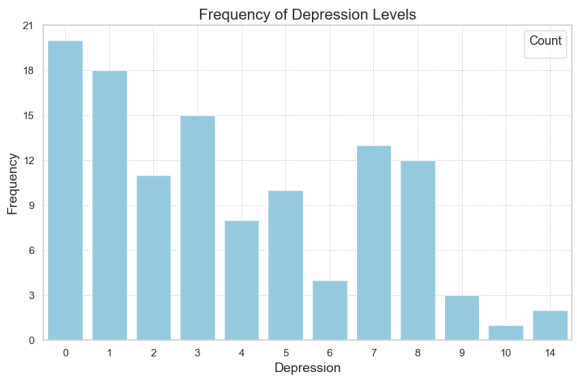

Figure 7. Distribution of the depressive symptom severity values.

Transformation of PCI Interaction Scores for Binary Classification of Depressive Symptoms The focus of the study was to determine whether or not audio and text features extracted from PCI could indicate whether a child exhibits depressive symptoms. To suit

the classification task, the original labels, which were scores ranging from 0 to 16, were converted into binary labels: True and False. “True” represents subjects considered to have depressive symptoms, while “False” denotes all subjects without depressive symptoms.

The methodology for this transformation relied on cut-off points. Initially, the scores were converted to Z-scores and following to T-scores using the formulas 5.1, 5.2. In Z-score's formula, μ indicated the population mean and σ the standard deviation of the population. Z-score and therefore T-scores as well, varied according to the subject's gender, as the mean and standard deviation differ for boys and girls in the US population. The specific values used in each case can be found in Table 2. After obtaining the T-scores for each subject, the next step was to convert these scores into labels.

$$Z = \frac{X - \mu}{\sigma} \qquad (5.1) \qquad\qquad T = 50 + 10 \times Z \qquad (5.2)$$

	Mean	Standard Deviation
Boys	1.1	1.6
Girls	1.4	1.7

Table 2. US population mean and std values

As given in the CBCL manual, two different thresholds can be applied on T-scores to determine the labels. The first threshold, set at 70, identifies clinical depression—subjects with scores of 70 or higher are classified as clinically depressed, while those below this threshold are not. The second threshold, set at 65, identifies cases where subjects exhibit borderline depressive symptoms.

Figure 8 and Figure 9 show the label distribution for both thresholds in the dataset. The dataset used for the classification tasks discussed later contains labels extracted with a threshold of 65. This threshold, as shown in the figures, results in a quite smaller imbalance between the two classes. Furthermore, this approach relaxes the task, allowing the model to identify cases in which subjects exhibit borderline symptoms, potentially preventing the development of more severe depressive symptoms.

5.2 Data preprocessing

Applying preprocessing techniques to the video data was an essential first step towards preparing for feature extraction and analysis of the language and audio modalities for depressive symptoms in children. Initially, video files, which were in mp4 format, were converted into audio files because this study focuses exclusively on language and audio modalities.

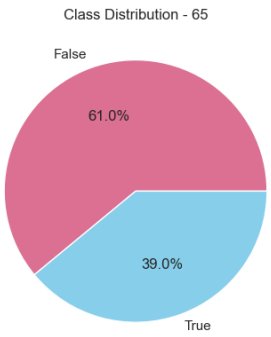

Figure 8. Label distribution for cut-off equal to 65.

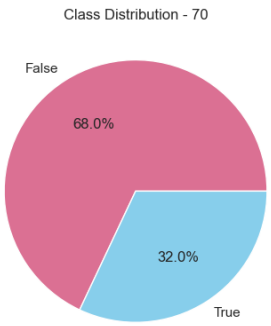

Figure 9. Label distribution for cut-off equal to 70.

MoviePy, a Python library for video editing that provides easy-to-use and high-level functions, was used to convert video files to audio. This library was used to extract audio clips from videos in mp3 format.

Later in the preprocessing, audio is required to be in wav form, so as to be diarized and transcribed and later proceeding with the feature extraction. Consequently, the next preprocessing step was converting the audio from mp3 to wav format, which served as the final audio format. For this conversion, the AudioSegment module from the pydub library [90] —a library designed for manipulating wav files was utilized.

Following the conversion of the participants' videos into a suitable audio format (wav), each modality's next steps can be performed. These processes, which will be explained below, are speaker diarization (see Section 5.2.1) and transcription (see Section 5.2.3).

Since this data is considered sensitive, all the operations described below were conducted exclusively on the Snellius server, where the data was initially located.

5.2.1 Speaker Diarization

After ensuring the audio is in the correct wav format, the next essential step is speaker diarization. Speaker diarization is the process of segmenting and grouping audio segments by speaker and determining the number of different speakers [106]. This procedure results in a record of events that provides information on "who spoke when" for an audio file [79]. This process does not require prior knowledge of the speakers' number or identity [79].

As this study explores scenarios, in which only children's or parents' segments are essential to be studied 2, it is crucial to eliminate these segments and exclude the other speakers, regarding the scenario. To achieve this isolation, speaker diarization techniques were

necessary to be applied and is discussed below.

For the speaker diarization task, the PyAnnote [14] speaker diarization model was chosen. Although there are various systems capable of performing this task, PyAnnote has been proven to perform better in similar tasks compared to GMM, Word Online, and the Google model in terms of Diarization Error Rate [114]. Thus, the audio files were diarized using the pyannote/speaker-diarization-3.1 function from the PyAnnote library.

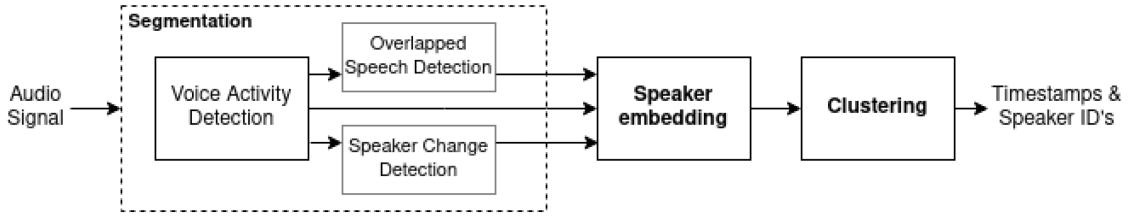

Figure 10. Pyannote pipeline of diarization. Figure by [79]

The pyannote pipeline, as illustrated in Figure 10, is composed by 3 main components: segmentation, speaker embedding and clustering. Initially, the audio is fed to the segmentation module, in which timestamps are created by detecting voice activity, speaker change and overlapped speech. The latter happens when two or more speakers talk simultaneously. This mechanism creates segments of audio that ensure that different speakers are separated from one another and overlapping parts are identified.

Next module in the diarization pipeline is the speaker embedding, which creates a vector representing the speaker's voice characteristics in a vector space containing all the speakers. In this vector space, a speaker's vector phrase is located closer to other vectors of the same speaker than to those of different speakers.

Finally, the clustering module groups the segments of the same speaker according to their embeddings, clustering similar speech characteristics in order to identify the speakers throughout the audio.

The output of this pipeline is a collection of timestamps and speaker IDs. Therefore, the child, the parent, and, in some instances, the instructor are the possible speakers. This information was also used in the pipeline's hyperparameter tuning, by setting the parameter `num_of_speakers` equal to 3. Each audio file was processed through the PyAnnote pipeline, which returned a list of IDs and their corresponding timestamps. In this study, the resulted files from the diarization belonged to either the child, the parent, or the instructor. For the next steps of the research, it was necessary to merge the diarization segments according to the IDs. This means that segments belonging to the child were combined into one audio file, and similarly for the parent.

5.2.2 Speaker identification

We need to identify the speakers in the audio after the speaker diarization process. Audio segments of the children are needed to answer Research Question 1 2 and Research Question 2, whereas audio segments of the parents are needed to answer Research Question 3. Audio segments of the instructor are not used. Even though the splitting of the speakers offered by the diarization pipeline is valuable, it does not provide any information about the speaker's identity - whether the audio corresponds to a child, parent, or instructor. In order to determine which diarised bits belong to the child and which ones to the parents or instructors, we carried out an inquiry into possible methods of doing so was carried out.

This task can be completed manually or automatically, with each having its own set of benefits and drawbacks. This task could have been completed with existing models (wav2vec-adult-child-cls) that detect age and classify audios as adult or child. Although these and other machine learning-based automatic age detection models have the potential to be extremely efficient and scalable, manual identification was deemed the best option based on the task's requirements.

Manual identification ensures that speakers are correctly identified when creating the final dataset, which only contains audio recordings of children and parents. It offers greater accuracy and interpretability than machine learning models, which may not provide exceptional accuracy.

Final dataset

During the manual identification process, various flaws were observed in the dataset. Since the study focuses on the Dutch language, audio recordings taken from an interview in which a mother and kid were speaking in German were removed from the dataset. Additionally, for some instances, diarization step did not yield good performance. Even though these segments are reprocessed, results were still unsatisfactory. Therefore, we decided to remove these instances from our dataset.

We noticed that several audio recordings had been improperly categorised or lacked a speaker throughout the detection procedure. Additionally, some recordings had parasitic noise and very poor sound quality, making it impossible to understand the speaker. Due to these issues, we excluded a total of 9 problematic audio recordings from the final dataset.

The final dataset contains 191 data points when it comes to the children dataset, which are represented by raw WAV audio files. In accordance to the children dataset, for the parents' dataset, the corresponding 191 datapoints were kept and only those were taken into account in the classification tasks.

5.2.3 Transcription

The second modality examined in this research is the text modality. First step to process this modality is to extract text of each session, the task known as transcription. Transcription is the process of converting spoken language into text, and there are a plethora of tools available, such as WhisperX [9], Word Online, and Google. According to [114], the WhisperX pipeline outperforms all three transcription metrics—word error rate (WER), match error rate (MER), and word insertion likelihood (WIL).

WhisperX utilizes Whisper as the primary transcription tool in its pipeline, as shown in Figure 11. Whisper [84] is an automatic speech recognition system that allows for accurate transcriptions in multiple languages, accents, and background noises. Whisper's success in transcription tasks is attributed to its training on 97 languages, including Dutch, with a total of 680,000 hours of multilingual noisy data (excluding English).

In terms of architecture, Whisper is an encoder-decoder Transformer. It takes 30-second audio segments as input, resamples them to 16,000 Hz, and converts them into a log-Mel spectrogram, which is then passed into the encoder. The decoder generates the text caption, including language identification and timestamps per phrase. Whisper was trained on large and diverse datasets, making it more robust and less error-prone than other transcription models.

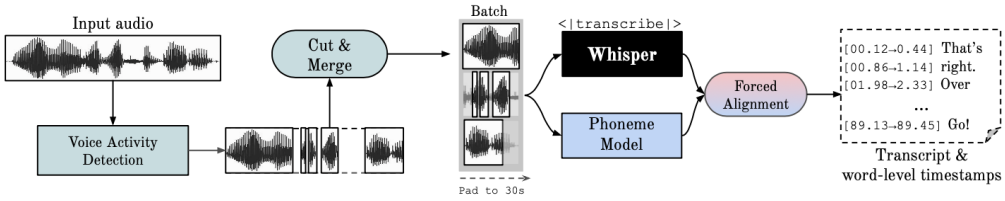

Figure 11. WhisperX transcription pipeline. Figure by [9]

WhisperX Architecture

The WhisperX model comprises several modules, as shown in Figure 11. The Voice Activity Detection (VAD) module detects parts of the audio containing speech, trimming parts with no speech to reduce errors and hallucinations. VAD is a sequence labeling task in which the input audio is represented by a vector of labels that indicate the presence or absence of speech.

The next step in the WhisperX pipeline is the cut and merge process. Segments with detected speech are evaluated based on their length. Segments longer than 30 seconds (the maximum duration for Whisper's training) are shortened. Similarly, very short sentences are merged until they are close to 30 seconds in duration, providing more contextual information for transcription.

Following preprocessing, each audio segment is independently fed into the Whisper model for transcription. This study used the large-v1 version of the Whisper model with a batch size of 16. There is no learning effect in the procedure to avoid hallucination.

After transcription, forced alignment is applied to the segments generated by Whisper. This task involves a phoneme recognition model that identifies the position of phonemes (any small unit of speech) within words. The model matches the words in the transcription to the exact parts of the audio segment.

The final output of the transcription task processed with WhisperX is a list of transcriptions with word-level timestamps. In this study, the pipeline was applied to audio segments created after diarization and merging of the children and parents' audios, as described in Section 5.2.2.

5.3 Audio feature extraction

Audio features were extracted using both the wav2vec2.0 [8] and CLAP [117] models. For wav2vec2.0, the *facebook/wav2vec2-large-xlsr-53-dutch* model [39] from Hugging Face was used for feature extraction. This model has been fine-tuned on Dutch language data using the validation splits of Common Voice 6.1 [7] and CSS10 [78]. The CLAP implementation used in this study is *laion/larger_clap_general* [117] from Hugging Face and will be further described in Section 5.3.3.

The following sections discuss the data preprocessing for the audio models and the models' architectures.

5.3.1 Audio Input setup

For children's audio recordings, the length ranged from 30 seconds to nearly 5 minutes, while for parents' audio recordings, it was between 1 minute and 7 minutes.

Regarding the embeddings extracted from the audio, two different approaches were explored in this research. The first approach utilized the entire audio of each session, regardless of its length. This resulted in vectors of varying sizes, with longer audios producing longer vectors and shorter audios producing shorter vectors.

The second approach, was the chunk approach. The rationale behind this experiment was that both Wav2Vec2.0 and CLAP were trained on smaller, fixed-size chunks, in contrast to our dataset. Various studies, as the one from Pepino et al. [81] that explored the task of emotion recognition with features extracted from Wav2Vec2.0 using these models for

feature extraction have similarly used smaller chunks as input. As a result, the models in this study employed smaller chunks as input, as a secondary experiment.

When chunk approach is followed, the number of samples extracted from a single audio increases. Huang et al. [46] performed a similar experiment, in which features extracted from Wav2Vec2.0 were utilised for depression detection with data extracted from DAIC-WOZ dataset [38]. The strategy followed in this thesis involved extracting five chunks of 12.5 seconds from each audio. The starting point of each window was randomly selected, with a check to ensure enough audio remained to cover the full 12.5 seconds. This resulted in a final dataset of 955 items, including audios for all subjects and different scenarios.

As the number of data points increased to 955, corresponding to chunks from the initial audios, a method to map the chunks' predictions to the actual audio was needed. For this reason, we followed two different strategies for calculating the final label: the primary and the at-least-1. The primary strategy selects the most common label out of the five predicted labels. The at-least-1 strategy requires at least one chunk to be predicted as True to label a subject as True. The randomness of the chunks was the main rationale behind the latter strategy. For example, it could be a case where the 12.5-second chunks are not informative enough regarding the depressive symptoms' detection task; thus, even if only one chunk was labeled as True, it would be sufficient.

Audios segments that are input into Wav2Vec2.0 had a sample rate of 16 kHz, while those given to CLAP had a sample rate of 48 kHz. The following sections detail the architectures and feature extraction processes of these models, starting with Wav2Vec2.0.

5.3.2 Wav2Vec2.0

Regardless of the length of the audio, the transformer-based Wav2Vec2.0 model [8] can create vector representations from raw audio files. A feature encoder, a contextualised transformer network, and a quantisation module compose its three primary phases (see Figure 12).

Feature encoder The feature encoder is the first component of Wav2Vec2.0's design, and its primary goal is to obtain low-dimensional representation of audio signal. It receives the 16 kHz sample-rate raw waveform as input and outputs a range of latent representations including the primary features of the audio input (shown as Z in Figure 12).

Normalising the waveform to have a zero mean and unit variance is the first step. Next, a 7-layer convolutional network with 512 channels in each layer receives this normalised waveform as input. Notably, as we move through the network, the

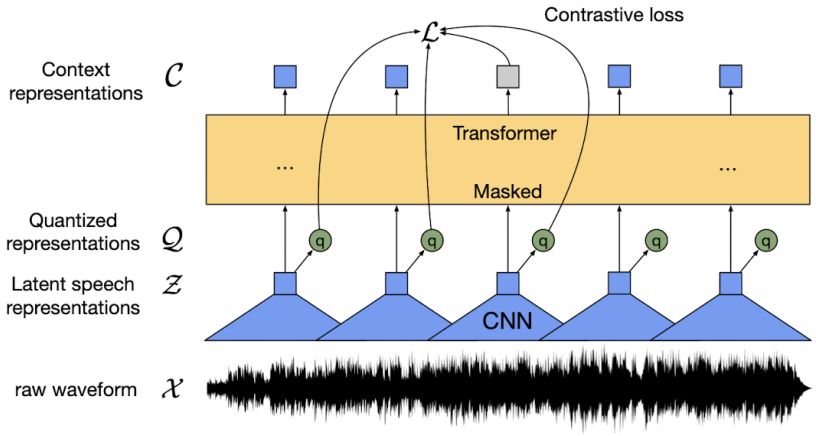

Figure 12. Illustration of the Wav2Vec2.0 architecture. Figure by [8]

kernel's size and width drop. The Gaussian Error Linear Unit (GELU) activation function, which comes after the convolution layers, generates the final output. The salient features of the input audio are captured in these representations.

Contextualized transformer The latent representations created by the encoder are processed by 24 Transformer blocks in the LARGE version of the model (12 blocks in the BASE version). The input sequence first goes through a feature projection layer, whose dimension is increased to 1024 for LARGE and 512 for BASE to match the size of the convolutional layers.

The transformer modules have two functions: a feed-forward neural network and an attention mechanism. The model is able to compute a sequence's representation while taking into account various positions within the sequence thanks to the self-attention mechanism. This allows attention to catch relationships from beginning to end across the whole series of latent representations. The feed-forward networks are responsible for managing the local representations of each latent representation, while the attention mechanism takes care of the global links between the representations.

In summary, the transformer's role in Wav2Vec2.0 is to construct a contextualized representation from the latent representations extracted by the feature encoder.

Quantization module The quantisation module completes the Wav2Vec2.0 architecture. Discrete unit representation of the data is necessary for self-supervised training. For instance, text is inherently discrete since it can be divided into individual words, sentences, and characters. The main job of the quantization module is to convert the latent representations into quantised discrete speech units, yet speech is continuous by nature.

These distinct units are taken out of codebooks, which have a predetermined list

of speech sounds in them. A speech unit is made up of these sounds combined. For Wav2Vec2.0, two codebooks having 320 potential sounds each yielded 102,400 possible speech units. In the figure 12, Q represents the final quantised latent representations.

Training and Finetuning Pre-training of the Wav2Vec2.0 model was conducted using 53.2k hours of unlabelled data with spoken English audios, from Libri-light dataset (LV-60k) [50] and and LibriVoice dataset (LS-960) [77]. This pre-training procedure is comparable to the method used to train the language model BERT - a language model discussed later in this thesis 5.4. The model aims to recreate the quantised encoder representations for the masked frames as the output of the contextualised encoder, where a number of timesteps from the latent representations of the feature encoder are masked. Equation 5.3 illustrates how contrastive loss is used to measure this portion of the training goal, where sim represents the cosine similarity between context c_t and quantized representations q_t. Specifically, given the context representations c_t. This set includes the true representations, which are the outputs of the context network, centered over the masked time step t, the model needs to identify the true quantized representations q_t from a set of $K + 1$ quantized candidate representations $\tilde{q} \in Q_t$. This set includes the true representation q_t and K distractors, which are uniformly sampled from other masked time steps within the same utterance.

In addition to contrastive loss, diversity loss, as this is depicted in Equation 5.4 assesses whether the model favours all of the words in the codebook equally. Specifically, this loss encourages the equal use of the V entries in each of the G codebooks by maximizing the entropy of the averaged softmax distribution over the codebook entries for each codebook $\bar{p}_g$ across a batch of utterances. Notably, the softmax distribution does not include the Gumbel noise or a temperature parameter. The model's training objective is made up of the total of these losses, the summation of contrastive and diversity loss.

$$\mathcal{L}_c = -\log \frac{\exp(\text{sim}(\mathbf{c}_t, \mathbf{q}_t)/\text{k})}{\sum_{\tilde{q}\in\mathcal{Q}} \exp(\text{sim}(\mathbf{c}_t, \tilde{\mathbf{q}})/\text{k})} \tag{5.3}$$

$$\mathcal{L}_d = -\frac{1}{GV} \sum_{g=1}^{G} \sum_{v=1}^{V} \bar{p}_{g,v} \log \bar{p}_{g,v} \tag{5.4}$$

Feature extraction from Wav2Vec2.0

After explaining how the architecture and training of Wav2Vec2.0 model, it is time to describe how the embeddings were extracted. The audio was fed in the model, either as a whole or in chunks. The embeddings, extracted from the transformer's final layer,

served as the feature representation for the classification task. The large-v2 version of the multilingual Wav2Vec2.0 is used in this study, thus the last layer size is equal to 1024.

As far as the first experiment is concerned, the varied length of the audio resulted in matrices of different size. By taking the mean representation over all segments, each audio recording was represented by a 1024 vector, regardless of its initial size.

For the chunked audio approach, each 12.5-second chunk produced a vector of the same size. The same procedure was followed, therefore each chunk was represented by one 1024-sized vector. In this case each subject was represented by 5 different feature representations, with the main goal of this approach is to provide more and smaller input data for the classifier.

In both cases described above, the feature vectors served as the input in the classification task explained later.

5.3.3 CLAP

The second model used in the audio modality for feature extraction is CLAP (Contrastive Language Audio Pretraining) [117]. CLAP, as its name suggests, is an audio-language foundation model that aligns text and audio modalities through contrastive learning. In other words, it maps similar instances together in a latent space and determines how similar or dissimilar they are in order to extract meaningful representations from unlabelled data.

CLAP is comparable to CLIP [83], which fulfills the same purpose for language and vision modalities. There are two primary encoders in the CLAP model, one for each modality. Two more modules are included in CLAP besides the encoders: the supervised audio classification module and the zero-shot audio classification module. This set of modules is shown in Figure 13. The audio encoder, which is the component of the architecture used in this study, is further discussed in the following section.

Encoders The audio encoder includes PANN, a CNN-based audio classification model with 14 blocks (7 downsampling and 7 upsampling), and HTSTAT, a transformer-based model with 4 swin-transformer blocks. The audio input, denoted as X_a in the figure 13, triggers the appropriate procedure based on its duration. If the audio is shorter than a fixed chunk duration $\mathtt{d}$, it is padded with zeros. This padded audio is then stacked four times in order to serve as the input. In case that the audio is longer than d seconds, three d-second clips are extracted from the original audio: one from the beginning, one from the middle, and one from the end. These three clips, along with the entire audio downsampled to d seconds, are combined. In both cases, the

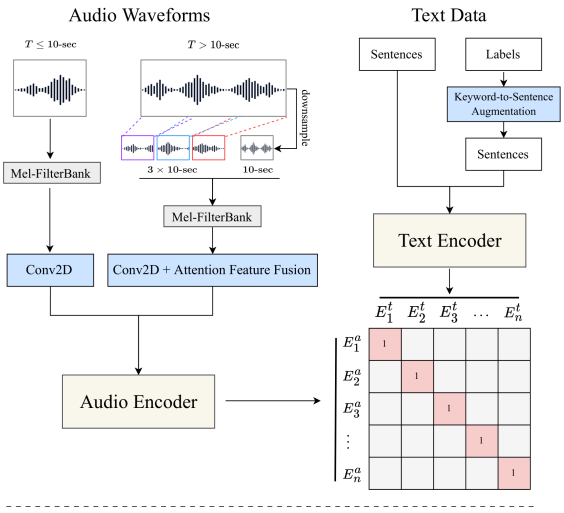

Figure 13. Audio and text encoders of CLAP architecture. Figure by [117].

input will be four stacked segments lasting d seconds each.

The text encoder consists of the CLIP text encoder, BERT [25], and RoBERTa [61]. The results of the three encoders are combined with a multilayer perceptron(MLP) head to produce a 512-dimensional output, which is the same size as the text and audio representation embeddings obtained through contrastive learning.

The embeddings obtained from the audio and text encoders are fed into a MLP with a ReLU activation function, which maps the encoder outputs to the appropriate dimension.

Training and Finetuning CLAP was trained using three different datasets: AudioCaps [53] + Clotho [27], LAION-Audio-630K [117], and Audioset [36]. The LAION-Audio-630K dataset [], which is composed of 633,526 text/audio pairs. It contains 4,325.39 hours of human activity, natural sounds, and audio effects. CLAP was trained with 10-second inputs.

The model was trained using the contrastive learning paradigm between the audio E_a and text E_t embeddings resulted by the MLP layer from audio waveforms and text data (see Figure 13), with the following contrastive loss function 5.5. Specifically, τ is a learnable temperature parameter for scaling the loss. Two logarithmic terms account for either audio-to-text logits or text-to-audio logits. During training, N represents the batch size instead of the total number of data points, as we update the model using batch gradient descent rather than computing the entire data matrix at once.

$$L = \frac{1}{2N} \sum_{i=1}^{N} \left(\log \frac{\exp(E_i^a \cdot E_i^t / \tau)}{\sum_{j=1}^{N} \exp(E_a^i \cdot E_j^t / \tau)} + \log \frac{\exp(E_t^i \cdot E_i^a / \tau)}{\sum_{j=1}^{N} \exp(E_t^i \cdot E_j^a / \tau)} \right) \tag{5.5}$$

Feature extraction from CLAP

For embedding extraction with CLAP, we use the representation produced by the MLP head. This representation has a size of 512, as we are using the large model. The audio was again fed in the model and multidimensional vectors were created, containing features the model deemed more representative.

Similar to Wav2Vec2.0, the duration of the audio affects the feature vector size. To produce vectors of consistent size from different-sized audio representations, the mean is calculated per feature over all segments. Zhao et al. [123] performed a similar procedure to their work, where they used CLAP for both text and audio modality. What they did was to integrate the sequence of features into a single vector for each modality by taking the mean along the sequence dimension. In case of the whole audio, each subject's audio is represented by a single vector fo size equal to 512. In the augmented approach, with chunks of audio, each subject is represented by five 512-dimensional feature vectors.

The feature vectors from the two previously mentioned cases were used as input for the classification task to evaluate whether a child exhibits depressive symptoms.

5.4 Text feature extraction

Similarly to the audio modality, feature extraction from the sessions' transcripts was performed using large language models. The features extracted from these models served as input to the classification tasks and were vectors representing each transcript. The features from the transcripts were extracted based on the tasks this study explored. Thus, transcripts from children-only and parent-only segments were necessary to address the research questions set in Chapter 2.

This approach was also essential for the multimodal part of the research. The textual features were combined with the audio features extracted as described earlier to determine whether this fusion improved predictions in the assessment of depressive symptoms in children.

Two models were leveraged for feature extraction from the sessions' transcripts: Sentence-Bert (SBERT) [87] and RobBERT [23]. For SBERT, the *NetherlandsForensicInstitute/robbert-2022-dutch-sentence-transformers* model [69] from Hugging Face was used — a finetuned Dutch-translated version of the Paraphrase dataset [33; 70], as it was the most appropriate implementation of SBERT.

The RobBERT implementation *DTAI-KULeuven/robbert-2023-dutch-large* [23], also from

Hugging Face, is a Dutch version of the BERT model [25], pretrained on the OSCAR2023 dataset [73; 74] and using a novel tokenizer [89].

The following paragraphs describe the architectures of SBERT and RobBERT, as well as the feature extraction procedure. A summary of BERT's architecture is provided first, as both SBERT and RobBERT are based on the BERT language model.

BERT Architecture BERT [25] is a language model which achieves state-of-the-art performance in numerous natural language processing tasks. Its architecture is based on transformers architecture and maintains bidirectional representations from unlabelled text by jointly combining both left and right context in all layers. The bidirectionality enables BERT to read both right to left and left to right context simultaneously allowing it to learn the context of each word and how it relates to other words in a sentence.

What makes BERT powerful is its ability to be trained on unlabelled data. During its pretraining BERT was trained on unlabelled data across different tasks, providing a significant advantage because labelled data is limited, whereas unlabelled data is vast. BERT got trained in BookCorpus [125] and English Wikipedia with 800M and 2500M words correspondingly. This resulted to the 350 million parameter large model and the 110 million parameter base model.

Training and Finetuning The first part of BERT's framework is pretraining, where the model is trained on unlabelled data over different pre-training tasks. This is followed by the fine-tuning in which the pre-trained parameters serve as initial, which are then fine-tuned based on the downstream task.

BERT's pre-training consists of two unsupervised tasks: the masked Language Model (MLM) and the next sentence prediction(NSP). For MLM, parts of the input are masked and the model has to predict what should be filled in the blank spots. Bidirectionality lets the network predict 15% of the target tokens that are masked. For NSP, the model's goal is to determine whether two input sentences belong together or not. This task involves training the model to understand the relationship between sentences, which is crucial for tasks like Question Answering and Natural Language Inference.

BERT accepts a text sequence as input, which can be a single sentence or two sentences packed together. A particular characteristic of BERT is that the first token of each sequence is [CLS], a special classification token that also provides an aggregated sequence

representation of the whole sentence. The latter one is located in final hidden state of CLS and is very useful in case of classification tasks. Another special characteristic of BERT's encoding is the [SEP] token, which is used to differentiate the sentences, along with a token indicating if it belongs to sentence A or sentence B. Figure 14 depicts how the input representations were constructed with BERT.

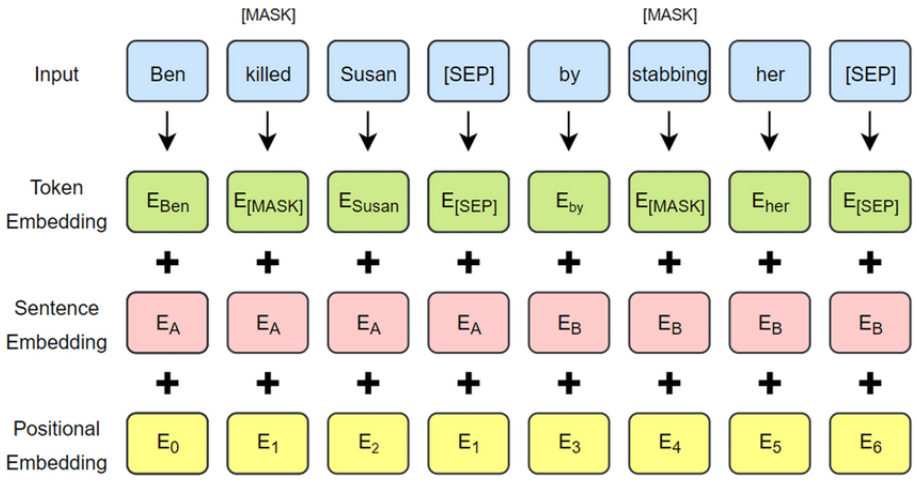

Figure 14. BERT input representations formation. Figure by [25]

BERT can perform very well in various downstream language tasks with minor hyperparameter tuning. The fine-tuning of BERT model can be achieved by only adding a layer on the main model, depending on the task. Thus, fine-tuning is straightforward, since the BERT's self-attention mechanism allows the model to adapt to the downstream tasks by swapping the corresponding inputs and outputs.

The original implementation of BERT was in English in two sizes: base and large. The base model has 12 hidden layers, each with a size of 768 and 12 self-attention heads. The large model, on the other hand, has 24 hidden layers with a size of 1024 and 16 self-attention heads. Due to BERT's success in various tasks, many language models based on BERT have been created, including RoBERTa [61], HubBERT [43], and RobBERT [23].

With an understanding of BERT's architecture, the subsequent sections delve into RobBERT and SBERT architectures and describe how the deep features were extracted using these models.

5.4.1 RobBERT Architecture

RobBERT [23] is the Dutch version of the RoBERTa model. Before delving into the RobBERT architecture it is important to understand how RoBERTa [61] improved BERT's performance through adaptations in the training procedure.

Liu et al. [61] replicated BERT's pretraining and adapted parts of the architecture. This

led to a better performing model, thus they proposed several changes that achieved state-of-the-art results. These changes, which comprise RoBERTa (Robustly optimised BERT approach), included longer training with bigger batches over more data, training on longer sequences, as well as excluding the NSP part from the training. Additionally, a dynamically changing masking technique was applied, where not just a single token but consecutive tokens were masked, forcing the model to rely more on the context to predict the blanks. RoBERTa was trained on three additional datasets, resulting in training on ten times more data than BERT without any signs of overfitting.

RobBERT is a model based on the RoBERTa architecture and training procedure, but trained with a Dutch corpus. There are two different version of the RobBERT model: the first version (v1) was pretrained on a Dutch corpus, while the second version (v2) used both a Dutch corpus and a Dutch tokenizer. This study utilised the first Dutch large model of 355 million parameters, trained on the OSCAR dataset with a novel Dutch tokenizer. OSCAR is a large multilingual corpus, with a Dutch corpus comprising 39GB of data, containing 6.6 billion words. Both versions of RobBERT outperforms the other Dutch BERT models, with v2 showing significantly better performance.

Feature extraction from RobBERT

For deep feature extraction with RobBERT, the procedure described below was followed. The entire transcript was used with the maximum sequence length for the model's input set to 512, which was consistent with BERT's pretraining procedure. Each group of tokens, with a maximum of 512 tokens,served as the input of the model. A tensor of size 1024, representing the hidden states for the last layer in the transformer, was extracted as the deep features of the corresponding group of tokens.

Since each transcript had different length, the number of vector representations also varied. Similar to the approach followed in the audio modality, the mean per feature was calculated, resulting in a 1024-dimensional vector per transcript. This vector contained the deep features derived from the model.

This feature vector was later utilised as the input of the classifier that was trained on the task studied in this thesis - predicting depressive symptoms in children.

5.4.2 SBERT Architecture

Similar to RoBERTa, and as the name implies, SBERT [87] is another variant of the BERT model. SBERT employs siamese and triplet network architectures, enabling it to produce sentence embeddings that can be compared using cosine similarity. This reduced computational time while maintaining accuracy. For instance, the most similar pair of sentences could be found in 5 seconds, whereas BERT or RoBERTa might need around

60 hours. SBERT addressed the need for using BERT at the sentence level by accepting sentences as input and creating fixed-sized vectors as output.

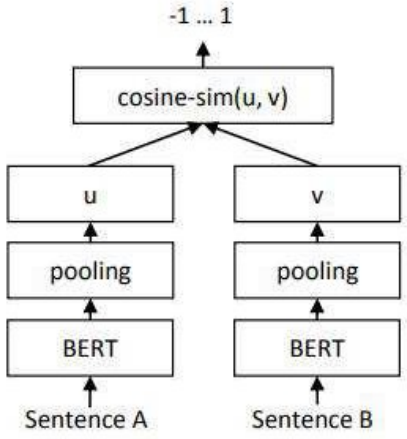

Figure 15. SBERT architecture for a classification task. Figure by [87]

Regarding SBERT's architecture (see Figure 15, a pooling operation was added to the output of BERT. This pooling layer allows us to create a fixed-size representation for the variable-length input phrases. Experiments with three pooling strategies were performed, using only the CLS token output, MEAN-strategy - calculating the mean of all vectors, MAX-strategy – computing a max-over-time of the output vectors. By default, MEAN-strategy was used.

Training and Finetuning For finetuning, Siamese and triplet networks were employed. The weights were updated to ensure sentence embeddings were meaningful, allowing similarity to be calculated using cosine similarity. The network structure varied based on the training data, and the following objective functions were utilized.

- **Classification Objective Function:** The embeddings of u and v are concatenated with the element-wise difference $|u - v|$. This value is then multiplied with a trainable weight and fed into a softmax layer to generate probabilities for each class.
- **Regression Objective Function:** This involves calculating the cosine similarity of two sentence embeddings and using mean-squared error as the objective function.
- **Triplet Objective Function:** A triplet loss tunes the network so that the distance between an anchor sentence a and a positive sentence p is smaller than the distance between a and a negative sentence n.

SBERT was trained on the SNLI [94] and multi-genre NLI [116] datasets, which contain collections of sentence pairs. In total, 1 million pairs composed SBERT's training data, covering a wide variety of genres of spoken and written text. Its performance was evaluated

on common Semantic Textual Similarity (STS) tasks. The goals behind training SBERT were to enable its use in various applications such as STS, Semantic Search, clustering, and Natural Language Inference (NLI).

Feature extraction from SBERT

For feature extraction from a sentence transformer from Hugging Face finetuned in Dutch laguage, each transcript was transformed into a list of sentences using the 'sent_tokenize' function from the NLTK library. This list of sentences was then fed to the sentence transformer model, and embeddings of size 768 were extracted. Each transcript was thus represented by a 768-dimensional representation, which served as the transformer input.

5.5 Multimodal feature fusion

In addition to working with audio and text features individually, it is interesting to explore how the fusion of these two modalities affect the performance of the model in the task of assessing depressive symptoms in children. The framework used for the multimodal fusion of the two modalities is as follows: Initially, two branches are used to extract features from audio and text, as discussed in the sections above. Feature vectors for each modality are then created, and their concatenation forms the multimodal fusion feature, which serves as the final feature representation of each subject in the classification task.

The proposed framework (see Figure 16) has been applied with all possible fusion combinations. We fused features from SBERT and RobBERT with features from both Wav2Vec2.0 and CLAP. The multimodal fusion was tested on the entire dataset as well as for each scenario explicitly. In Table 3, we present a summary of the feature vectors per model, the fused feature vectors, and their corresponding dimensions.

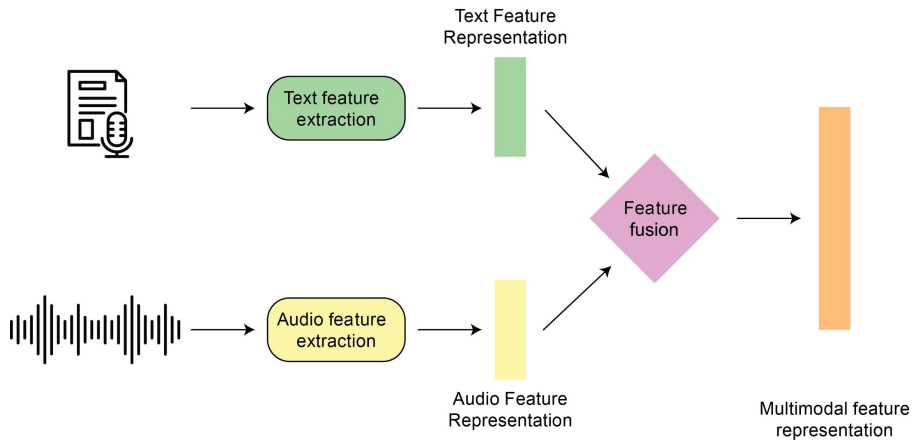

Figure 16. Fusion framework of the audio and text modality.

The following abbreviations will be used in this report to represent the fusion of the models: SBERT + Wav2Vec2.0 (S + W), RobBERT + Wav2Vec2.0 (R + W), SBERT + CLAP (S +

C), RobBERT + CLAP (R + C).

Model	Feature Vector Dimension
Audio Models	
Wav2Vec2.0	1024
CLAP	512
Language Models	
SBERT	768
RobBERT	1024
Fused Features	
(S + W)	1792
(R + W)	2048
(S + C)	1280
(R + C)	1536

Table 3. Feature vector dimensions for audio and language models and their fusion.

5.6 Models and parameters

This section discusses the model used for the classification of depressive symptoms utilizing the features extracted as outlined in the earlier sections. XGBoost classifier [17] was used in this research, as it is effective in case of smaller datasets, as the one used here for the automated assessment of depressive symptoms in children. Next to model fitting and prediction functions, it provides model instantiation with several hyperparameter options. In every setup discussed below, the data split followed 70/30 split, with 70% of the data utilised for training and 30% for testing.

The most crucial hyperparameters for each approach that are relevant to the experiment are briefly discussed below:
XGBoost classifier:

- **Maximum depth** (default =6): This variable `max_depth` represents the maximum depth of the tree. A larger number could result to more complex model prone to overfitting.
- **Number of estimators**: The variable `n_estimators` represent the number of trees to fit in a random forest.
- **Class weight**: This variable `scale_pos_weight` serves a balancing variable for positive and negative weights in case of imbalanced datasets.
- **Gamma**: Gamma variable `gamma` represents the minimum loss reduction required in order to make a further partition on a leaf node of a tree.

5.6.1 Hyperparameter tuning

The optimization of hyperparameters was necessary to produce the optimal model from XGBoost classifier. The performance of a model can be improved by selecting the optimal set of hyperparameters through hyperparameter tuning. Cross-validation was performed in this experiment to fine-tune the model hyperparameters. Specifically, the grid search method, which creates a grid of predetermined values for hyperparameters, was applied. In each iteration, a different set of defined hyperparameter values was tested in a particular order. The model was fitted using every conceivable set of hyperparameters. To obtain more informative results, each combination was repeated several times, depending on the number of folds, and the performance of the splits was averaged for each combination. The combination that performed best was selected after this process.

On the implementation level, the function GridSearchCV, which is part of the model_selection package of scikit-learn [80], was used. The number of folds for each combination and the values of the hyperparameters that needed to be assessed were passed to this function. Consequently, to perform cross-validation, the tested hyperparameters and their exact values were defined for each model.

Table 4 list each algorithm's hyperparameters along with the defined values considered during the initial cross-validation run. A 5-fold cross-validation was carried out for the GridSearchCV run.

Hyperparameter	*Values*
max_depth	[3, 4, 5, 6, 8, 10, 12, 15]
n_estimators	[10, 20, 30]
gamma	[0.0, 0.1, 0.2, 0.3, 0.4]

Table 4. Hyperparameters and defined values for the classification task.

The variable `scale_pos_weight` was intentionally excluded from the grid search, as it was explicitly calculated based on the number of positives and negatives in the dataset. The proportions of positives and negatives varied depending on the research scenario. When the entire dataset was used to address the main research question, which considered both scenarios, `scale_pos_weight` was set to 1.42. When the study focused specifically on the vacation scenario or the conflict scenario, `scale_pos_weight` was set to 1.59 and 0.83, respectively.

5.6.2 Cross validation techniques

This study explored the results produced using two different cross-validation algorithms, considering the special features of the dataset. Since each subject may appear twice in the

dataset—once in the vacation scenario and once in the conflict scenario—special treatment in the data splitting was required. It was essential to ensure that both occurrences of the same subject ID remained in the same split. Therefore, three methods that satisfied this condition were explored: GroupShuffleSplit, Group K-Folds, and StratifiedGroupKFold, all from the scikit-learn library [80], with the number of folds set to 5.

GroupShuffleSplit This method provides a randomized train/test split according to the provided group.

Group K-Folds This method splits the data according to individual groups, ensuring no overlap among groups. In this case, data points from the same subject, whether in the form of audio or transcripts, remained in the same split. Additionally, each test fold contained almost the same number of samples, indicating that the folds were roughly balanced.

StratifiedGroupKFold This method returns stratified folds with non-overlapping groups. It is similar to the Group K-Folds method, but it adds the functionality of attempting to create balanced folds, where the class distribution is preserved as much as possible.

5.6.3 Model evaluation

For the evaluation of the model's results, various evaluation methods were utilized. The F1-score, accuracy, precision, and recall were calculated. Additionally, the Area Under the Curve (AUC) score, which represents the area under the Receiver Operating Characteristic (ROC) curve, was calculated. The ROC curve The ROC curve is plotted with True Positive Rate (TPR), in other words Recall 5.8 or Sensitivity, against the False Positive Rate (FPR) 5.11 where TPR is on the y-axis and FPR is on the x-axis. As demonstrated by Laszlo et al. [48], the AUC score is the only metric unaffected by skewed data distribution. This characteristic made it particularly suitable for this thesis, given the imbalanced class distribution.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5.6) \qquad \text{Precision} = \frac{TP}{TP + FP} \quad (5.7)$$

$$\text{Recall/TPR} = \frac{TP}{TP + FN} \quad (5.8) \qquad \text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5.9)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (5.10) \qquad \text{FPR} = 1 - \text{Specificity} \quad (5.11)$$

The AUC score reflects the model's performance, with a perfect model achieving an AUC of 1 and a random model achieving an AUC of 0.5. As every other metric, it indicates how well the classifier can distinguish between positive and negative classes. But more specifically, AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [68]. All the metrics described above were calculated using functions from the scikit-learn library [80] and were computed for all the different scenarios to address the research questions.

6. Results

This chapter presents the results of all the experiments described in the methodology section, organized according to the research questions.

First, the results of the classification task related to the main research question are presented. Section 6.1 evaluates how the multimodal features – text and audio – detect depressive symptoms in children. The dataset offers a thorough examination of the multimodal method since it contains both scenarios for each subject.

Second, the classification results are presented for each scenario separately. Section 6.2 investigates how the performance of both text and audio modalities individually changes per scenario – cooperative and conflicting. Experiments related to the second sub-research question focused on the unimodal approach. Section 6.3 presents the performance of audio and text features independently, without fusion, in assessing depressive symptoms.

Last but not least, the study extends beyond the child's data, exploring also how features extracted from parents' audios and transcripts relate to depressive symptoms in children. Section 6.4 evaluates the classification performance of deep features derived from the parents' segments, as part of the exploration of which indicators could be valuable for the prediction task.

All results presented in this chapter represent the best outcomes obtained using the optimal parameters and cross-validation algorithms for the deep features. Detailed results from all runs are provided in the Appendix (see Section 10).

6.1 Results of multimodal fusion

This section describes the best results the combined features produced in terms of AUC scores. The fusion of the results performed has been discussed in Section 5.5. The features per modality have been extracted from two separate models, thus in this part of the experiment, all the possible four combinations have been tried for the fusion of the features. Besides using different models for feature extraction, various dataset splits have been explored.

Tables 5 and 6 present the AUC scores achieved for each fusion combination with random split and k-fold split, accordingly.

Model	AUC (Default Weight)	AUC (Class Weight)
(C + S)	0.612	0.584
(W + S)	0.550	0.593
(C + R)	0.568	0.528
(W + R)	0.571	0.567

Table 5. Comparison of AUC scores of each fusion combination using random split with default and class weights.

Model	AUC (Default Weight)	AUC (Class Weight)
(C + S)	0.632	0.680
(W + S)	0.582	0.554
(C + R)	0.561	0.587
(W + R)	0.447	0.489

Table 6. Comparison of AUC scores of each fusion combination using k-fold split with default and class weights.

Figure 17 shows the overall predictive performance of random and k-fold splits for all possible fusion combinations, namely SBERT and CLAP (S + C), SBERT and Wav2Vec2.0 (S + W), RobBERT and CLAP (R + C) and RobBERT and Wav2Vec2.0 (R + W).

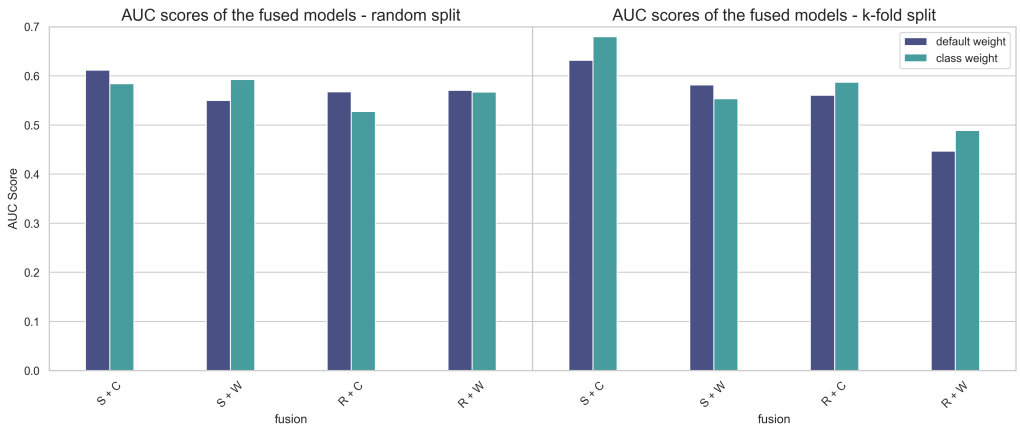

Figure 17. AUC scores obtained from fusion of all models with a random split (left Figure). AUC scores obtained from fusion of all models with a k-fold split (right Figure).

6.2 Results per scenario

In this section, we present the results of the classification tasks, when the features were separated by scenario. Specifically, the extracted features related to the cooperative scenario, in which parent and child were asked to plan a vacation, fed to a classifier

separate from those obtained from the session where the parent and child discussed about a conflict they had.

The results are presented below, categorized by modality. First, the AUC scores from the classification of the audio features –extracted from Wav2Vec2.0 and CLAP – can be seen in Table 7 and Table 8. A notable difference in performance across each model's features per scenario is observed. In both splits, CLAP features perform significantly better in the cooperative scenario, specifically in case of random split CLAP reaches an AUC score equal to 0.746, while in case of Wav2Vec2.0, AUC score is only 0.393.

An opposite trend can be observed in case of the conflicting scenario, in which Wav2Vec2.0 embeddings resulted to higher AUC scores. For instance, when we performed a random split, Wav2Vec2.0 features achieved AUC score equal to 0.611, while CLAP's score was only 0.483. Similar behaviour is noticed in the k-fold split case, with 0.610 and 0.386 being the AUC scores Wav2Vec2.0 and CLAP features produced, respectively.

Model	Scenario	AUC (Default Weight)	AUC (Class Weight)
Wav2Vec2.0	**Vacation**	0.367	0.393
	Conflict	0.611	0.597
CLAP	**Vacation**	0.713	0.746
	Conflict	0.483	0.427

Table 7. AUC scores from feature selection per scenario using random split in audio modality.

Model	Scenario	AUC (Default Weight)	AUC (Class Weight)
Wav2Vec2.0	**Vacation**	0.461	0.460
	Conflict	0.610	0.599
CLAP	**Vacation**	0.568	0.601
	Conflict	0.386	415

Table 8. AUC scores from feature selection per scenario using k-fold split in audio modality.

The AUC scores for the text modality are shown in Tables 9 and 10, based on the features extracted from SBERT and RobBERT models. A similar pattern to the one observed in the audio modality is seen in the text modality as well. In the cooperative scenario, SBERT features specifically obtained a high AUC score of 0.755, and in the conflict task, 0.552. However, RobBERT features achieved an AUC value of 0.592 in the vacation task and outperformed in the conflict task with an AUC score of 0.720.

The overall performance of the two modalities per scenario when data split in a random

Model	Scenario	AUC (Default Weight)	AUC (Class Weight)
SBERT	**Vacation**	0.644	0.670
	Conflict	0.495	0.479
RobBERT	**Vacation**	0.490	0.592
	Conflict	0.551	0.506

Table 9. AUC scores from feature selection per scenario using random split in text modality.

Model	Scenario	AUC (Default Weight)	AUC (Class Weight)
SBERT	**Vacation**	0.676	0.680
	Conflict	0.552	0.494
RobBERT	**Vacation**	0.481	0.424
	Conflict	0.720	0.672

Table 10. AUC scores from feature selection per scenario using k-fold split in text modality.

way is shown in Figure 18. The highest AUC score is obtained by CLAP features in the cooperative scenario and it is followed by the one obtained by RobBERT features in text modality.

Except of the unimodal features, also the multimodal fusion features were used as the features fed in the classification task of assessing depressive symptoms. The performance of the fused features per scenario is shown in Tables 11 and 12. We also present the overall performance of the combined features in Figure 19. The bar plot highlights the superior performance of the (S + C) fusion, which achieves an AUC score at least 10 points higher than other combined models in the cooperative scenario. It is also noticeable that the fusion models perform better in the cooperative scenario compared to the conflicting one.

6.3 Unimodal results

The unimodal results for the audio models, when the scenarios were not taken into account can be found in Table 13 and Table 14. In the main experiment, the deep features correspond to the whole audio, CLAP features perform slightly better with AUC equal to 0.597 while Wav2Vec2.0 reached a AUC score equal to 0.570.

In the secondary experiment, which utilized features extracted from the audio chunks, the model performance was marginally better (see Figure 20). For example, in case we apply a random split with class weight, the chunked segments result to an AUC score equal to 0.620 for Wav2Vec2.0 and 0.668 for CLAP features, with CLAP again outperforming

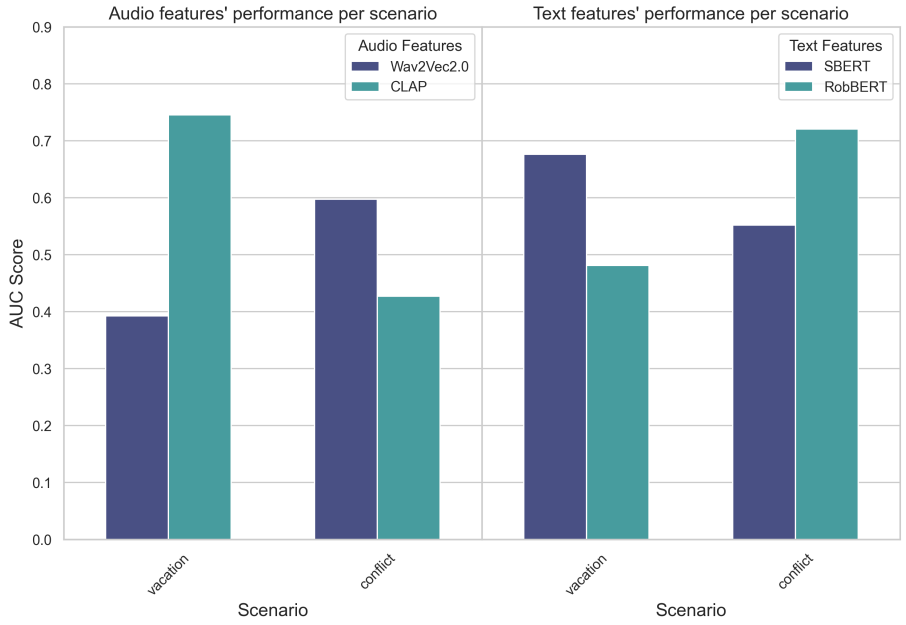

Figure 18. AUC scores produced by each model per scenario in case of a random split.

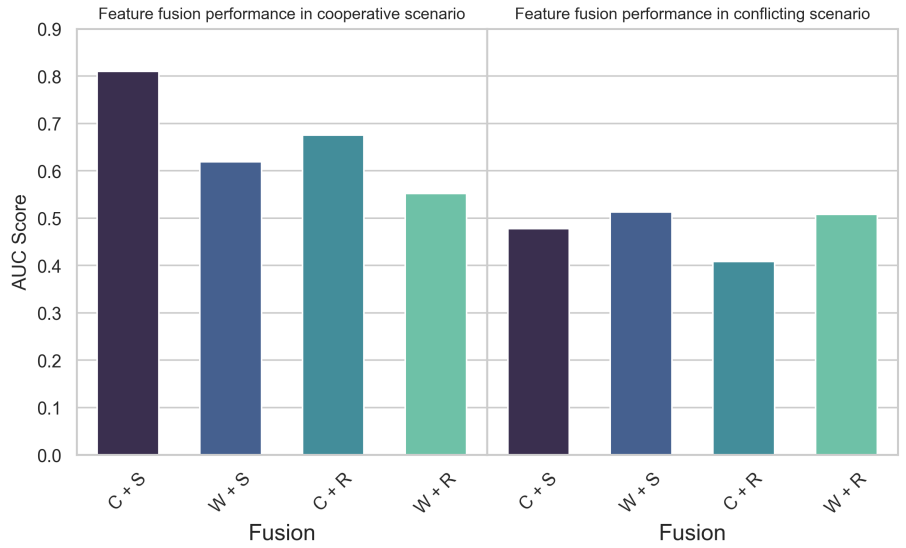

Figure 19. AUC scores from the fused features per scenario in case of a random split.

Model	Scenario	AUC (Default Weight)	AUC (Class Weight)
C + S	**Vacation**	0.810	0.778
	Conflict	0.478	0.523
W + S	**Vacation**	0.619	0.648
	Conflict	0.513	0.525
C + R	**Vacation**	0.675	0.560
	Conflict	0.408	0.449
W + R	**Vacation**	0.552	0.558
	Conflict	0.508	0.575

Table 11. AUC scores from feature selection per scenario using random split in multimodal fusion.

Model	Scenario	AUC (Default Weight)	AUC (Class Weight)
C + S	**Vacation**	0.648	0.632
	Conflict	0.480	0.481
W + S	**Vacation**	0.493	0.465
	Conflict	0.490	0.518
C + R	**Vacation**	0.508	0.531
	Conflict	0.324	0.451
W + R	**Vacation**	0.426	0.438
	Conflict	0.456	0.468

Table 12. AUC scores from feature selection per scenario using k-fold split in multimodal fusion.

Wav2Vec2.0.

Tables 15 and 16 provide information about the results for the text modality, showing the classification performance of deep features from SBERT and RobBERT. The highest AUC scores in this case are 0.681 from SBERT features, while AUC score from RobBERT features is significantly lower equal to 0.404 in case of a k-fold split is applied. As Figure 21 shows, across scores obtained by different splits, SBERT outperforms RobBERT.

Figure 21 represents the performance of the unimodal features, for text and audio in this case. Each bar corresponds to the highest performance per modality and feature representations. The comparison is crucial for understanding how the modalities perform individually and understanding the strengths and limitations of each modality.

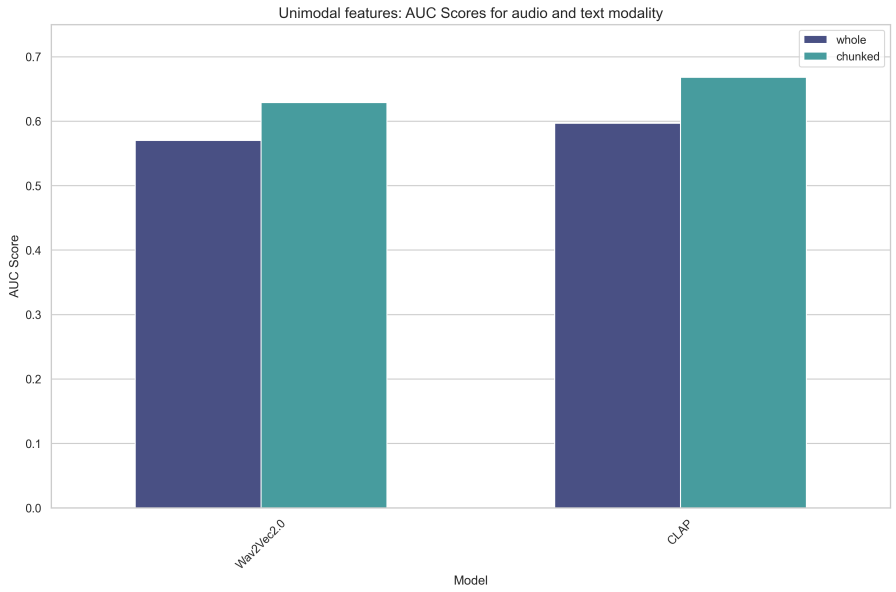

Figure 20. AUC scores from different feature extraction strategy from audio models.

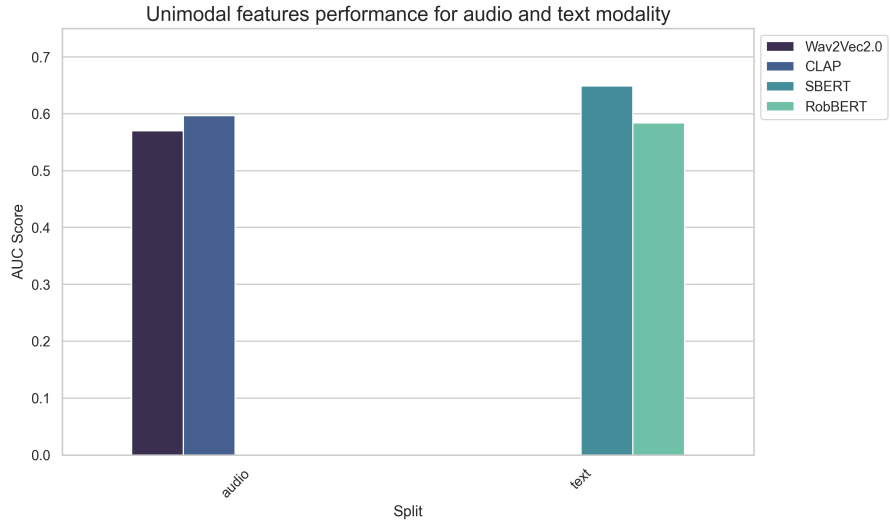

Figure 21. AUC scores from features per modality across all scenarios.

Model	Audio	AUC (Default Weight)	AUC (Class Weight)
Wav2Vec2.0	**whole**	0.569	0.570
	chunked	0.609	0.629
CLAP	**whole**	0.562	0.597
	chunked	0.651	0.668

Table 13. Comparison of AUC scores of audio features using random split with default and class weights.

Model	Audio	AUC (Default Weight)	AUC (Class Weight)
Wav2Vec2.0	**whole**	0.510	0.504
	chunked	0.551	0.581
CLAP	**whole**	0.398	0.404
	chunked	0.447	0.553

Table 14. Comparison of AUC scores of each fusion combination using k-fold split with default and class weights.

6.4 Results of using parent segments

The parent segments did not perform as well as the child segments. The labels in this task are still the same as the ones before, namely the presence or absence of children's depressive symptoms. while the features were extracted from the parents' segments. Regarding the audio modality, the features extracted from the models produced AUC scores of 0.603 and 0.589 for Wav2Vec2.0 and CLAP, respectively. For the text modality, the scores were marginally lower than those from the audio modality. SBERT features resulted in an AUC score of 0.528, while RobBERT achieved 0.550.

Tables 37 and 38 present parents' segments performance in audio modality and Tables 35 and 36 the performance as this resulted from the text modality.

The overall image of the parents' segments results can be seen in Figure 22.

Model	AUC (Default Weight)	AUC (Class Weight)
SBERT	0.647	0.649
RobBERT	0.551	0.584

Table 15. Comparison of AUC scores of text features using random split with default and class weights.

Model	AUC (Default Weight)	AUC (Class Weight)
SBERT	0.681	0.649
RobBERT	0.404	0.554

Table 16. Comparison of AUC scores of text features using k-fold split with default and class weights.

Model	AUC (Default Weight)	AUC (Class Weight)
Wav2Vec2.0	0.575	0.548
CLAP	0.571	0.554

Table 17. Comparison of AUC scores of audio features extracted from parents' segments using random split with default and class weights.

Model	AUC (Default Weight)	AUC (Class Weight)
Wav2Vec2.0	0.603	0.582
CLAP	0.589	0.572

Table 18. Comparison of AUC scores of audio features extracted from parents' segments using k-fold split with default and class weights.

Model	AUC (Default Weight)	AUC (Class Weight)
SBERT	0.418	0.418
RobBERT	0.469	0.415

Table 19. Comparison of AUC scores of text features extracted from parents' segments using random split with default and class weights.

Model	AUC (Default Weight)	AUC (Class Weight)
SBERT	0.505	0.528
RobBERT	0.550	0.568

Table 20. Comparison of AUC scores of text features extracted from parents' segments using k-fold split with default and class weights.

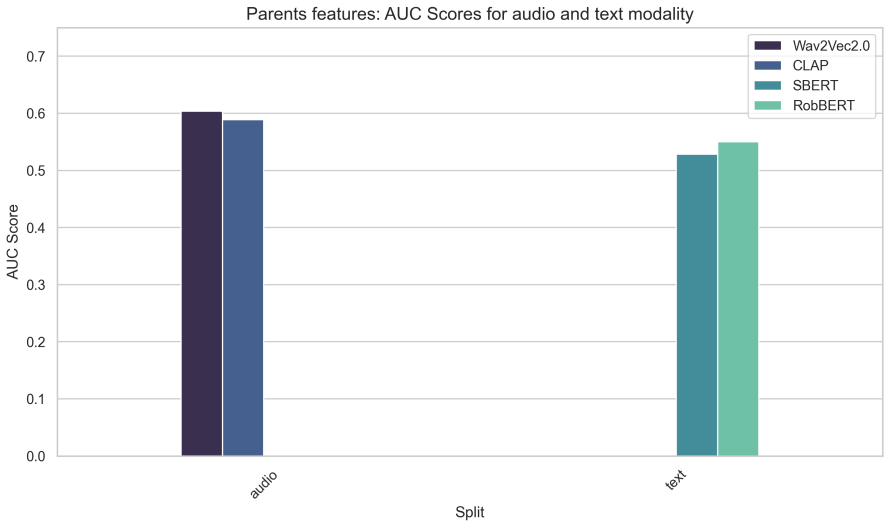

Figure 22. AUC scores from features of the best performing models from parents' segments features.

7. Discussion

The main goal of this research was to find an objective way to assess depression symptoms in young children. Two modalities were explored as indicators of this research, namely deep embeddings from audio segments and text transcripts were used as features in the classification task . Assessing depressive symptoms in children is a quite challenging task and this research contributes in the effort of early assessment of depressive symptoms. As Allgaier et al. [4] mentioned, depression in children is rarely studied despite the importance of early detection.

Identifying the first depressive symptoms in young children is particularly challenging due to the heterogeneity of the disease—there is no single symptomatology associated with depression. Rather than that, depression could be overshadowed by comorbid mental disorders, leading to it being undetected and untreated. In this study, detecting depressive symptoms can significantly improve the early assessment of the disease by specialists and potentially reduce the number of undetected cases.

Accurately assessing depressive symptoms could serve as a referral for a specialist, who can then confirm or refute the suspicion of depression. It is interesting to see how the deep features from large language and audio models can contribute to this field and how effectively deep learning can identify patterns of depression. The results of this study show that some specific scenarios could offer an AUC score greater than 0.7. Interesting findings derived from the results will be explained in the following sections, where each research question will be explored in detail.

7.1 Addressing the research questions

Research question: To what extent can multimodal embeddings obtained from large audio and language models detect depressive symptoms in children from parent-child interaction videos?

We found that embeddings obtained from large language and large audio models can moderately predict the depressive symptoms in children. No certain conclusion can be drawn, as most of the results are slightly better than predicting by chance. A main reason for this is the size of the dataset. The best results were obtained when the fusion of deep representations from the CLAP and SBERT were fed in the XGBoost classifier in the cooperative scenario. These embeddings reached a significant AUC score of 0.810.

Similarly to the study of Shen et al. [100], combining the features of the audio and text modality performed with slightly higher AUC scores than the individual modalities.

The combination of models (S + C) performed better in both splits, as shown in the Figure 19. This combination results in a fused feature vector with 1280 dimension, the smallest among the other fused feature vectors. Additionally, the chunked approach performed with higher AUC scores, implying that the model learns better when the features were extracted from smaller parts of audio. Further discussion of these points follows later along with the addressing of the sub-research questions.

Sub-research question 1: To what extent can depressive symptoms in children be assessed separately in both cooperative and conflicting scenarios?

As mentioned earlier, we achieved the highest AUC scores when both cooperative and conflicting scenarios were studied independently. The fusion of SBERT and CLAP reached an AUC score equal to 0.810. In the audio modality, the classification task with CLAP embeddings in the cooperative scenario reached an AUC score equal to 0.746. Both scores imply that the model does not predict by chance and the performance is considerable, according to Çorbacıoğlu et al. [21]. In their work they set some thresholds in the interpretation of AUC score, with AUC scores greater or equal to 80 till 90 to be characterized as considerable.

Each scenario was explored independently in this task, yielding the best and most interesting results. In the audio modality, CLAP features overperformed Wav2Vec2.0 features with more than double AUC score. Specifically, in the cooperative scenario, CLAP embeddings reached AUC score equal to 0.746, while Wav2Vec2.0 scored 0.393 in the random split and class weight setup. Similar scores and differences were observed in the default weight (see Table 7). In the k-fold split setup, the differences in AUC scores were smoother, but CLAP still outperformed Wav2Vec2.0 in the cooperative scenario.

In the conflicting scenario, the opposite trend was observed, with Wav2Vec2.0 embeddings resulting in higher AUC scores than CLAP features. In a random split setup with default weight, Wav2Vec2.0 features achieved an AUC score of 0.612, while CLAP reached 0.483. This trend was consistent in the conflicting scenario.

For the text modality, the highest score in a k-fold setup was obtained by SBERT in the cooperative scenario, with an AUC of 0.680, outperforming RobBERT by more than 0.20 (AUC equal to 0.424). SBERT outperformed RobBERT in the cooperating scenario, while RobBERT performed better in the conflicting scenario. For instance, in the k-fold setup,

RobBERT features achieved an AUC score of 0.720 in the conflicting scenario, compared to SBERT's 0.552 (see Table 10).

Lastly, multimodal fusion vectors performed exceptionally well, reaching an AUC score of 0.810 in the cooperative scenario. This score was obtained using a random split setup, which utilizes only a part of the dataset for testing. In contrast, a k-fold setup would provide a better indication of the model's generalization ability, as it evaluates the model across multiple subsets of the data. Models with an AUC score above 0.8 are very good at distinguishing the positive class from the negative. Overall, fusion performed better when the PCI task was related to planning a vacation in a random split setup, with its performance varying in AUC score terms from 0.810 to 0.551. The poorest performance was achieved by (W + R). An opposite pattern was observed in the conflict-related PCI task, with (W + S) and (W + R) performing better than (S + C).

The patterns suggest that depressive symptoms can better be predicted from the cooperative scenario with smaller feature vectors, while they can be better predicted with large feature vectors from conflict scenario. Models with smaller feature vectors, such CLAP and SBERT, were better at predicting depressed symptoms than RobBERT and Wav2Vec2.0 in both modalities. In addition to achieving an outstanding score of 0.810, the combination of CLAP and SBERT was the best in predicting the cooperative task. With 1280 dimensions, the combined feature vector of (S + C) is the smallest.

For the conflict task, the opposite trend has been observed. Larger feature vectors resulted in higher AUC scores. (R + W), with 2048 dimensions, achieved the best AUC score for the conflict scenario in both random split and k-fold split setups. RobBERT features obtained a high AUC score of 0.720 in the conflicting scenario, the highest score achieved by RobBERT features in a k-fold split.

In the audio modality, Wav2Vec2.0 embeddings achieved better scores in the conflicting scenario than the cooperative one, and the Wav2Vec2.0 feature vector is larger than the CLAP one. After considering all the results above and comparing them among modalities and splits, it is noticeable that depressive symptoms can better be predicted from the cooperative scenario with smaller feature vectors, while the prediction of depressive symptoms can better be predicted with large feature vectors from conflict scenario.

Sub-research question 2: To what extent can depressive symptoms in children be accurately assessed using solely the audio and text embeddings?

This task explores each modality separately. Regarding the audio modality, CLAP is

performing better in tasks across all the scenarios. Although it achieved an AUC score of 0.597 in a random split, this is much lower than the performance CLAP features achieved per scenario. Similarly, in a random split setup, SBERT performed better than RobBERT features, with the first one reaching an AUC of 0.649 and RobBERT achieving 0.584. When comparing the results in the same split setup between the combined features and the unimodal features, the combined features performed better. Both modalities combined proved to be more effective in assessing depressive symptoms in children.

This part of the results yields another interesting finding in audio modality only. As shown in Tables 13 and 14, the performance of features extracted from the whole audio versus the chunked audio was compared. In both splits, the chunked version of the features outperformed the whole audio.

Sub-research question 3: How does examining only the parent's segment impact the assessment of depressive symptoms in children?

Apart from the segments extracted from the child, the parent's segment and transcript were explored as well. As Allgaier et al. [4] indicated, children of depressive are exposed to a vulnerable environment and thus are at a higher risk of developing psychopathology. That was the motivation behind this last part of the thesis. In this task, the features extracted from parents' segments were utilised and the ground truth labels remained the same as before, coming from the symptoms parents reported for their children. The overall performance of audio modality was higher than the one achieved by text modality in both random and k-fold split setups. For instance, Wav2Vec2.0 and SBERT features reached AUC scores of 0.603 and 0.589 respectively, in comparison to SBERT and RobBERT, which achieved 0.505 and 0.550. No special patterns noticed in this group of results, thus exploring the parents' segments did not add any significant value in the exploration of the task.

7.2 General observations and implications

The overall performance of CLAP embeddings compared to Wav2Vec2.0 ones is very impressive and surprising, as CLAP has not been trained on dataset containing human speech, it mostly contained pairs of text and audio with human activity among natural sounds and audio effects. Also, the CLAP model used as a feature extractor in this thesis was not finetuned in Dutch, while the Wav2Vec2.0 version was finetuned in Dutch, which should have given it an advantage in performance. This is an interesting finding, as studies in similar tasks have never utilised CLAP embeddings. On the other hand, there are studies which utilise Wav2Vec2.0 features for depression assessment task and other tasks like this thesis.

Returning to the main research topic, our study shows that it is quite feasible to evaluate children's depressive symptoms using multimodal embeddings from large audio and language models, particularly when the cooperative and conflicting scenarios are analysed separately. While our dataset was limited to 191 data points, the multimodal fusion of characteristics generated a significant AUC score of 0.81 in a default split setup. Thus, in that case the model was able to distinguish between different levels of depression symptoms with a reasonable accuracy, but only in a part of the database.

This shows that multimodal techniques combining both audio and text deep representations hold significant promise for diagnosing depression symptoms in children. Importantly, our findings suggest that given a larger and more balanced dataset, these models may recognize even more intricate patterns, resulting in more accurate and robust predictions.

Future research should focus on increasing the dataset and fine-tuning these multimodal models to improve their prediction ability. Furthermore, combining additional modalities like video or image, may provide deeper insights and improve the assessment of depressed symptoms in children. This work demonstrates the potential of multimodal embeddings in mental health evaluation and emphasises the need for additional research in this promising area.

7.3 Limitations & Future Work

There were several limiting factors in this research. First and most important limitation was the limited dataset size. Machine learning models require substantial amounts of data to effectively learn and recognize patterns. As a result, having more data to train the model would greatly enhance the performance of the proposed framework.

Another limitation relating to the dataset, was the quality of the recordings. As the study has followed different generations of children, the recordings were made using different setups and technologies over time. Some recordings are of poor quality, missing subtle changes in voice, tone or volume, details that could contribute to the results.

Focusing now on another part of the dataset, the annotations, which are made by the parents' perception of their children's behavior. A more accurate ground truth might be obtained if children reported the answers for themselves. In such cases, the ground truth labels would reflect the children's self-perceptions.

Another limitation occurs also in case of the mean values used for the label transformation (see Section 5.1). The dataset transformation to binary labels based on cut-off points

used mean values from the US population. However, this research focuses on the Dutch population and mean values for the Dutch population were not available to us. Cultural differences imply that cut-off points for the Dutch population would likely differ, potentially impacting the results.

Last but not least, regarding the models used, many models were not finetuned in the Dutch language. Features extracted could be more accurate with a Dutch-specific fine-tuned model, such as a Dutch-fine-tuned CLAP. Additionally, the popular Dutch language model Fietje from HuggingFace was not used as it did not provide any feature extraction module. It would be interesting to utilize a model finetuned in such a big amount of Dutch data and compare the results.

An interesting addition for future work could be the inclusion of the video modality. Examining how video data contributes to the overall performance of the model could provide valuable insights and potentially improve the accuracy of detecting depressive symptoms in children. For example, in their recent work, Zhang et al. [122] utilized the fusion of three modalities: video, audio, and text. The video modality offered valuable information related to facial expressions, which contributed to better performance compared to unimodal features or other combinations. This was particularly evident in their interview process when participants were asked, "Can you evaluate yourself?". Similarly, the videos available in this dataset could be utilized to enhance the task.

Another task to consider as part of future work is to integrate explainability mechanisms to better understand the meaning of the deep representations. Since these features were extracted by the large language and audio models for each audio and transcript, there is no clear meaning behind them in human terms. Therefore, it would be valuable to have an interpretation behind these numerical vectors, in a way similar to how handcrafted features can be interpreted, providing more transparency and insight into the models' learned representations.

8. Conclusion

To conclude, this thesis has studied the assessment of depressive symptoms in children through PCI interaction. Deep features were extracted from large audio and language models and combined across two modalities for different scenarios: cooperative and conflicting. The predictions were performed with an XGBoost classifier model, which served the different tasks explored, such as feature fusion, scenario- specific predictions, or modality-specific predictions.

We observed various interesting patterns in both the cooperative and conflicting scenarios. Notably, in the cooperative scenario, the fusion model performed well, regardless the limited data available. Additionally, an improvement in performance was noticed when the features were combined. However, we were not able to answer the research question "To what extent can multimodal embeddings obtained from large audio and language models detect depressive symptoms in children from parent-child interaction videos?" posed in the beginning of this study with great certainty, due to the limited data, especially for training models with deep features. Nevertheless, our findings contribute to the research in this area and may improve with more data.

In conclusion, our study suggests that depressive symptoms are better assessed when there is a specific task behind the PCI, either conflicting or cooperative. Better results were obtained with the cooperative scenario, but both scenarios provided valuable information for the final prediction. Moreover, given the multi-faceted nature of depressive symptoms, multimodal fusion enhances the accuracy of depressive symptom detection. The findings are promising, indicating that with more data, there could be a great contribution in the assessment of depressive symptoms in children.

9. Acknowledgments

I would like to thank my main supervisor, Dr. Itir Önal Ertugrul, for her invaluable support, innovative ideas, and the insightful discussions during our weekly meetings. I highly appreciated the calmness and the grouding she offered me when I needed that. Additionally, I am thankful to my daily supervisor, Mang Ning, M.Sc., for his technical expertise and constructive feedback on my ideas. Your mentorship has greatly enriched my project—thank you both! I am also grateful to Prof. Albert Salah for his insightful feedback on my thesis proposal and his interest in my work.

To my dear friends Max, Ian, and Admitos, thank you for your support, the long study hours we shared, and the discussions that helped me through moments of doubt. A big thanks to Menandros, Lydia, my friends and my beloved family—Kyriaki, Stavroula, my parents, and my sweet grandmothers. Your support has been invaluable.

10. Appendix: Additional results

In the following section, Random Split refers to GroupShuffleSplit method, k-fold Split refers to GroupKFold method and Stratified Split refers to StratifiedGroupKFold.

10.1 Results of multimodal fusion

Tables 21 and 22 shows the results from the fusion of all the models for all the different splits performed in this thesis. Table 21 shows the results when the default weight was used in the XGBoost classifier, while Table 22 shows the results when the class weight was set depending on the task.

10.2 Results per scenario

This section presents the results from the experiments ran for each scenario independently, highlighting the performance metrics of various classifiers utilized in the analysis. In the tables below, we present the results from both the modalities explicitly and the multimodal fusion.

Tables 23 and 24 shows the results of the classification tasks per scenario when it trained on SBERT and RobBERT features for all the different splits performed in this thesis. 23 shows the results when the default weight was used in the XGBoost classifier, while the 24 shows the results when the class weight was set depending on the task.

In the following part of the section, results from the audio modality for the scenario study case are presented. Tables 25 and 26 show the results from the classification tasks, when the model trained on Wav2Vec2.0 and CLAP features, for each scenario explicitly.

Last but not least, the results of the multimodal fusion for the scenario study case are presented in Table 27 for different split and default weights, while in Table 28 we present the results different splits produced with the class weights.

10.3 Unimodal results

This section presents the comprehensive results of all classification tasks conducted utilizing unimodal features, specifically focusing on audio-only and language-only data. The outcomes are organized in a series of tables.

Table 21. Classification results using the fusion of the models with default weights.

Model	Classifier	Accuracy	Precision	Recall	F1	AUC
C + S	Random Split	0.6238	0.5203	0.3887	0.4449	0.6115
	k-fold Split	0.6067	0.4801	0.3156	0.3808	0.6316
	Stratified Split	0.6132	0.4776	0.2790	0.3522	0.5960
W + S	Random Split	0.5474	0.4158	0.3043	0.3514	0.5500
	k-fold Split	0.5551	0.3802	0.2834	0.3247	0.5818
	Stratified Split	0.6072	0.4724	0.3448	0.3986	0.6359
C + R	Random Split	0.5713	0.4029	0.2927	0.3391	0.5677
	k-fold Split	0.5599	0.3891	0.2798	0.3255	0.5608
	Stratified Split	0.6444	0.5833	0.2371	0.3372	0.5533
W + R	Random Split	0.5646	0.4084	0.3345	0.3678	0.5708
	k-fold Split	0.5713	0.3970	0.2703	0.3216	0.4466
	Stratified Split	0.5762	0.3804	0.1819	0.2461	0.5141

Table 22. Classification results using the fusion of various models with class weights.

Model	Classifier	Accuracy	Precision	Recall	F1	AUC
C + S	Random Split	0.5784	0.4718	0.3742	0.4174	0.5844
	k-fold Split	0.6174	0.4731	0.3930	0.4293	0.6801
	Stratified Split	0.6181	0.4871	0.3886	0.4323	0.6672
W + S	Random Split	0.5432	0.4084	0.3545	0.3795	0.5925
	k-fold Split	0.5602	0.4455	0.3357	0.3829	0.5536
	Stratified Split	0.5707	0.4095	0.2781	0.3312	0.5742
C + R	Random Split	0.5152	0.3528	0.3270	0.3394	0.5277
	k-fold Split	0.5449	0.3907	0.3396	0.3634	0.5872
	Stratified Split	0.6333	0.5309	0.3324	0.4088	0.5598
W + R	Random Split	0.5332	0.3516	0.2388	0.2844	0.5670
	k-fold Split	0.5493	0.4030	0.3043	0.3468	0.4889
	Stratified Split	0.5287	0.3489	0.1810	0.2384	0.4175

Table 23. Classification results per scenario using SBERT and RobBERT features with default weights.

Model	Scenario	Classifier	Accuracy	Precision	Recall	F1	AUC
SBERT	Vacation	Random Split	0.6067	0.4028	0.3509	0.3751	0.6443
		k-fold Split	0.6442	0.6192	0.5173	0.5637	0.6760
		Stratified Split	0.6426	0.5440	0.4429	0.4883	0.7545
	Conflict	Random Split	0.5310	0.5036	0.2881	0.3665	0.4946
		k-fold Split	0.5611	0.4405	0.1900	0.2655	0.5520
		Stratified Split	0.5526	0.2000	0.1107	0.1425	0.4236
RobBERT	Vacation	Random Split	0.5600	0.3572	0.2541	0.2970	0.4896
		k-fold Split	0.5621	0.3143	0.1896	0.2365	0.4811
		Stratified Split	0.5611	0.3655	0.2571	0.3019	0.5303
	Conflict	Random Split	0.5517	0.4622	0.3295	0.3847	0.5506
		k-fold Split	0.5621	0.4667	0.2033	0.2832	0.7204
		Stratified Split	0.5405	0.2950	0.2179	0.2507	0.5357

Table 24. Classification results per scenario using SBERT and RobBERT features with class weights.

Model	Scenario	Classifier	Accuracy	Precision	Recall	F1	AUC
SBERT	Vacation	Random Split	0.6267	0.4414	0.5006	0.4691	0.6699
		k-fold Split	0.6637	0.6579	0.4506	0.9011	0.6803
		Stratified Split	0.6011	0.4786	0.4714	0.4750	0.7237
	Conflict	Random Split	0.5103	0.4979	0.1989	0.2842	0.4786
		k-fold Split	0.5826	0.4757	0.2233	0.3039	0.4942
		Stratified Split	0.5216	0.2067	0.0821	0.1175	0.3001
RobBERT	Vacation	Random Split	0.5867	0.4171	0.2907	0.3426	0.5917
		k-fold Split	0.4989	0.3405	0.2411	0.2823	0.4241
		Stratified Split	0.5100	0.3138	0.2607	0.2848	0.4688
	Conflict	Random Split	0.5448	0.4083	0.2780	0.3308	0.5060
		k-fold Split	0.5411	0.3833	0.1833	0.2480	0.6721
		Stratified Split	0.5516	0.3000	0.1929	0.2348	0.5823

Table 25. Classification results per scenario using Wav2Vec2.0 and CLAP features with default weights.

Model	Scenario	Classifier	Accuracy	Precision	Recall	F1	AUC
Wav2Vec2.0	Vacation	Random Split	0.5655	0.3533	0.2001	0.2555	0.3670
		k-fold Split	0.5516	0.1400	0.0650	0.0888	0.4609
		Stratified Split	0.5000	0.2155	0.2571	0.2345	0.2858
	Conflict	Random Split	0.5931	0.3671	0.3054	0.3334	0.6108
		k-fold Split	0.6526	0.6267	0.3780	0.4716	0.6103
		Stratified Split	0.6211	0.5010	0.4286	0.4620	0.5516
CLAP	Vacation	Random Split	0.6138	0.3032	0.2027	0.2430	0.7132
		k-fold Split	0.6032	0.3700	0.1893	0.2505	0.5680
		Stratified Split	0.5747	0.4367	0.3500	0.3886	0.3518
	Conflict	Random Split	0.5793	0.2236	0.1706	0.1935	0.4832
		k-fold Split	0.5368	0.0667	0.0400	0.0500	0.3860
		Stratified Split	0.5684	0.3817	0.2571	0.3072	0.4127

Table 26. Classification results per scenario using Wav2Vec2.0 and CLAP features with class weights.

Model	Scenario	Classifier	Accuracy	Precision	Recall	F1	AUC
Wav2Vec2.0	Vacation	Random Split	0.4483	0.2525	0.1909	0.2174	0.3925
		k-fold Split	0.4889	0.3384	0.3236	0.3308	0.4600
		Stratified Split	0.4689	0.2738	0.2107	0.2381	0.2858
	Conflict	Random Split	0.6000	0.2971	0.2571	0.2757	0.5974
		k-fold Split	0.6105	0.4583	0.2130	0.2908	0.5990
		Stratified Split	0.5789	0.4167	0.2286	0.2952	0.5516
CLAP	Vacation	Random Split	0.5931	0.5029	0.6044	0.5490	0.7456
		k-fold Split	0.6453	0.4476	0.4586	0.4530	0.6014
		Stratified Split	0.5842	0.4610	0.4286	0.4442	0.3287
	Conflict	Random Split	0.5931	0.2178	0.1071	0.1436	0.4268
		k-fold Split	0.6105	0.2000	0.0622	0.0949	0.4147
		Stratified Split	0.5789	0.3133	0.1429	0.1963	0.5040

Table 27. Classification results per scenario using multimodal fusion features with default weights.

Model	Scenario	Classifier	Accuracy	Precision	Recall	F1	AUC
C + S	Vacation	Random Split	0.6552	0.5823	0.4660	0.5177	0.8100
		k-fold Split	0.6347	0.5238	0.4072	0.4582	0.6483
		Stratified Split	0.6153	0.4914	0.3786	0.4277	0.5216
	Conflict	Random Split	0.5793	0.3209	0.2409	0.2752	0.4780
		k-fold Split	0.5864	0.3583	0.2012	0.2577	0.4802
		Stratified Split	0.5368	0.4200	0.2000	0.2710	0.3095
W + S	Vacation	Random Split	0.5793	0.4624	0.3774	0.4156	0.6190
		k-fold Split	0.5511	0.4967	0.3006	0.3745	0.4929
		Stratified Split	0.5926	0.5452	0.2929	0.3811	0.6139
	Conflict	Random Split	0.5793	0.3306	0.2465	0.2824	0.5125
		k-fold Split	0.5684	0.3024	0.1726	0.2198	0.4899
		Stratified Split	0.5789	0.3833	0.2000	0.2628	0.4841
C + R	Vacation	Random Split	0.5931	0.4608	0.3758	0.4140	0.6752
		k-fold Split	0.6053	0.5750	0.3561	0.4398	0.5081
		Stratified Split	0.5742	0.4976	0.2714	0.3512	0.3599
	Conflict	Random Split	0.4897	0.2517	0.2331	0.2420	0.4084
		k-fold Split	0.5579	0.4400	0.2060	0.2806	0.3241
		Stratified Split	0.5263	0.2738	0.1714	0.2108	0.4960
W + R	Vacation	Random Split	0.5517	0.3838	0.2226	0.2818	0.4997
		k-fold Split	0.6363	0.6000	0.2483	0.3512	0.4258
		Stratified Split	0.5000	0.1721	0.1679	0.1670	0.3516
	Conflict	Random Split	0.5724	0.3222	0.2753	0.2970	0.5077
		k-fold Split	0.5263	0.3244	0.3631	0.3427	0.4560
		Stratified Split	0.6000	0.4605	0.2857	0.3526	0.5833

Table 28. Classification results per scenario using multimodal fusion features with class weights.

Model	Scenario	Classifier	Accuracy	Precision	Recall	F1	AUC
C + S	Vacation	Random Split	0.6621	0.5784	0.5720	0.5752	0.7777
		k-fold Split	0.6242	0.5143	0.4294	0.4680	0.6320
		Stratified Split	0.6474	0.5707	0.4357	0.4941	0.4972
	Conflict	Random Split	0.6000	0.4244	0.2076	0.2788	0.5233
		k-fold Split	0.5368	0.1917	0.1393	0.1614	0.4805
		Stratified Split	0.5684	0.1800	0.1429	0.1593	0.3571
W + S	Vacation	Random Split	0.6069	0.4802	0.4695	0.4748	0.6475
		k-fold Split	0.5405	0.4694	0.2783	0.3494	0.4647
		Stratified Split	0.5195	0.3394	0.2929	0.3144	0.4975
	Conflict	Random Split	0.5517	0.3296	0.2498	0.2842	0.5249
		k-fold Split	0.6316	0.5524	0.2262	0.3210	0.5177
		Stratified Split	0.5684	0.3967	0.2000	0.2659	0.5040
C + R	Vacation	Random Split	0.5517	0.3711	0.3249	0.3465	0.5604
		k-fold Split	0.5953	0.5514	0.4478	0.4942	0.5313
		Stratified Split	0.5426	0.3644	0.3250	0.3436	0.3830
	Conflict	Random Split	0.4828	0.1739	0.2268	0.1969	0.4492
		k-fold Split	0.5789	0.3800	0.1726	0.2374	0.4508
		Stratified Split	0.5789	0.3600	0.2000	0.2571	0.5040
W + R	Vacation	Random Split	0.5655	0.4014	0.2907	0.3372	0.5575
		k-fold Split	0.6279	0.5433	0.3644	0.4362	0.4377
		Stratified Split	0.4474	0.1221	0.1107	0.1161	0.3810
	Conflict	Random Split	0.5931	0.3611	0.3240	0.3415	0.5754
		k-fold Split	0.5895	0.4500	0.3786	0.4112	0.4675
		Stratified Split	0.6632	0.5667	0.3429	0.4273	0.6032

Tables 29 and 30 present the results of the classification tasks trained on RobBERT and SBERT features on various splits conducted in this thesis, across all scenarios. Same convention as before with Table 29 displays the outcomes when the default weight was applied in the XGBoost classifier, while Table 30 illustrates the results when class weights were adjusted according to the specific tasks.

Similarly, to the results presented from the language features, results from classification of audio features are presented below. Tables below show the performance of Wav2Vec2.0 and CLAP features when these were input to XGBoost classifier in both cases of default weights 31 and class weights 32.

For the audio modality, a second experiment has been conducted with chunked audio. Results from the second experiment can be seen in the Tables 33 and 34

10.4 Results of using parent segments

This section presents the comprehensive results of all classification tasks conducted utilizing features extracted from the parents' audios and transcripts.

Tables 35 and 36 present the results of the classification tasks trained on SBERT and RobBERT features across various splits conducted for transcripts extracted from parents' segments with default and class weights.

The classification results of the audio features extracted from Wav2Vec2.0 and CLAP for the parents' audio segments can be found in the tables below. Table 37 contains the results when default weight was utilised, while Table 38 shows the results of the task with the corresponding class weight.

Table 29. Classification results using SBERT and RobBERT with default weights.

Model	Classifier	Accuracy	Precision	Recall	F1	AUC
SBERT	Random Split	0.5845	0.4726	0.3245	0.3848	0.6474
	k-fold Split	0.6495	0.6139	0.3186	0.4195	0.6807
	Stratified Split	0.6289	0.5249	0.2971	0.3794	0.6452
RobBERT - last layer	Random Split	0.5676	0.3633	0.1482	0.2105	0.5514
	k-fold Split	0.5105	0.2328	0.1368	0.1723	0.4039
	Stratified Split	0.5568	0.3233	0.1476	0.2027	0.5269

Table 30. Classification results using SBERT and RobBERT with class weights.

Model	Classifier	Accuracy	Precision	Recall	F1	AUC
SBERT	Random Split	0.6158	0.5448	0.4083	0.4668	0.6491
	k-fold Split	0.6081	0.4899	0.3762	0.4256	0.6486
	Stratified Split	0.5978	0.4709	0.3781	0.4194	0.6073
RobBERT - last layer	Random Split	0.5707	0.4346	0.2563	0.3224	0.5841
	k-fold Split	0.5468	0.3787	0.2331	0.2886	0.5537
	Stratified Split	0.5467	0.3501	0.2438	0.2874	0.5246

Table 31. Classification results using Wav2Vec2.0 and CLAP with default weights.

Model	Classifier	Accuracy	Precision	Recall	F1	AUC
Wav2Vec2.0	Random Split	0.5754	0.3282	0.1502	0.2061	0.5689
	k-fold Split	0.5598	0.3624	0.2250	0.2776	0.5098
	Stratified Split	0.5762	0.3988	0.2229	0.2860	0.4241
CLAP	Random Split	0.5574	0.2955	0.2272	0.2569	0.5619
	k-fold Split	0.5232	0.3269	0.1922	0.2421	0.3980
	Stratified Split	0.5973	0.4524	0.2343	0.3087	0.5458

Table 32. Classification results using Wav2Vec2.0 and CLAP with class weights.

Model	Classifier	Accuracy	Precision	Recall	F1	AUC
Wav2Vec2.0	Random Split	0.5611	0.4374	0.3494	0.3885	0.5702
	k-fold Split	0.5336	0.3397	0.2553	0.2915	0.5040
	Stratified Split	0.5660	0.4293	0.3067	0.3578	0.4041
CLAP	Random Split	0.5683	0.4144	0.3742	0.3933	0.5967
	k-fold Split	0.4814	0.2613	0.2136	0.2351	0.4044
	Stratified Split	0.5973	0.4333	0.1914	0.2655	0.4944

Table 33. Classification results using Wav2Vec2.0 and CLAP features extracted from chunked audios with default weights.

Model	Strategy	Classifier	Accuracy	Precision	Recall	F1	std. acc	std. prec	std. rec	AUC
Wav2Vec2.0	Primary	Random Split	0.5818	0.2802	0.1464	0.1923	0.0444	0.2083	0.1062	0.609
		k-fold Split	0.5862	0.3810	0.1121	0.1732	0.0361	0.2903	0.0998	0.5509
		Stratified Split	0.5653	0.2567	0.0695	0.1094	0.0368	0.0760	0.0026	0.5403
Wav2Vec2.0	At-least-1	Random Split	0.4567	0.3753	0.5303	0.4395	0.0609	0.0953	0.1189	0.609
		k-fold Split	0.4343	0.3659	0.5400	0.4362	0.1170	0.1293	0.0563	0.5509
		Stratified Split	0.4609	0.3712	0.6114	0.4619	0.0983	0.0794	0.1862	0.5403
CLAP	Primary	Random Split	0.5963	0.4482	0.2564	0.3262	0.0616	0.1722	0.1536	0.6512
		k-fold Split	0.5497	0.3683	0.2122	0.2693	0.0575	0.2329	0.2015	0.447
		Stratified Split	0.6175	0.3595	0.2314	0.2816	0.0700	0.3367	0.2172	0.6278
CLAP	At-least-1	Random Split	0.4428	0.3754	0.6004	0.4620	0.0514	0.0714	0.1407	0.6512
		k-fold Split	0.4185	0.3471	0.6150	0.4438	0.1752	0.1569	0.1485	0.447
		Stratified Split	0.4821	0.3975	0.6374	0.4896	0.0515	0.0360	0.2018	0.6278

Table 34. Classification results using Wav2Vec2.0 and CLAP features extracted from chunked audios with class weights.

Model	Strategy	Classifier	Accuracy	Precision	Recall	F1	std. acc	std. prec	std. rec	AUC
Wav2Vec2.0	Primary	Random Split	0.5614	0.2413	0.1557	0.1893	0.0935	0.1878	0.1503	0.6291
		k-fold Split	0.5700	0.4226	0.2793	0.3363	0.0744	0.2172	0.1992	0.5814
		Stratified Split	0.5551	0.3155	0.1657	0.2173	0.0354	0.0923	0.0758	0.5677
Wav2Vec2.0	At-least-1	Random Split	0.4604	0.3962	0.5919	0.4747	0.1198	0.1392	0.0452	0.6291
		k-fold Split	0.4343	0.3706	0.6120	0.4616	0.0822	0.8739	0.1147	0.5814
		Stratified Split	0.4811	0.4004	0.7132	0.5129	0.1191	0.0906	0.2137	0.5677
CLAP	Primary	Random Split	0.5926	0.4744	0.2856	0.3565	0.0651	0.1736	0.1528	0.668
		k-fold Split	0.5599	0.4374	0.3018	0.3572	0.0675	0.1430	0.1629	0.5527
		Stratified Split	0.5549	0.3659	0.2476	0.2953	0.0561	0.0991	0.1259	0.5685
CLAP	At-least-1	Random Split	0.4118	0.3602	0.6268	0.4575	0.0626	0.0553	0.1748	0.668
		k-fold Split	0.4501	0.3900	0.6859	0.4973	0.1257	0.0963	0.1976	0.5527
		Stratified Split	0.4139	0.3550	0.6359	0.4556	0.0932	0.0704	0.2032	0.5685

Table 35. Classification results using SBERT and RobBERT with default weights on features extracted from parents' segments.

Model	Classifier	Accuracy	Precision	Recall	F1	AUC
SBERT	Random Split	0.5055	0.3790	0.1665	0.2314	0.4184
	k-fold Split	0.5806	0.5175	0.2358	0.3240	0.5054
	Stratified Split	0.5533	0.3116	0.1857	0.2327	0.4402
RobBERT	Random Split	0.5347	0.4634	0.1642	0.2425	0.4688
	k-fold Split	0.5964	0.4340	0.2953	0.3515	0.5502
	Stratified Split	0.5336	0.1778	0.0857	0.1157	0.4538

Table 36. Classification results using SBERT and RobBERT with class weights on features extracted from parents' segments.

Model	Classifier	Accuracy	Precision	Recall	F1	AUC
SBERT	Random Split	0.4760	0.3006	0.1543	0.2039	0.4176
	k-fold Split	0.5652	0.4438	0.2211	0.2952	0.5284
	Stratified Split	0.4914	0.2312	0.1714	0.1969	0.3681
RobBERT	Random Split	0.5117	0.3570	0.1107	0.1690	0.4146
	k-fold Split	0.5856	0.4698	0.4229	0.4451	0.5680
	Stratified Split	0.5540	0.3068	0.1714	0.2199	0.4355

Table 37. Classification results using Wav2Vec2.0 and CLAP features from parents' segments with default weights.

Model	Classifier	Accuracy	Precision	Recall	F1	AUC
Wav2Vec2.0	Random Split	0.5315	0.4293	0.1453	0.2171	0.5749
	k-fold Split	0.6232	0.4903	0.3463	0.4059	0.6034
	Stratified Split	0.5816	0.3888	0.3143	0.3476	0.5653
CLAP	Random Split	0.5348	0.3100	0.1449	0.1975	0.5705
	k-fold Split	0.6027	0.4933	0.2373	0.3204	0.5889
	Stratified Split	0.6020	0.4207	0.2571	0.3192	0.5505

Table 38. Classification results per scenario using Wav2Vec2.0 and CLAP features from parent's segments with class weights.

Model	Classifier	Accuracy	Precision	Recall	F1	AUC
Wav2Vec2.0	Random Split	0.4886	0.3153	0.1969	0.2424	0.5481
	k-fold Split	0.6023	0.4622	0.4232	0.4418	0.5821
	Stratified Split	0.5912	0.4411	0.3857	0.4115	0.5832
CLAP	Random Split	0.5378	0.3821	0.2363	0.2920	0.5542
	k-fold Split	0.6282	0.4951	0.3878	0.4349	0.5724
	Stratified Split	0.6227	0.4869	0.3571	0.4120	0.5855

Bibliography

Achenbach, T. (1991). *Manual for the Child Behavior Checklist/4-18 and 1991 Profile*. Department of Psychiatry, University of Vermont.

Agrawal, A. (2024). Illuminate: A novel approach for depression detection with explainable analysis and proactive therapy using prompt engineering.

Al Hanai, T., Ghassemi, M., and Glass, J. (2018). Detecting depression with audio/text sequence modeling of interviews. In *Interspeech 2018*, ISCA. ISCA.

Allgaier, A.-K., Krick, K., Opitz, A., Saravo, B., Romanos, M., and Schulte-Körne, G. (2014). Improving early detection of childhood depression in mental health care: The childrens depression screener (child-s). *Psychiatry Research*, 217(3):248–252.

Aloshban, N., Esposito, A., and Vinciarelli, A. (2022). What you say or how you say it? depression detection through joint modeling of linguistic and acoustic aspects of speech. *Cognit. Comput.*, 14(5):1585–1598.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders (DSM-5 (R))*. American Psychiatric Association Publishing, Arlington, TX, 5 edition.

Ardila, R., Branson, M., Davis, K., et al. (2020). Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Baevski, A., Zhou, H., Mohamed, A., et al. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations.

Bain, M., Huh, J., Han, T., and Zisserman, A. (2023). WhisperX: Time-accurate speech transcription of long-form audio.

Bernaras, E., Jaureguizar, J., and Garaigordobil, M. (2019). Child and adolescent depression: A review of theories, evaluation instruments, prevention programs, and treatments. *Front. Psychol.*, 10:543.

Bilalpur, M., Hinduja, S., Cariola, L., Sheeber, L., Allen, N., Morency, L.-P., and Cohn, J. F. (2023). SHAP-based prediction of mother's history of depression to understand the influence on child behavior. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION*, New York, NY, USA. ACM.

Birleson, P. (1981). The validity of depressive disorder in childhood and the development of a self-rating scale: a research report. *J. Child Psychol. Psychiatry*, 22(1):73–88.

Bommasani, R., Hudson, D. A., Adeli, E., et al. (2022). On the opportunities and risks of foundation models.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., and Gill, M.-P. (2020). Pyannote.Audio: Neural building blocks for speaker diarization. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Caligiuri, M. P. and Ellwanger, J. (2000). Motor and cognitive aspects of motor retardation in depression. *J. Affect. Disord.*, 57(1-3):83–93.

Chen, Q., Du, W., Gao, Y., Ma, C., Ban, C., and Meng, F. (2017). Analysis of family functioning and parent-child relationship between adolescents with depression and their parents. *Shanghai Arch. Psychiatry*, 29(6):365–372.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Chiong, R., Budhi, G. S., Dhakal, S., and Chiong, F. (2021a). A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. *Comput. Biol. Med.*, 135(104499):104499.

Chiong, R., Budhi, G. S., Dhakal, S., and Chiong, F. (2021b). A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. *Comput. Biol. Med.*, 135(104499):104499.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling.

Çorbacıoğlu, Ş. K. and Aksel, G. (2023). Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. *Turk. J. Emerg. Med.*, 23(4):195–198.

Darby, J. K., Simmons, N., and Berger, P. A. (1984). Speech and voice parameters of depression: a pilot study. *J. Commun. Disord.*, 17(2):75–85.

Delobelle, P. and Remy, F. (2023). Robbert-2023: Keeping dutch language models up-to-date at a lower cost thanks to model conversion.

Deshmukh, S., Elizalde, B., Singh, R., and Wang, H. (2023). Pengi: An audio language model for audio tasks.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding.

Dibeklioglu, H., Hammal, Z., and Cohn, J. F. (2018). Dynamic multimodal measurement of depression severity using deep autoencoding. *IEEE J. Biomed. Health Inform.*, 22(2):525–536.

Drossos, K., Lipping, S., and Virtanen, T. (2020). Clotho: an audio captioning dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Elhai, J. D., Contractor, A. A., Tamburrino, M., Fine, T. H., Prescott, M. R., Shirley, E., Chan, P. K., Slembarski, R., Liberzon, I., Galea, S., and Calabrese, J. R. (2012a). The factor structure of major depression symptoms: a test of four competing models using the patient health questionnaire-9. *Psychiatry Res.*, 199(3):169–173.

Elhai, J. D., Contractor, A. A., Tamburrino, M., Fine, T. H., Prescott, M. R., Shirley, E., Chan, P. K., Slembarski, R., Liberzon, I., Galea, S., and Calabrese, J. R. (2012b). The factor structure of major depression symptoms: a test of four competing models using the patient health questionnaire-9. *Psychiatry Res.*, 199(3):169–173.

Elizalde, B., Deshmukh, S., Ismail, M. A., and Wang, H. (2022). CLAP: Learning audio concepts from natural language supervision.

et al., T. (2023). Llama 2: Open foundation and fine-tuned chat models.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile. In *Proceedings of the 18th ACM international conference on Multimedia*, New York, NY, USA. ACM.

Faruqui, M., Pavlick, E., Tenney, I., et al. (2018). WikiAtomicEdits: A Multilingual Corpus of Wikipedia Edits for Modeling Language and Discourse. In *Proc. of EMNLP*.

France, D., Shiavi, R., Silverman, S., et al. (2000). Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering*, 47(7):829–837.

Gemini Team and et al., A. (2023). Gemini: A family of highly capable multimodal models.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., et al. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Goldman, L. S., Nielsen, N. H., and Champion, H. C. (1999). Awareness, diagnosis, and treatment of depression. *J. Gen. Intern. Med.*, 14(9):569–580.

Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D., Rizzo, S., and Morency, L.-P. (2014). The distress analysis interview corpus of human and computer interviews. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).

Grosman, J. (2021). Fine-tuned XLSR-53 large model for speech recognition in Dutch. `https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-dutch`.

Havigerová, J. M., Haviger, J., Kučera, D., and Hoffmannová, P. (2019). Text-based detection of the risk of depression. *Front. Psychol.*, 10:513.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units.

Huang, J., Lu, C., Ping, G., Sun, L., and Ye, X. (2020). TCN-ATT: A non-recurrent model for sequence-based malware detection. In *Advances in Knowledge Discovery and Data Mining*, Lecture notes in computer science, pages 178–190. Springer International Publishing, Cham.

Huang, R., Li, M., Yang, D., Shi, J., Chang, X., Ye, Z., Wu, Y., Hong, Z., Huang, J., Liu, J., Ren, Y., Zhao, Z., and Watanabe, S. (2023). AudioGPT: Understanding and generating speech, music, sound, and talking head.

Huang, X., Wang, F., Gao, Y., Liao, Y., Zhang, W., Zhang, L., and Xu, Z. (2024). Depression recognition using voice-based pre-training model. *Sci. Rep.*, 14(1):12734.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging.

Jeni, L. A., Cohn, J. F., and De La Torre, F. (2013). Facing imbalanced data–recommendations for the use of performance metrics. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2016). Bag of tricks for efficient text classification.

Kahn, J., Rivière, M., Zheng, W., et al. (2019). Libri-light: A benchmark for asr with limited or no supervision.

Kessler, R. C. and Bromet, E. J. (2013). The epidemiology of depression across cultures. *Annual Review of Public Health*, 34(1):119–138. PMID: 23514317.

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

Kim, C. D., Kim, B., Lee, H., et al. (2019). AudioCaps: Generating Captions for Audios in The Wild. In *NAACL-HLT*.

Kim, J., Jung, J., Lee, J., and Woo, S. H. (2024). EnCLAP: Combining neural audio codec and audio-text joint embedding for automated audio captioning.

Kohler, K. J. (2017). Speech communication in human interaction. In *Communicative Functions and Linguistic Forms in Speech Interaction*, pages 18–70. Cambridge University Press, Cambridge.

Korczak, D. J., Westwell-Roper, C., and Sassi, R. (2023). Diagnosis and management of depression in adolescents. *CMAJ*, 195(21):E739–E746.

Lang, M. and Tisher, M. (1974). Children's depression scale. *Firenze: Organizzioni Speciali*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Lima, N. N. R., do Nascimento, V. B., de Carvalho, S. M. F., de Abreu, L. C., Neto, M. L. R., Brasil, A. Q., Junior, F. T. C., de Oliveira, G. F., and Reis, A. O. A. (2013). Childhood depression: a systematic review. *Neuropsychiatr. Dis. Treat.*, 9:1417–1425.

Lin, D., Nazreen, T., Rutowski, T., Lu, Y., Harati, A., Shriberg, E., Chlebek, P., and Aratow, M. (2022). Feasibility of a machine learning-based smartphone application in detecting depression and anxiety in a generally senior population. *Front. Psychol.*, 13:811517.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach.

Liu, Y.-L. (2003). Parent-child interaction and children's depression: the relationships between parent-child interaction and children's depressive symptoms in taiwan. *J. Adolesc.*, 26(4):447–457.

McGinnis, E. W., Anderau, S. P., Hruschak, J., Gurchiek, R. D., Lopez-Duran, N. L., Fitzgerald, K., Rosenblum, K. L., Muzik, M., and McGinnis, R. S. (2019). Giving voice to vulnerable children: Machine learning analysis of speech detects anxiety and depression in early childhood. *IEEE J. Biomed. Health Inform.*, 23(6):2294–2301.

Mental HealthFoundation (2024). `https://www.mentalhealth.org.uk/explore-mental-health/statistics`.

Milintsevich, K., Sirts, K., and Dias, G. (2023). Towards automatic text-based estimation of depression through symptom prediction. *Brain Inform.*, 10(1):4.

Mitchell, A. J., Vaze, A., and Rao, S. (2009). Clinical diagnosis of depression in primary care: a meta-analysis. *Lancet*, 374(9690):609–619.

Moreau, D. L. (1990). Major depression in childhood and adolescence. *Psychiatr. Clin. North Am.*, 13(2):355–368.

Muschelli, J. (2020). ROC and AUC with a binary predictor: A potentially misleading metric. *J. Classif.*, 37(3):696–708.

Netherlands Forensic Institute (2024). robbert-2022-dutch-sentence-transformers (revision cdf42f6).

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. (2016). MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

Onland-Moret, N. C., Buizer-Voskamp, J. E., Albers, M. E. W. A., Brouwer, R. M., Buimer, E. E. L., Hessels, R. S., de Heus, R., Huijding, J., Junge, C. M. M., Mandl, R. C. W., Pas, P., Vink, M., van der Wal, J. J. M., Hulshoff Pol, H. E., and Kemner, C. (2020). The YOUth study: Rationale, design, and study procedures. *Dev. Cogn. Neurosci.*, 46(100868):100868.

OpenAI and et al., A. (2023). GPT-4 technical report.

Ortiz Su'arez, P. J., Romary, L., and Sagot, B. (2020). A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Ortiz Su'arez, P. J., Sagot, B., and Romary, L. (2019). Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim. Leibniz-Institut f"ur Deutsche Sprache.

Orvaschel, H., Weissman, M. M., and Kidd, K. K. (1980). Children and depression: The children of depressed parents; the childhood of depressed patients; depression in children. *Journal of Affective Disorders*, 2(1):1–16.

Pan, W., Flint, J., Shenhav, L., Liu, T., Liu, M., Hu, B., and Zhu, T. (2019). Re-examining the robustness of voice features in predicting depression: Compared with baseline of confounders. *PLoS One*, 14(6):e0218172.

Panayotov, V., Chen, G., Povey, D., et al. (2015). Librispeech: an asr corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pages 5206–5210. IEEE.

Park, K. and Mulc, T. (2019). Css10: A collection of single speaker speech datasets for 10 languages. *Interspeech*.

Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., and Narayanan, S. (2021). A review of speaker diarization: Recent advances with deep learning.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Pepino, L., Riera, P., and Ferrer, L. (2021). Emotion recognition from speech using wav2vec 2.0 embeddings.

Pessanha, F., Kaya, H., Akdag Salah, A. A., and Salah, A. A. (2022). Towards using breathing features for multimodal estimation of depression severity. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, New York, NY, USA. ACM.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision.

Radford, A., Kim, J. W., Xu, T., et al. (2022). Robust speech recognition via large-scale weak supervision.

Radovic, M., Ghalwash, M., Filipovic, N., and Obradovic, Z. (2017). Minimum redundancy maximum relevance feature selection approach for temporal gene expression data. *BMC Bioinformatics*, 18(1):9.

Reimers, N. and Gurevych, I. (2019a). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Reimers, N. and Gurevych, I. (2019b). Sentence-bert: Sentence embeddings using siamese bert-networks.

Rejaibi, E., Komaty, A., Meriaudeau, F., et al. (2022). Mfcc-based recurrent neural network for automatic clinical depression recognition and assessment from speech. *Biomedical Signal Processing and Control*, 71:103107.

Remy, F., Delobelle, P., Berendt, B., et al. (2023). Tik-to-tok: Translating language models one token at a time: An embedding initialization strategy for efficient language adaptation.

Robert, J., Webbie, M., et al. (2018). Pydub.

Rodrigues Makiuchi, M., Warnita, T., Uto, K., and Shinoda, K. (2019). Multimodal fusion of BERT-CNN and gated CNN representations for depression detection. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, New York, NY, USA. ACM.

Rolim Neto, M. L., Silva, T. d. N., Assunção Filho, J. K. M., Carvalho, R. d. S., Teixeira, S. A., Lima, N. N. R., Pedroso, D., Cartaxo, J. d. S., Demarzo, M. M. P., Duarte Júnior, J. A., and Reis, A. O. A. (2011). Childhood depression and psychocognitive development: description of causality relationships. `http://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S0104-12822011000300016&lng=pt&tlng=en`. Accessed: 2024-3-11.

Rude, S., Gortner, E.-M., and Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cogn. Emot.*, 18(8):1121–1133.

Samuel, R., Bowman, G., Angeli, C., and Potts, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. EMNLP.

Sardari, S., Nakisa, B., Rastgoo, M. N., and Eklund, P. (2022). Audio based depression detection using convolutional autoencoder. *Expert Syst. Appl.*, 189(116076):116076.

Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., and Pantic, M. (2011). AVEC 2011–the first international audio/visual emotion challenge. In *Affective Computing and Intelligent Interaction*, Lecture notes in computer science, pages 415–424. Springer Berlin Heidelberg, Berlin, Heidelberg.

Schuller, B. W., Batliner, A., Bergler, C., Messner, E.-M., Hamilton, A., Amiriparian, S., Baird, A., Rizos, G., Schmitt, M., Stappen, L., Baumeister, H., MacIntyre, A. D., and Hantke, S. (2020). The INTERSPEECH 2020 computational paralinguistics challenge: Elderly emotion, breathing & masks. In *Interspeech 2020*, ISCA. ISCA.

Seneviratne, N., Williamson, J. R., Lammert, A. C., et al. (2020a). Extended study on the use of vocal tract variables to quantify neuromotor coordination in depression. In *Interspeech*, pages 4551–4555.

Seneviratne, N., Williamson, J. R., Lammert, A. C., et al. (2020b). Extended study on the use of vocal tract variables to quantify neuromotor coordination in depression. In *Interspeech*, pages 4551–4555.

Shen, Y., Yang, H., and Lin, L. (2022). Automatic depression detection: an emotional audio-textual corpus and a gru/bilstm-based model. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6247–6251.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition.

Singh, P., Singh, G., Singh, A., and Singh, J. (2024). Intelligent mental depression recognition model with ensemble learning through social media tweet resources. *Cybern. Syst.*, 55(2):471–510.

Smith, M., Dietrich, B. J., Bai, E.-W., and Bockholt, H. J. (2020). Vocal pattern detection of depression among older adults. *Int. J. Ment. Health Nurs.*, 29(3):440–449.

Soares, I., Belsky, J., Oliveira, P., Silva, J., Marques, S., Baptista, J., and Martins, C. (2014). Does early family risk and current quality of care predict indiscriminate social behavior in institutionalized portuguese children? *Attach. Hum. Dev.*, 16(2):137–148.

Sobin, C. and Sackeim, H. A. (1997). Psychomotor symptoms of depression. *Am. J. Psychiatry*, 154(1):4–17.

Stolcke, A. (2019). Improving diarization robustness using diversification, randomization and the DOVER algorithm.

Susilo, S. (2020). The role of families in cultivating children's personality values: An analysis of social psychology education.

Tejaswini, V., Babu, K. S., and Sahoo, B. (2022). Depression detection from social media text analysis using natural language processing techniques and hybrid deep learning model. *ACM Trans. Asian Low-resour. Lang. Inf. Process.*

Thapar, A., Eyre, O., Patel, V., and Brent, D. (2022). Depression in young people. *Lancet*, 400(10352):617–631.

Trifu, R. N., Iuliu Hatieganu University of Medicine and Pharmacy, Cluj-Napoca, Romania, raluca.trifu@yahoo.com, Nemeș, B., Bodea-Hațegan, C., Cozman, D., Iuliu Hatieganu University of Medicine and Pharmacy, Cluj-Napoca, Romania, nemes.bogdan@umfcluj.ro, Babes-Bolyai University, Cluj-Napoca, Romania, carolina.bodea@ubbluj.ro, and Iuliu Hatieganu University of Medicine and Pharmacy, Cluj-Napoca, Romania, dcosman@umfcluj.ro (2017). Linguistic indicators of language in major depressive disorder (MDD). an evidence based research. *J. Evid.-Based Psychother.*, 17(1):105–128.

Valstar, M., Schuller, B., Smith, K., Almaev, T., Eyben, F., Krajewski, J., Cowie, R., and Pantic, M. (2014). AVEC 2014. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*, New York, NY, USA. ACM.

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., and Pantic, M. (2013). AVEC 2013. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, New York, NY, USA. ACM.

Vincent, P., Larochelle, H., Lajoie, I., et al. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.*, 11:3371–3408.

Vollebregt, R. (2023). A multimodal approach to working alliance detection in therapist-patient psychotherapy using deep learning models. Available at `https://studenttheses.uu.nl/handle/20.500.12932/45323`.

Wang, P. S., Demler, O., Olfson, M., Pincus, H. A., Wells, K. B., and Kessler, R. C. (2006). Changing profiles of service sectors used for mental health care in the united states. *Am. J. Psychiatry*, 163(7):1187–1198.

Williams, A., Nangia, N., and Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Wu, Y., Chen, K., Zhang, T., et al. (2024). Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation.

Xu, H., Liu, W., Liu, J., Li, M., Feng, Y., Peng, Y., Shi, Y., Sun, X., and Wang, M. (2022). Hybrid multimodal fusion for humor detection. In *Proceedings of the 3rd*

International on Multimodal Sentiment Analysis Workshop and Challenge, New York, NY, USA. ACM.

Xu, Y., Chen, H., Yu, J., Huang, Q., Wu, Z., Zhang, S., Li, G., Luo, Y., and Gu, R. (2023). SECap: Speech emotion captioning with large language model.

Yang, H., Kim, H., Lee, J. H., and Shin, D. (2022). Implementation of an AI chatbot as an english conversation partner in EFL speaking classes. *ReCALL*, 34(3):327–343.

Ye, J., Yu, Y., Wang, Q., Li, W., Liang, H., Zheng, Y., and Fu, G. (2021). Multi-modal depression detection based on emotional audio and evaluation text. *J. Affect. Disord.*, 295:904–913.

Zhang, Z., Zhang, S., Ni, D., Wei, Z., Yang, K., Jin, S., Huang, G., Liang, Z., Zhang, L., Li, L., Ding, H., Zhang, Z., and Wang, J. (2024). Multimodal sensing for depression risk detection: Integrating audio, video, and text data. *Sensors*, 24(12):3714.

Zhao, T., Kong, M., Liang, T., et al. (2023). Clap: Contrastive language-audio pre-training model for multi-modal sentiment analysis. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, ICMR '23, page 622–626, New York, NY, USA. Association for Computing Machinery.

Zhao, Z., Li, Q., Cummins, N., Liu, B., Wang, H., Tao, J., and Schuller, B. W. (2020). Hybrid network feature extraction for depression assessment from speech. In *Interspeech 2020*, ISCA. ISCA.

Zhu, Y., Kiros, R., Zemel, R., et al. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.