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

AUTOMATED ASSESSMENT OF SYMPTOMS OF ANXIETY IN CHILDREN FROM PARENT-CHILD INTERACTION VIDEOS

Master Thesis for MSc Artificial Intelligence
INFOMAI1

A THESIS BY
Aristophanes Albertus Alvin
2928175

Project supervisor Dr. Itir Önal Ertuğrul
Daily supervisor Mang Ning, MSc
Second examiner Prof. Dr. Albert Ali Salah

Abstract

Burgeoning economic crises, growing political instabilities, and the recent pandemic have caused mental health deterioration in many parts of the world. The most prevalent mental health affliction is anxiety disorder and it has been affecting a growing number of adults and children alike. In this thesis, we investigated the feasibility and approaches of applying machine learning for the detection of anxiety symptoms in 9-year-old children. These symptoms might be expressed more saliently through certain modes of communications and in specific interactional contexts. Hence, we experimented with models that were trained on unimodal and multimodal features extracted from video recordings of conflictual and cooperative interactions between nine-year-olds and their parents in a laboratory setting. Results suggest that anxiety symptoms manifest most noticeably during tense, conflictual interactions and are conveyed through the hand movements, facial expressions—particularly the mouth area—and word choice. Moreover, training with multimodal features demonstrated better performance compared to unimodal approaches. Although the resulting performance of the models was moderate, this study establishes the feasibility of detecting symptoms of anxiety using machine learning applied to multimodal dataset.

Table of Contents

1 Introduction 4

1.1 Relevance . . . 4

1.2 Machine learning to detect anxiety symptoms . . . 5

1.3 Research questions . . . 6

1.4 Contribution . . . 7

1.5 Ethical and privacy concerns . . . 8

2 Related work 10

2.1 Anxiety emotion . . . 10

2.1.1 Anxiety in children . . . 11

2.2 Anxiety symptoms detection . . . 12

2.2.1 Visual modality . . . 13

2.2.2 Acoustic modality . . . 15

2.2.3 Linguistic modality . . . 18

2.3 Clinical application . . . 18

2.4 Previous works on the YOUth dataset . . . 20

3 Methodology 21

3.1 Dataset . . . 21

3.2 Feature extraction . . . 23

3.2.1 Acoustic modality . . . 24

3.2.2 Linguistic modality . . . 26

3.2.3 Visual modality . . . 28

3.3 Pipeline . . . 35

3.3.1 Data preparation . . . 36

3.3.2 Model architectures . . . 38

3.3.3 Model training . . . 42

3.3.4 Performance metrics . . . 43

4 Results 46

4.1 Comparative analysis between modalities . . . 46

4.2 Comparative analysis between scenarios . . . 49

4.3 Important predictors . . . 51

4.3.1 Acoustic features . . . 51

4.3.2 Linguistic features . . . 53

4.3.3 Facial expressions features 56
4.3.4 Affective body pose features 58
4.3.5 Combined modality 60

5 Discussion 63
5.1 Better context through multimodality 63
5.2 Interaction scenarios and anxiety expressions 63
5.3 Advantage of the GBC approach 64
5.3.1 Curse of dimensionality, kernel methods, and tree-based approaches 65
5.3.2 Hyperparameter optimization 65
5.3.3 Early vs. late fusion 65
5.4 Suitability of the MCC metric 66
5.5 Answers to the research questions 66

6 Limitations &future work 69

7 Conclusion &final remarks 72

8 Acknowledgements 73

9 Appendix 74
9.1 Model performance 74
9.2 Confusion matrices 78
9.3 Contingency tables 79
9.4 Hyperparameters 80

1. Introduction

1.1 Relevance

Anxiety is a prevalent emotion experienced by people of all ages and is characterized by a sense of restlessness and worry in anticipation of a forthcoming event. There are at least two factors affecting the intensity of the emotion: the significance of the event and the individual's preparedness to face its worst possible outcome (Hinds et al., 2010). Just as any other emotions, anxiety evolves as an adaptation that aids the individual to survive and thrive in the environment. However, when dysregulated, it transforms into a debilitating force that interferes with the individual's well-being and becomes a disorder (Marks and Nesse, 1994).

In 2019, it was estimated that 301 million people in the world suffer from anxiety disorder. This estimate increased to 322 million or 4% of global population by 2023, making anxiety disorder the most prevalent mental issue.[1] Unfortunately, despite being highly treatable, only one in four sufferer received treatment due to factors such as lack of awareness, inadequate investment in mental health services, low number of trained health care providers, and social stigma. Anxiety disorder is not only detrimental to health and–in severe cases–potentially life-threatening, but also inflicts sizeable economic loss. One estimate suggests that 2.08% of global healthcare costs and 0.22% of global GDP is spent on treating the disease, whereas the indirect costs, such as loss of productivity and stunted education, corresponded to 0.23% of GDP (Konnopka and König, 2020).

Mental health issues in general, and anxiety disorder in particular, also affect children and adolescent. While encountering fear and learning to overcome it are normal parts of childhood, some children may fail to outgrow these emotions such that their education, socialization, and family life are disrupted. Some symptoms of anxiety disorder in children includes being terrified when away from parents (separation anxiety), extremely fearful of certain objects or animals (phobias), refusing to go to school and crowded places (social anxiety), having episodes of sudden and intense fear accompanied with dizziness or shaking (panic disorder), or unable to speak in certain situations (selective mutism).[2]

Children mental health issue is becoming more widespread and has been on the rise even prior to the COVID-19 pandemic. It was estimated that between 2016 and 2019, the number

[1] https://www.who.int/news-room/fact-sheets/detail/anxiety-disorders

[2] https://www.cdc.gov/childrensmentalhealth/depression.html

of children and adolescents in the United States who suffered from anxiety and depression rose by 27% and 24% respectively. By 2020, 5.6 million kids (9.2%) had been affected by anxiety problems and 2.4 million (4%) had been affected by depression (Lebrun-Harris et al., 2022). Although the statistics may vary between countries, prioritizing children's mental health is becoming an urgent global concern.

Current methods to assess mental disorders depend almost entirely on clinical interviews and self-reports. These approaches are not only prone to subjective biases, but also hinge on the sufferer's awareness of psychological disorders, articulation capacity, and sincerity in reporting. Already this presents a barrier to segments of sufferers such as youth, less educated population, and those with diminished mental and cognitive capacity. Moreover, while these measures have been proven useful, they lack objective and efficient ways of incorporating nonverbal behavioral observations that can be strong indicators of the presence of mental disorders. This presents a missed opportunity that will be explored further in this study.

Recently, automated assessment of symptoms of mental disorders from multimodal behavioral cues has become an important topic in the field of affective computing. Several studies have focused on automated assessment of symptoms of depression, obsessive-compulsive disorder and anxiety disorders from verbal and nonverbal behavior of adults. Most studies have focused on assessment during clinical interviews, interaction with a virtual agent, and reading aloud. Broadly speaking, there are two types of detection methods: intrusive and non-intrusive. Intrusive detections rely on biosignals such as heartbeat and blood volume. As such, the subject has to be intruded with apparatus that is attached on or in the body. In contrast, non-intrusive methods rely on observations made from a distance with little to no physical contact with the subject. It utilizes features of outward expressions conveyed through multiple modes of communications such as visual (facial and head movements), auditory (vocal characteristics), and linguistic (words statistics) modalities. In a unimodal approach, automated assessment is produced using features from a single modality, such as auditory features or linguistic features. In a multimodal approach, features from multiple modalities are combined to make the predictions.

1.2 Machine learning to detect anxiety symptoms

In this thesis, we developed a machine learning approach to automatically assess symptoms of anxiety in 9-year-old children, as defined by items of Child Behavior Checklist (CBCL; Verhulst et al. (1996)). Our models were trained using verbal and nonverbal behaviors that were captured in video recordings of the children's interactions with their parents. These interaction sessions were part of the YOUth Cohort Study (Onland-Moret et al., 2020), a large scale longitudinal cohort study following nearly 4,000 Dutch children in their development from pregnancy until early adulthood. This study utilized the interaction videos that recorded their behaviors in two settings: planning a vacation and arguing.

We employed three types of modeling approaches (Support Vector Machine, Gradient Boosting Classifier, and Light Gradient Boosting Machine) and applied grid-search cross-validation for hyperparameter tuning. To better approximate the capability of the models to generalize to unseen data, we ran stratified k-fold cross-validation and averaged the model's performance across different folds. We analyzed the performances of models that were trained on unimodal features and compared them with models trained on multimodal features. We also analyzed the performance difference between models trained on features from conflictual interaction videos and from cooperative videos. Furthermore, we leveraged Shapley Additive Explanations (SHAP) to analyze the relative importance and impact of features in a model's predictions. By doing so, we were able to suggest the outward manifestations of anxiety symptoms based on the features most impactful to our models.

1.3 Research questions

The aim of this study is to investigate the feasibility and effectiveness of machine-learning-based methodologies of automated anxiety symptoms detection on children. More specifically, we aim to answer the following questions:

Main question: *To what extent can anxiety symptoms be detected in 9 year-old children using unimodal and multimodal approaches?*

The above question is very broad and contains many supporting sub-aspects that are formulated below.

Sub-question 1: *To what extent can unimodal methods detect symptoms of anxiety compared to multimodal approaches?*

The outward behavior of our participants was observed through multiple modalities. In this sub-question, we compared the predictive power of features extracted from individual and combined modalities. By answering this question, we were able to suggest whether anxiety symptoms are more salient in some modalities than others, and whether a holistic observation through combined modalities is more beneficial.

Sub-question 2: *How do videos recorded during a conflicting and a cooperative task compare to detect symptoms of anxiety?*

The situation in which a person finds him/herself may have a strong influence on their expressiveness. Likewise, outward symptoms of anxiety may surface more quickly/prominently in certain situations more than others. YOUth Cohort Study presented two contrasting themes of interaction between the child and parent. One was peaceful, almost frictionless, and likely delightful for the children (i.e., vacation planning), while the other was tense and perhaps intimidating (i.e., arguing). By observing how the children behaved in these opposite atmospheres, we investigated whether different types of interaction scenarios elicit different levels of expressiveness of anxiety. To do this, we performed experiments using data from individual scenarios, and both, and compared their performances to detect anxiety symptoms.

Sub-question 3: *What are the important verbal and non-verbal indicators of anxiety symptoms?*

Following sub-question 1, we dived deeper into the models' prediction mechanism to discover the features that most strongly influenced the outcome. To do this, we applied SHAP and analyzed the pattern of influence each feature had on the models' predictions of the test sets. Surely, non-linearity and feature interactions affected how a particular behavior would be weighed differently from one participant to another. However, by observing the general patterns, we were able to identify specific actions of the participants that were presumed as manifestations of anxiety symptoms.

Note that this study is not aimed at discovering the best-suited model architecture for the problem domain. While hyperparameter tuning was involved, the independent variables were mainly the feature sets and the type of interactions displayed in the videos. We implemented a commonly used model architecture that have been proven effective in previous studies, as explained in Section 2.2

1.4 Contribution

Current anxiety diagnostic methods rely mostly on the reported level of subjective distress and suffering to distinguish normal anxiety from pathological anxiety (Knepley et al., 2019; Read et al., 2015). This approach can introduce bias into diagnoses and makes early detection less accessible to those unfamiliar with psychological disorders or less capable to articulate their emotions, such as young children, less informed populations, and individuals with diminished mental or cognitive capacities. This thesis aims to demonstrate the feasibility of applying technological measurement tools in psychiatry. In turn, it may encourage the adoption of technology, not as a replacement for human experts but as a preliminary screening aid or as a supplementary second opinion, serving the overarching purpose of making mental health care more accessible.

In the field of Psychology, this study contributes by utilizing physical expressions as anxiety indicators, thus furthering the understanding of the correlation between mental disorders and their outward expressions. This thesis also contributes to the field of Artificial Intelligence by investigating how different modalities and their combinations affect the performance of various models, as well as exploring ways to optimize them. We also showcased multiple pre-trained models by leveraging them to process raw videos and extract features, which were then used by other models to learn the patterns of anxiety symptoms.

Several studies have attempted to detect emotional distress using physical expressions captured through a single modality. For instance, Han et al. (2020) and Baird et al. (2020) used audio features, while Giannakakis et al. (2017), Holmes et al. (2006), Pearlstein et al. (2019), Metaxas et al. (2004), and Pediaditis et al. (2015) focused on visual features. There are also studies that combined multiple modalities, such as Lin et al. (2023) which integrated audio and visual features (body poses), and Hinduja et al. (2024) which utilized audio, visual (facial expressions), and linguistic features. However, to my knowledge, this is the first study that extracts and combines features from facial expressions, body language, verbal content, and voice characteristics, while also considering how the context of the interaction atmosphere (i.e., conflicting vs. cooperative) affects the manifestation of anxiety symptoms.

1.5 Ethical and privacy concerns

Since we worked with human data containing unaltered voices and images of children and parents, the security and protection of the participants' privacy is of utmost priority. At the time of this writing, the European Union (EU) has released the final draft[3] of a law that will govern the deployment and usage of artificial intelligence technology (AI) across EU, aptly named The EU AI Act. The purpose of the act is to ensure that AI systems are safe, transparent, traceable, non-discriminatory and environmentally friendly.[4] Title II Article 5 point 1-dc of the act in particular is relevant for this work as it regulates the us of emotion recognition systems[5]. The precise wording of the law states:

> *The following artificial intelligence practices shall be prohibited: ... (dc) the placing on the market, putting into service for this specific purpose, or use*

[3] https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf

[4] https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

[5] https://artificialintelligenceact.eu/high-level-summary/

> *of AI systems to infer emotions of a natural person in the areas of workplace and education institutions except in cases where the use of the AI system is intended to be put in place or into the market for medical or safety reasons;*[6]

This study protected the privacy of the participants in a number of ways. First, we did not recruit the participants and never had any contact with any of them. Their identities were unknowable to any of the authors and supervisors of this thesis. Second, all of the recordings were stored in a protected server that prevented any unauthorized party, ourselves included, from exporting the data. Our access privilege was strictly for data processing and modelling experimentation for the purpose of this thesis. We did not work with samples from the original dataset on our local machines; instead, we worked with dummy data obtained from various other sources. Third, we did not do this study in a joint manner with any external organizations or persons outside of the team of authors and supervisors, thus eliminating the risk of unwarranted access from third-parties. Fourth, this work was done for the purpose of studying the feasibility of automated anxiety symptoms detection, and not intended to be deployed in any institutions or workplaces, let alone for commercial purpose. Moreover, Utrecht University had ensured the ethical and privacy aspect of this study through its master's thesis screening.[7]

[6] https://artificialintelligenceact.eu/article/5/

[7] https://www.uu.nl/en/research/institute-of-information-and-computing-sciences/ethics-and-privacy/guidance-for-research-master-thesis-students

2. Related work

This chapter discusses the literature and studies relevant to this work. It explores the anxiety emotion from psychological perspective, describes a number of works that applies artificial intelligence to detect the symptoms of anxiety disorder, and shows several studies that had been conducted on the YOUth dataset.

2.1 Anxiety emotion

Over the course of its lifetime, an organism constantly navigates through an environment that is continuously undergoing gradual or drastic shifts. This dynamic exposes the organism to potential threats that may lurk in the unfamiliar and unknown parts of the environment. The emotion of anxiety serves to alert the organism to the level of risk and uncertainties in its surroundings, which in turn prepares it to respond appropriately in anticipation of the impending danger (Marks and Nesse, 1994; Bateson et al., 2011). Anxiety produced by the brain's security motivation system is an enduring motivational state that urges the organism to perform threat-reducing actions (Woody and Szechtman, 2005; Woody et al., 2005; Szechtman and Woody, 2004). Experiments have shown that concrete and tangible action is highly effective in reducing the activation to baseline, while post-contact reappraisal has negligible effect. In other words, the anxiety signal has little correlation to whether the threat is real or sham; what is relevant is the organism's readiness to face the threat if it were to occur (Hinds et al., 2010). The vast evolutionary timespan shaping the anxiety mechanism has caused humans to be more anxious to largely archaic threats, like snakes (~138 thousand deaths per year), than to modern hazards, like automobiles (~1.19 million deaths per year).[1] That said, humans have not failed to extrapolate the mechanism to detect threats to relationships, property, social status, reputation, capability, and anything else that supports our thriving and well-being in the modern society.

The methodologies used in the literature that are presented here have shown that anxiety-exhibiting behaviors can be induced in laboratory settings and be detected through small time windows. However, when the same behaviors is displayed over a longer period of time and in non-anxiety-inducing situations, it may be categorized as a disorder. Like other disorders, anxiety disorders are failure of regulation resulting in excessive or deficient responses (Marks and Nesse, 1994). Generalized anxiety disorder (GAD) is among the

[1] www.who.int/news-room/fact-sheets/detail/snakebite-envenoming and www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

most common disorders, affecting 6.8 million adults or 3.1% of the U.S. population.[2] It is characterized by excessive anxiety of at least six months' duration, occurring most days, hard to control, not focused on a specific situation or objects, and not triggered by recent stressful events. GAD is comorbid with other psychopathological disorders in majority of cases. Other disorders include: depressive disorder, bipolar disorder, mood disorder, separation anxiety, panic disorder, agoraphobia, social phobia, simple phobias, obsessive-compulsive disorder (OCD), attention deficit/hyperactivity disorder, and bipolar disorder (Masi et al., 2004).

2.1.1 Anxiety in children

Although there are developmental periods where increased anxiety is normative (e.g., separation anxiety at 9 months - 18 months, fear of storms in toddlerhood), the excessive and overwhelming presence of the emotion can morph into disorder that affects people even at a young age. Unfortunately, data on the prevalence of this disease is hard to come by due to various factors (e.g., under-reported, undetected, and variations between regions and cultures). Estimates of the prevalence rates of anxiety disorder among young children vary from 9% - 10% (Egger and Angold, 2006) and up to 22% (Paulus et al., 2015). Physical complaints are frequent and early signs of anxiety, both in children and in adolescents. Children with somatic symptoms are more fearful of novelty, more prone to separation anxiety disorder (SAD), and more likely to miss school or refuse to go to school (Bernstein et al., 1997). Although children might be less verbally coherent than adults, they might exhibit similar but less articulated symptoms such as alteration in tone and volume of voice, increased use of filler sounds, reduced word counts, increased negative words, Laukka et al. (2008) or broader change in body language such as crying, tantrums, freezing, clinging, repeated frightening dreams, blushing, muscle tension, or trembling (Rockhill et al., 2010). Preschool children with an anxiety disorder often have comorbid depression, attention deficit-hyperactivity disorder, oppositional defiant disorder (ODD), or conduct disorder (Rockhill et al., 2010). Moreover, once a child is diagnosed with an anxiety disorder, that child is at increased future risk for the same disorder and for additional anxiety and depressive disorders. In turn, it may jeopardize their education and relationship, leading to less social acceptability and support, in the long term may lead to criminality and societal disharmony.

Separation anxiety disorder (SAD) is one of the most common childhood anxiety disorders. SAD involves significant distress when the child is unexpectantly separated from home or a close attachment figure (Vaughan et al., 2017). Separation anxiety is developmentally normal for infants and toddlers but becomes a disorder when it continues past toddlerhood. Some risk factors from parents: parents with panic disorder or other anxiety disorders,

[2] adaa.org/understanding-anxiety/facts-statistics

maternal smoking, alcohol consumption, low birth weight. Risk factors from life circumstances: divorce, military leaves such as parental deployments during war, foster care, adoption, incarceration, parental death, and relocation due to occupation. Children whose parents stayed with them during the onset of sleep, such as co-sleeping, were at an increased risk of developing anxiety or depressive symptoms. Some consequences: social withdrawal, apathy, sadness, or difficulty concentrating on work or play, which in turn leads to academic social difficulties. SAD also precedes other disorders, such as panic disorder with agoraphobia, social phobias, obsessive–compulsive disorder, bipolar disorder, pain disorder, depressive disorders, and alcohol dependence. SAD may also strain the relationship with other caregivers, leading to resentment and low self-efficacy among caregivers and family members. Children with SAD often present with comorbid conditions such as depression and attention deficit hyperactivity disorder, and thus it is important to consider these comorbid disorders when screening for and diagnosing SAD.

2.2 Anxiety symptoms detection

Studies have shown that stress and anxiety have physical and behavioral manifestations. Marks and Nesse (1994) identify at least four ways in which anxiety prepares an organism to deal with threat. First, it urges flight/escape or preflight/avoidance that distances the individual from the threat. Sweating prevents the body from overheating and makes the skin slippery, preparing the organism for evasion. Vomiting, disgust, diarrhea, coughing, and sneezing are some of the ways the organism creates physical distance from the pathogen. Second, it empowers the organism to do aggressive defense by neutralizing the source of threat just as the immune system attacks the bacteria. Hyperventilation raises oxygen intake and carbon dioxide removal, increasing energy production and preparing the body to exert high amount of power. Third, it ceases the organism's movement (freeze) which may conceal it, prevent the predator's attack reflex, prevent detrimental moves (e.g., fear of height), and aid the location and assessment of the threat. Fourth, it inhibits the organisms impulses and induce submission/appeasement which may be beneficial when the threat comes from one's own group.

In the more severe cases, stress and anxiety can negatively affect our physical health. The symptoms of illnesses resulting from anxiety includes headache, hypertension, muscular pain especially in the neck and lower back, and gastrointestinal problems such as gastritis and irritable bowel or colic (Giannakakis et al., 2017). Anxiety is often comorbid with other mental disorders such as depression and can be fatal when it leads to social isolation and suicidal thoughts (Gorman, 1996).

Given the physical manifestations, it is feasible to objectively detect anxiety from the outward symptoms using external detectors, rather than relying solely on self-reports of the sufferers. Although studies aimed specifically at detecting anxiety are rare, there are substantial number of relevant works aimed at detecting other mental disorders like OCD and depression from multimodal behavioral indicators. The following sections will discuss various methods to automate detection of symptoms of mental disorders found in the literature, as well as the challenges in applying them in clinical practices.

2.2.1 Visual modality

Several works suggest that psychological disorders are expressed visually, mainly through head and facial movements. Several studies have found that emotionally distressed people tend to make greater overall head movements with more frequency and rapidity (Dinges et al., 2005; Liao et al., 2005; Hadar et al., 1983). People with anxiety disorder, for example, tend to show symptoms such as increased blink rate, eye aperture/pupil diameter, eyelid response, gaze distribution, and variation in pupil size (Giannakakis et al., 2017). Anxious people also tend to be more alerted by and make more saccades to images of angry or fearful expressions (Holmes et al., 2006; Mogg et al., 2007). A study by Harrigan and O'Connell (1996) reports that raising the eyebrows and drawing them together is the second most reliable facial indicator of anxiety. Other indicators mentioned in the same report are widening the eye by raising the upper eyelid, non-enjoyment smile (lacking eye movement compared to sincere smile), and decreased eye contact. Another study by Pearlstein et al. (2019) makes the same observation where people diagnosed with SAD have reduced smile reciprocity and express insincere smiles, i.e., one with minimal involvement of the eye region. Moreover, they tend to have decreased smile frequency and intensity, and they tend to smile less when listening than when speaking in a conversation.

Regarding the mouth region, it is found that asymmetric lip deformations and movements are related to stress/anxiety (Metaxas et al., 2004). Giannakakis et al. (2017) reports that the frequency of mouth opening is inversely proportional to stress level, as indexed by higher cognitive workload. Another study by Harrigan and O'Connell (1996) found that stretching the lips horizontally to form a rectangular mouth shape is the most predominant indicator of the fear emotion. Pediaditis et al. (2015) uses the position and velocity of the upper and lower lips to distinguish stressed or anxious state versus relaxed state. Interestingly, facial skin tone, eye blink rate, and mouth opening rate were not found to be among the most useful predictors.

The facial skin tone is also found to be correlated to stress/anxiety (Giannakakis et al., 2017), with blushing linked to anger and pallor linked to fright or embarrassment. The change in skin tone, as well as other symptoms such as sweating palms and shortness of breath, are the result of the change of blood volume transferred from the heart (Mauss et al., 2004).

Beyond the head and facial region, emotional distress like anxiety and depression also manifests itself throughout other parts of the body and can be read from body postures and gestures of the limbs. In a study by Lin et al. (2023), audio features are combined with generic body features (overall movements, hands, legs, and head movements) and fidgeting behavior (cross hand, single-hand, and leg/feet fidgeting) to predict the level of anxiety and depression. The authors find that the length of gestures in the head and legs is most predictive of mental distress, while hand fidgeting is observed across individuals irrespective of mental condition. Another pertinent study by Kasap and Tanhan (2019) explores the effect of open and closed body postures on anxiety levels. Their work draws from the well-established reciprocal relationship between mental states and body language, as elucidated by previous research (Hargrave, 2008; Pease and Pease, 2008).

Several studies have suggested that the predictive power of methods that use visual features are comparable to those that use biosignals such as blood volume pulse, galvanic skin response, pupil diameter, and skin temperature (Dinges et al., 2005; Gao et al., 2014; Setz et al., 2010; Zhai and Barreto, 2006; Barreto et al., 2007). However, it is worth noting that facial expressions are also affected by factors independent of the emotional and cognitive state of the subject. For example, the ambient conditions such as lighting, temperature, and humidity, or disorders such as depression, schizophrenia, and Parkinson's disease (Giannakakis et al., 2017). Moreover, people tend to blink more during speaking, reading or memorizing (Harrigan and O'Connell, 1996).

There have been a number of studies that employ computer vision to detect symptoms of psychological disorders. Adams et al. (2015), for example, uses head shakes and nods among other features to identify complex emotional situations. In the work by Dibeklioğlu et al. (2018), vocal features are combined with visual features (facial movements and head pose) extracted using a fully automatic, person-independent, generic approach called ZFace (Jeni et al., 2015). The library works by constructing 3D registration from 2D videos, tracking 49 facial landmarks (fiducial points) and 3 degrees of out-of-plane rigid head movements (pitch, yaw, and roll). For the predictor, the authors choose to use Stacked Denoising Autoencoders (SDAE), a deep network architecture based on stacking layers of denoising autoencoders which has been proven to be highly effective in discovering unknown non-linear mappings between features and outcomes, while coping with high dimensionality and redundancy. Separate SDAEs are used to encode per-frame representations for facial and head movements, each comprises 3 hidden layers. The number of units per each hidden layer and other hyperparameters are determined during training by minimizing the prediction error.

ZFace is also employed in a study by Hinduja et al. (2024) to extract facial Action Units (AUs), head pose, Eye Aspect Ratio, and Mouth Aspect Ratio, which are then combined with acoustic and linguistic features to predict OCD and comorbid depression severity. This work specifically uses the AFAR (Ertugrul et al., 2019) AU detector module of ZFace which tracks a set of facial muscles instrumental in the display of emotion according to the Facial Action Coding System (FACS, Ekman and Friesen (1978)). As the classifier, they train a number of Mixed-Effect Random Forests (MERF) on multimodal features. The performance is evaluated using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-square (R^2), Intraclass correlation (ICC), Normalized MAE, and feature contribution assessed through Shapley Additive Explanations (SHAP). With this method, the authors show results that have the potential to rival the interrater reliabilities of trained clinicians, suggesting the opportunity to diagnose the severity of OCD and depression without the necessity for formal clinical interviews.

2.2.2 Acoustic modality

The human voice is proven to be highly expressive in conveying the speaker's emotion to the listener (Pope et al., 1970). A smiling speaker alters her vocal tract such that "the smile is audible" (Tartter, 1980), while an anxious speaker produces vocal disturbances and varied speech-rate that can be interpreted as hesitation (Cook, 1969; Kasl and Mahl, 1965). Depressed speakers, on the other hand, tend to present reduced speech variability and monotonicity in loudness and pitch, reduced speech and articulation rate, and increased pause duration (Dibeklioğlu et al., 2018).

Although these auditory cues may not be as easily recognizable to untrained ears as the speaker might have believed (Goberman et al., 2011), an ample number of studies has proven the efficacy of predicting anxiety from adult voice and speech. Some examples of the conventional and well established approach are provided in Han et al. (2020) and Baird et al. (2020). In these works, the OpenSMILE toolkit is utilized to extract the Computational Paralinguistics Challenge (ComParE) and Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) feature sets from a set of audio samples.[3] ComParE is a large-scale brute-force feature set used in a number of INTERSPEECH Computational Paralinguistics Challenges since 2013. It contains 6373 dimensional static features produced by statistical computations over 65 low-level descriptor (LLD) contours, which themselves consist of spectral (e.g., relative spectra auditory bands 1-26, spectral energy, spectral slope, spectral sharpness, spectral centroid), cepstral (i.e., Mel frequency cepstral coefficient 1-14), prosodic (e.g., loudness, root mean square energy, zero-crossing rate, F_0 via subharmonic summation), and voice quality features (probability of voicing, jitter, shimmer and harmonics-to-noise ratio) (Schuller et al., 2013). In contrast, the eGeMAPS

[3] https://audeering.github.io/opensmile-python/api/opensmile.FeatureSet.html

feature set is considerably smaller, comprising 88 features derived from 25 LLDs that are also proven descriptive for identifying affective psychological effect on voice production (Eyben et al., 2015).

Han et al. (2020) utilize the aforementioned audio features to predict the level of anxiety, fatigue, and sleep quality of COVID-19 patients. They preprocess the voice segments data by employing speaker diarisation to annotate speaker identities for each voice segment, and speech transcription which converts voice segments from the targeted speakers into text transcriptions. As the classifiers, the authors opt for Support Vector Machine (SVM) with a linear kernel based on the scikit-learn library. Leave-One-Subject-Out (LOSO) cross-validation is chosen as the evaluation strategy as it is particularly well-suited for dealing with small datasets. The classifiers are trained to optimize the following metrics: Unweighted Average Recall (UAR), overall accuracy (also referred to as Weighted Average Recall or WAR), and the F1 Score (alternatively known as F-score or F-measure). The best performances were achieved in predicting sleep quality and anxiety level, while the performance on fatigue prediction was unsatisfactory although still above chance. Further, when comparing two selected feature sets, the compact eGeMAPS set consistently outperformed the large-scale ComParE feature set.

In a similar work by Baird et al. (2020), besides ComParE and eGeMAPS, the authors also utilize the DEEP SPECTRUM toolkit which has shown success in similar audio- and speech-based tasks in the INTERSPEECH challenge. For the classifier, the authors choose to train a series of epsilon-support vector regressor (SVR) models with a linear kernel, optimizing the complexity parameters C. Their performance are evaluated using *Spearman's correlation coefficient* (ρ) due to the ordinal nature of the raw Beck Anxiety Inventory (BAI) scores that the models predict. This study finds that an anxious person produces more variations in speech characteristics, which is most detectable when the speaker is smiling. In terms of technique, late-fusion of best results from eGeMAPS and DEEP SPECTRUM feature sets is found to produce the best performance.

Hinduja et al. (2024) provide a more elaborated example where the authors utilize OpenSMILE and Collaborative Voice Analysis Repository (COVAREP) (Degottex et al., 2014) to extract acoustic feature sets before combining them with visual and linguistic modalities to predict OCD severity and comorbid depression severity. OpenSMILE is used to extract eGeMAPS feature set that contains 62 features: arithmetic mean and coefficient of variation of 18 low-level descriptors (LLD), 8 functionals applied to loudness and pitch LLD, and 6 temporal features. COVAREP extracts 72 low-level speech acoustic features derived from the speech signal including pitch, energy, spectral envelope, loudness, voice quality and other characteristics. Interestingly, while multimodal models seem to perform best, unimodal model that uses voice acoustics alone approached the accuracy of the best multimodal model. The authors hypothesize that this is because of the inherent dynamism of voice and its connection to the vagus nerve, rendering it highly expressive in capturing variations in arousal and stress level.

Dibeklioğlu et al. (2018) explore a number of vocal features, namely fundamental frequency (F_0), Maxima Dispersion Quotient (MDQ), Peak Slope (PS), Normalized Amplitude Quotient (NAQ), Quasi Open Quotient (QOQ), and switching pause durations. In this paper, the authors find that only switching pause durations (or latency to speak; the pause duration between the end of one speaker's utterance and the start of the other speaker's utterance) and F_0 are correlated with depression severity.

Studies by Laukka et al. (2008) and McGinnis et al. (2019) employ techniques that can function with a limited computational resources and a small number of audio features. The former uses seven acoustic features, namely mean, standard deviation, and maximum of F_0, mean voice intensity, HF 500, speech rate, and ratio of silence, to predict social anxiety and nervousness during speech. This work indicates that anxiety impacts the nonverbal qualities of speech reflected in the features of the voice, and that tone of anxiety in speech is perceivable by the listener. Similarly, the latter study computes Davies-Bouldin Index-based feature selection to produce eight features from a small subset of three-minutes audio data and use them as input to a logistic regression model. While the performance in detecting internalizing disorders (which includes anxiety, depression, and somatic symptoms) is moderate, it can potentially be improved with higher quality audio data.

Kwon et al. (2022) show that their approach can detect depression and anxiety using acoustic features alone by capturing 23 Mel Frequency Cepstral Coefficients (MFCCs) from phone conversations with participats diagnosed with depression and anxiety. The features are then fed into a pre-trained SRE16 X-vector Model for feature extraction, which is followed by a trainable classifier model. The classifier architecture is a convolutional neural network (CNN) with three 1-D convolutions (each is followed by batch ReLU activation, batch normalization, and dropout) followed by a fully connected layer with 32 nodes, and an output layer with two nodes and softmax activation. In total, the X-vector and the classifier contains 4.5 million parameters, of which 0.28 million are trainable. This approach manage to achieve comparable sensitivity and specificity to self-reports such as General Anxiety Disorder-7 (GAD-7) and Patient Health Questionnaire-8 (PHQ-8).

Diep et al. (2022) performs a similar experiment that combines deep-learned acoustic and linguistic features—extracted using Wav2Vec 2.0 (Baevski et al., 2020) and RoBERTa, respectively—with 16 hand-crafted features proposed by domain experts to train an AudiBERT-like model. The model comprises a two-layer Bi-directional Long Short-Term Memory (biLSTM) connected to a multi-head attention layer with two heads. Their work shows that augmenting hand-crafted features with deep-learned features improves the overall classification F1 score in depression and anxiety detection (compared to a baseline of hand-crafted features alone).

2.2.3 Linguistic modality

Various studies from the field of personality psychology (Fast and Funder, 2008; Ireland and Mehl, 2014; Schwartz et al., 2013; Yarkoni, 2010) have found associations between the content of speech and the speaker's traits. As such, this modality can also be used as an indicator to the speaker's psychological condition. One canonical method of psychological analysis of speech is using the Linguistic Inquiry and Word Count (LIWC) program developed by James Pennebaker and colleagues (Boyd et al., 2022), which efficiently maps the content of text bodies on a number of psychological dimensions.

Bilalpur et al. (2023) employ text analysis to detect depression by capturing concepts such as analytic capabilities, focus on past events, negations, comparisons, humor and quantitative aspects of the spoken content. Among the 92 verbal features extracted using the LIWC framework are usage of negations and comparisons, informal language and humor, and referencing past events and agreements. The authors choose SVM as the classifier and determine the best hyperparameters (linear and polynomial kernels and C value in the range of 105 to 103 on a log-scale) with a grid-search and five-fold cross-validation (CV). The best hyperparameters were used for Leave-One Subject-Out (LOO) cross-validation to report accuracy (ACC) as well as positive and negative agreement. This study found that this modality is the best performing unimodal and the most contributing modality (6 out of 15 features) to the best performing multimodal model. Additionally, dictionaries from LIWC are also utilized in a number of other experiments, e.g., to measure the prevalence of absolutist words in speeches produced by people who are suspected to have anxiety, depression, and suicidal ideation (Al-Mosaiwi and Johnstone, 2018), to find correlation between linguistic expression in blogs and symptoms of depression, generalized anxiety, and suicidal ideation (O'Dea et al., 2021), and to find linguistic markers of grandiose narcissism (Holtzman et al., 2019).

2.3 Clinical application

Successful development of a predictive system does not always entail its application in clinical practices. Ahn and Busemeyer (2016) went further and stated that "no computational methods have been successfully translated into clinical settings." The authors identify at least three major issues. First, it is often difficult to precisely characterize the neurocognitive processes underlying human behaviors and decision-making, and more cross-discipline collaborations are needed between the field of psychology, mathematics, and computer science. Second, the design of diagnostic tests are sometimes too difficult or contain confounding factors that hinder maximal decomposition of the underlying processes and interactions between decision-making systems. Third, the size of sample data that is often insufficient to reflect real-world condition.

In order to improve the clinical utility of predictive systems, Ruberg et al. (2023) propose that developers adopt the same rigorous clinical trials that are used in the development of other medical interventions. Furthermore, they also recommend a more comprehensive way in assessing the performance of the system. Currently, developers rely too much on the area under the receiver operating characteristic curve (AUC) as a performance metric. The authors argue that the prevalence (P) of the disease or prognosis in question, which affects positive and negative predictive value (PPV and NPV), are just as critical to clinicians and should be taken into account.

Aspect	Manual detection	Automated detection
Indicator	Subjective distress/self report.	Objective presence of symptoms.
Proneness to bias	More prone to untruthful report, self suppression, and bias.	Less prone to self suppression, able to pick up on small cues that are hard to detect, and more consistent.
Cost of access	More expensive as it involves the labor of experts and takes substantially more time.	Less expensive in terms of labor and time, thus more accessible.
Scale	Limited to a smaller audience with resources to afford the service.	Because of the low barrier, the service can be scaled up to reach a wider population.
Early detection	Less opportunity, as it requires manual detection and diagnosis.	More opportunity, as it is more accessible and inexpensive.

Table 1. Potential benefits of automated detection.

Having explored the challenges, the current psychiatric approach to anxiety detection leaves rooms for enhancements through computational methods (see Table 1). First, the various manual diagnostic methods such as Diagnostic and Statistical Manual (DSM), Child Behavior Checklist (CBCL) including its various subscales, Zung Self-Rating Anxiety Scale[4], Beck Anxiety Inventory (BAI), Generalized Anxiety Disorder Scale, or Yale-Brown Obsessive-Compulsive Scale (YBOCS) relies on the level of subjective distress and

[4] www.healthline.com/health/anxiety-diagnosis#diagnostic-tests

suffering as the basis to discern normal from pathological anxiety (Knepley et al., 2019; Read et al., 2015). This introduces biases and inaccuracies in the diagnosis.

Second, it is dilemmatic to assume that anxiety is pathological when it impairs quality of life. A differing and equally valid view, for example from Bateson et al. (2011), suggests that the level of anxiety is the result of interaction between the probability of threat occurring and the cost that the threat may incur. Thus, in some cases where the level of danger is overwhelming, severe anxiety may in fact be the appropriate level of response. As a corollary, insensitive anxiety mechanism could be a common dysfunction which goes largely unreported because those affected tend to not seek help. The sufferer may also try to deny and suppress the emotion and not making a truthful report, thus preventing the opportunity for early detection.

Third, and more pertinent to this study, is that because the initiative and measurement relies on patient's self-report or expert assessment, early detection becomes less accessible to sufferers who lack familiarity with psychological disorders and/or capability to articulate emotions, such as young children, less informed population, or those with diminished mental and cognitive capacity. This is where this study comes in, not as a replacement of human experts but as a technological aid for preliminary screening, serving the overarching purpose of making mental health diagnosis more accessible.

2.4 Previous works on the YOUth dataset

Child and teenager data from the YOUth cohort study had been the basis of number of works. Holleman et al. (2021) study the characteristics of gaze and its relation to speech behavior during video-mediated face-to-face interactions between 81 pairs of parents and their preadolescent children. The study utilizes the eye-tracking, frontal video, and audio data from the YOUth cohort. In another study, Buimer et al. (2022) investigate the impact adverse childhood experiences (ACEs) on the morphology of the developing brain of children aged between 8 to 11 years old. They utilized the data on subcortical volumes, cortical thickness, cortical surface area and fractional anisotropy in regions of interest in brain scans acquired in 1,184 children from the YOUth cohort.

3. Methodology

This chapter describes the dataset and the methods applied to extract the features from various modalities. Further, it explains the modeling pipeline and performance evaluation metrics used in this study.

3.1 Dataset

The data used in this thesis come from YOUth Cohort Study[1] (Onland-Moret et al., 2020), an extensive longitudinal study in child development conducted in the Netherlands. This study tracks the development of nearly 4,000 Dutch children from pregnancy to early adulthood. The study focus on neurocognitive development involved in two core characteristics of behavioural development: social competence and behavioural control. There are two primary cohorts of participants: Baby & Child, covering the period from pregnancy to 7 years, and Child & Adolescent, spanning ages 8 to 16 years. With a specific emphasis on neurocognitive development, the research focuses on unraveling the intricacies of two fundamental aspects of behavioral development: social competence and behavioral control. This rich database encompasses an array of sophisticated measures, including 3D-ultrasound scans of the fetal brain, eyetracking, EEG, fMRI, computer-based tasks, cognitive assessments, and parent-child observations. Also included in the database is a diverse observational data such as behavior, personality, health, lifestyle, parenting dynamics, child development, and the use of (social) media. Furthermore, data from biomaterials like (umbilical) blood samples, buccal swabs, saliva, and hair samples are also provided.

We used a sample set of 200 video recordings from 100 parent-child dyads. The children comprised 52 females and 48 males aged 9 years. Each dyad appeared in 2 videos, showing the child and his/her parent interacting in a laboratory setting for around 10 minutes. They were instructed to conduct two types of interactions: cooperative (i.e., planning a vacation) and conflicting (i.e., arguing).

Before the sessions, each parent filled the Child Behavior Checklist (CBCL; Verhulst et al. (1996)) questionnaire, which is a widely used questionnaire to assess behavioral and emotional problems in children (Mazefsky et al., 2011). It is generally divided into two age brackets, the pre-school age (1.5 to 5 years) and the school age (6 to 18 years) with each further divided into various subscales corresponding to specific disorders (Knepley

[1] https://www.uu.nl/en/research/youth-cohort-study

et al., 2019; Read et al., 2015). The school age questionnaire contains a list of 118 observational items, 13 of which are related to anxiety (also known as the CBCL-A/D or the anxious/depressed subscale, see Table 2). The score on each item is an integer between 0 and 2, with 0 means *Not True (as far as you know)*, 1 means *Somewhat or Sometimes True*, and 2 means *Very True or Often True*. Hence, the maximum possible score of anxiety-related symptoms is 26.

Item	**Description**
14	Cries a lot
29	Fears certain animals, situations, or places other than school
30	Fears going to school
31	Fears he/she might think or do something bad
32	Feels he/she has to be perfect
33	Feels or complains that no one loves him/her
35	Feels worthless or inferior
45	Nervous, high-strung, or tense
50	Too fearful or anxious
52	Feels too guilty
71	Self-conscious or easily embarrassed
91	Talks about killing self
112	Worries

Table 2. Anxiety-related CBCL items.

The average score of anxiety-related symptoms in the normative sample for the boys is 2.8 (SD = 2.7) and for the girls is 3.2 (SD = 2.9). For the purpose of classification, we first calculated the Z-score, which is the individual score's distance away from the mean in terms of standard deviation (Equation 3.1; x is the anxiety score of the individual observation, μ and σ are the mean and standard deviation of the respective gender group in the population). The Z-scores were then converted to T-scores by scaling to have a mean of 50 and standard deviation of 10 (Equation 3.2).

$$Z = \frac{x - \mu}{\sigma} \tag{3.1}$$

$$T = (Z \times 10) + 50 \tag{3.2}$$

Once the T-scores were obtained, they were classified into two classes: class 0 for the healthy participants and class 1 for the clinically anxious and the borderline clinically anxious participants. The thresholds for classification were: score >= 70 is in clinical range, 65 <= score < 70 is the borderline clinical range, and score < 65 is healthy. The score distribution is shown in Figure 1.

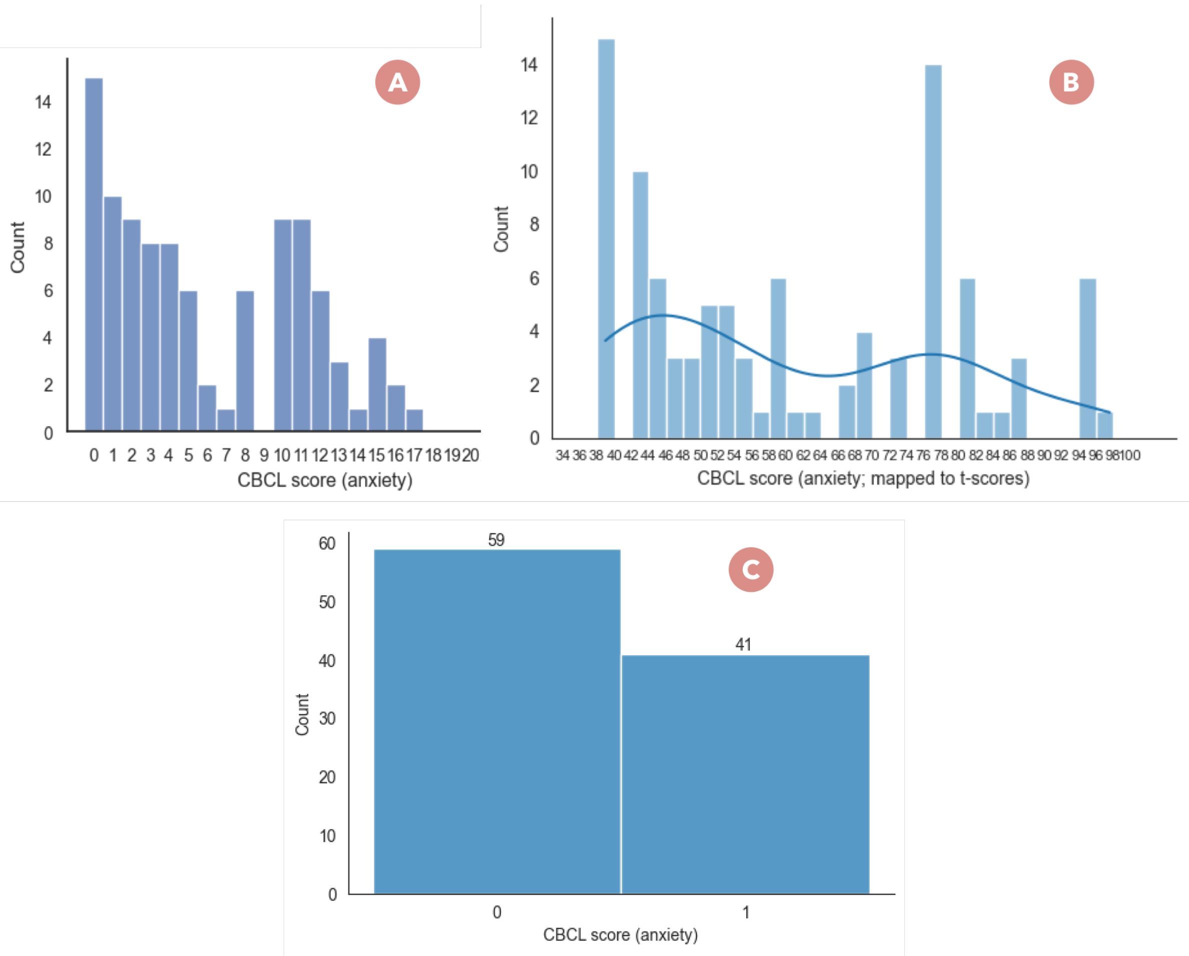


Figure 1. Distribution of the participants' CBCL scores. (A) Anxiety scores from the CBCL questionnaires. (B) Anxiety scores transformed into T-scores, which are the deviations from the mean in terms of standard deviations, scaled to a mean of 50 and a standard deviation of 10. (C) The T-scores are classified into two classes, class 0 (healthy) and class 1 (borderline and clinically anxious), using a threshold of 65.

3.2 Feature extraction

This chapter describes our feature engineering process. Using the audio and visual data from the videos, we were able to extract features for four modalities. The acoustic characteristics of the audio produced the acoustic features, and analysis of the content of the speech produced the linguistic features. We also tracked the facial expressions and the body postures depicted on the video frame images and convert them into facial expression and body pose feature sets. While we studied numerous approaches of feature extraction, the main inspirations behind our method are the work of Hinduja et al. (2024) and Jung et al. (2023).

3.2.1 Acoustic modality

To analyze the emotional states expressed in the voice of the speakers, the sound signals were converted into a set of 88 acoustic parameters of the *extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS; Eyben et al. (2015)) of OpenSMILE. The set was then used as training features for the model. The outline of the process is depicted in Figure 2. In this section we will discuss the steps and tools that were applied during the process in greater details.

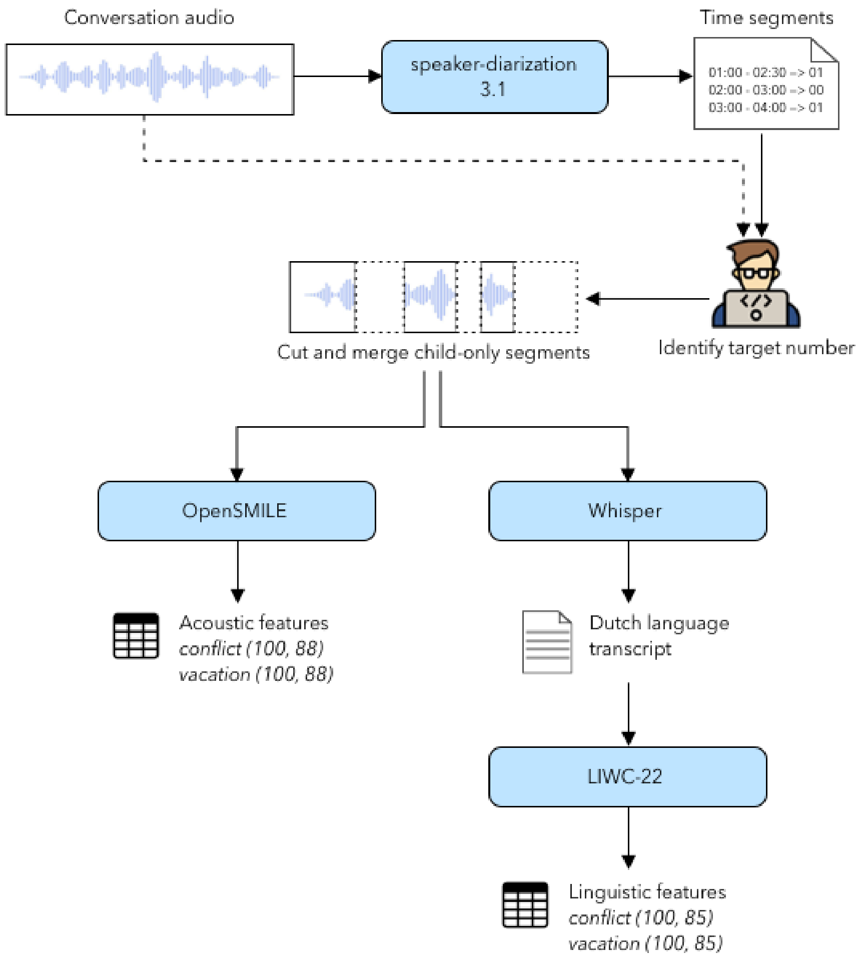


Figure 2. Acoustic and linguistic features extraction pipeline.

Speaker diarization with PyAnnote

Since the audio samples contained multiple speakers, a number of preprocessing steps were taken before the acoustic features could be extracted. The first step was speaker diarization, which is the task of partitioning an audio stream into homogeneous temporal segments according to the identity of the speaker (Plaquet and Bredin, 2023). This was done using the pretrained model *speaker-diarization version 3.1* from PyAnnote[2]. The

[2] https://huggingface.co/pyannote/speaker-diarization-3.1

speaker diarization process pipeline is outlined in Figure 3.

PyAnnnote works by first collecting fixed-length sub-sequences at random from the training set to form mini-batches. This improves the variability of the training data and make for more efficient training time. Then, it performs a number of detections, namely: *voice activity detection* where the model predicts for each time step whether it contains a speech or noise, *speaker change detection* in which the model marks each time step where it predicts a speaker switch as well as 200ms surrounding the switching point, and *overlapped speech detection* in which it detects the time steps where there are more than one speaker. The time segments and labels are refined through an unsupervised process where a new model is trained from scratch and applied on the very same file it was trained on. The next step involves calculating the distances between the speaker embeddings, which reflects the similarity between any two embeddings (commonly done using cosine distance). Finally, based on these distances, a clustering algorithm groups similar embeddings together and assigns it to a speaker identity (Bredin et al., 2020; Bredin and Laurent, 2021; Bredin, 2023).

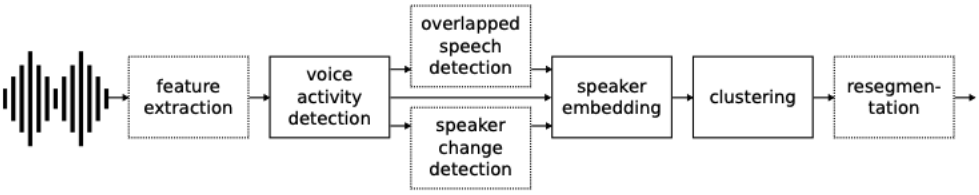


Figure 3. Working pipeline of the PyAnnote speaker diarization tool (Bredin et al., 2020).

We specifically set the *num_speakers* parameter to 2 as this increased the accuracy of the model. Although the moderators could sometimes be heard, we did not set the parameter to 3 as the moderator segments were much shorter compared to the parent-child's segments. We would also like to avoid the risk of losing some of the child's segments which, in some cases, were rather short.

The output for each audio file was a list of time segments, each assigned a speaker ID. Since this number is assigned randomly, the child's ID differs between videos. Therefore, a manual process was required to identify the child's ID in each recording. This was done by carefully listening to each recording and cross-referencing with the time segments produced by PyAnnote. Once identified, the audio file was truncated by removing time segments that were not assigned the child's ID. Some segments contained voices from multiple speakers, i.e., the child's voice overlapped with either the parent or the moderator. If their voices exceed the amplitude threshold, the segment was assigned to all actors (Bredin, 2023). Such segments were also kept as it was not feasible to separate the actors using our available tools.

eGeMAPS extraction with openSMILE

After separating the child's audio segments from other speakers, the resulting audio file was processed to extract the audio features. To achieve this, we leveraged a toolkit called openSMILE (Open-source Speech and Music Interpretation by Large-space Extraction).[3][4] The tool offers the capability to extract various pre-defined acoustic feature sets, including ComParE, GeMAPS, and eGeMAPS.

After conducting some preliminary benchmarking and considering our limited sample size, we decided to work with eGeMAPS v2. This set is a popular acoustic feature set commonly used as a baseline for analyzing speech and other vocal sounds. It is lauded for its ability to capture physiological changes in voice production during emotional states as well as its successful applications in past research. The set was developed in two configurations: minimalistic (GeMAPS) and extended (eGeMAPS). The minimalistic set comprises 62 parameters that capture various aspects of speech, including frequency, energy, spectral balance, loudness, pitch, and temporal characteristics (Eyben et al., 2015). The extended set adds 26 more features related to spectral flux, Mel-Frequency Cepstral Coefficients (MFCC), and sound level. Despite its relatively small size (88 parameters), eGeMAPS has demonstrated performance comparable to much larger brute-forced baseline acoustic feature sets of the INTERSPEECH 2009 Emotion Challenge (ranging from 384 to 6,373 parameters) in tasks such as automatic recognition in binary valence and binary arousal dimensions (and in some cases, outperformed the bigger set; for example see Han et al. (2020)).

3.2.2 Linguistic modality

To get the content of the subject's speech, his or her voice was transcribed using OpenAI's Web-scale Supervised Pretraining for Speech Recognition (WSPSR, or *Whisper*).[5] Then, the Linguistic Inquiry and Word Count (LIWC, or the latest software iteration, LIWC-22)[6] text analysis toolkit was applied to analyze the affective state of the speaker. The output is a linguistic feature set comprising 85 parameters. The procedure and tools that were applied are further described below.

Speech transcription with Whisper

Following the speaker diarization process (explained in Subsection 3.2.1), the resulting audio files were transcribed using Whisper. Specifically, we utilized the multilingual model to perform transcription in the Dutch language and opted for the largest model size in order to obtain the highest accuracy.

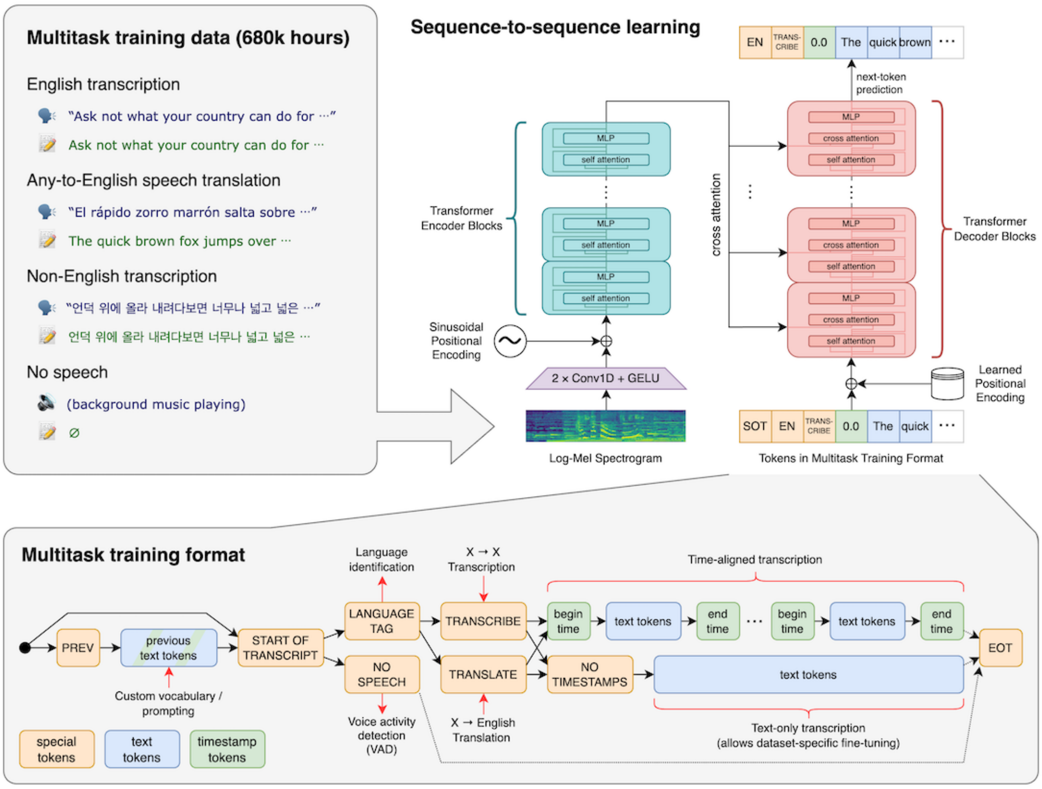
[3] https://github.com/audeering/opensmile

[4] https://audeering.github.io/opensmile-python/index.html

[5] https://github.com/openai/whisper/tree/main

[6] https://liwc.app

Whisper is a general-purpose speech recognition model trained on a large dataset of diverse audio samples. It offers multiple capabilities such as multilingual speech recognition, speech translation, and language identification. The model is based on the encoder-decoder architecture (transformer) and uses a special input embedding with tokens that represent the combination of tasks for which the model was trained (see Figure 4). This allows the model to perform multiple tasks (e.g., voice activity detection, alignment, speaker diarization, inverse text normalization, transcription, translation, and language identification) that traditionally requires many different stages (Radford et al., 2023).



Multitask training data (680k hours)
English transcription
"Ask not what your country can do for ⋯"
Ask not what your country can do for ⋯
Any-to-English speech translation
"El rápido zorro marrón salta sobre ⋯"
The quick brown fox jumps over ⋯
Non-English transcription
"언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯"
언덕 위에 올라 내려다보면 너무나 넓고 넓은 ⋯
No speech
(background music playing)
∅
Sequence-to-sequence learning
EN
TRANS-CRIBE
0.0
The
quick
brown
⋯
next-token prediction
MLP
cross attention
self attention
Transformer Encoder Blocks
Transformer Decoder Blocks
cross attention
Sinusoidal Positional Encoding
2 x Conv1D + GELU
Log-Mel Spectrogram
Learned Positional Encoding
SOT
EN
TRANS-CRIBE
0.0
The
quick
⋯
Tokens in Multitask Training Format
Multitask training format
Language identification
X → X Transcription
Time-aligned transcription
PREV
previous text tokens
START OF TRANSCRIPT
LANGUAGE TAG
NO SPEECH
TRANSCRIBE
TRANSLATE
begin time
text tokens
end time
⋯
begin time
text tokens
end time
NO TIMESTAMPS
text tokens
EOT
Custom vocabulary / prompting
Voice activity detection (VAD)
X → English Translation
Text-only transcription (allows dataset-specific fine-tuning)
special tokens
text tokens
timestamp tokens

Figure 4. Overview of Whisper's multitask speech recognition learning approach (Radford et al., 2023).

Feature extraction using LIWC

LIWC-22 is a text analysis tool based on the theory that word choice carries significant psychological value and reflects the mental construct of the speaker. At the heart of LIWC's analysis strategy is a series of dictionaries, essentially mappings between sets of textual components (e.g., words, word stems, phrases, emoticons) and important psychological, emotional, and social concepts and theories (Boyd et al., 2022). In total, there are over 12,000 textual components spread over many categories. Each entry can be part of multiple dictionaries. For example, the word *cried* is part of ten dictionaries: affect, tone_pos, emotion, emo_neg, emo_sad, verbs, focuspast, communication, linguistic, and cognition. Thus, each time the word appear in the input text, these ten dictionaries would gain increased relevance. The dictionaries themselves are arranged hierarchically (e.g., the emo_sad dictionary is part of the emo_neg dictionary, which is part of the emotion dictionary, which in turn is part of the broader affect category). LIWC has been applied successfully in a number of works, such as Zhao et al. (2016); Yu et al. (2023); Kane and van Swol (2023).

This thesis used the Dutch translation of LIWC dictionary (2015 version) created by Peter Boot.[7] As shown in Figure 2, after the speech transcription, the output text was analyzed using LIWC-22 to extract 85 linguistic features. Broadly, these features consisted of summary features (e.g., word count, average number of words per sentence), percentages of punctuation marks (e.g., question mark, exclamation mark), and percentages of psychological concepts present in the input text (i.e., concepts LIWC's dictionaries). To illustrate, if a speaker uttered 100 words of which 5 were expressions of sadness (dictionary: emo_sad) and 20 were 3rd person plural pronouns (dictionary: they), then LIWC would report them as: 0.05 emo_sad and 0.2 they (as well as other related dictionaries).

3.2.3 Visual modality

The facial expressions and body gestures of the subjects throughout the interactions were analyzed to predict their anxiety level. We employed the Py-Feat library[8] to detect the presence and probability of 20 Facial Action Units (AUs) on the face of the subject. We also applied OpenPose[9] to detect the presence and locations of 15 Key Points on the subject's body. These datasets were then used to train our models. Note that although both facial expressions and body gestures come from visual data, they were treated as separate modalities, as illustrated in Figure 5.

Sub-frame extraction

Just as the audio files required some preprocessing, the videos also required some preparation work before the feature extraction could be implemented. Each video frame comprises four sub-frames: two single-subject frames that captured the child and the parent, one frame that had the dyad together, and one empty frame. Since the location of the child's sub-frame varied across videos, a manual process was required to note the child's sub-frame in each video. These target frames were then cropped and stored as images.

While each video had around 10,000 to 20,000 frames, the small time interval between between subsequent frames (shot at 60 fps) meant that the amount of differences between adjacent frames was negligible. For efficiency, only one in every five frames was captured. This way, we were able to significantly shorten the computing cost without notable information loss since at this rate, the time difference between adjacent frames was still below 100 ms.

[7] Contact listed on the site: peter.boot@huygens.knaw.nl

[8] https://py-feat.org

[9] https://github.com/CMU-Perceptual-Computing-Lab/openpose

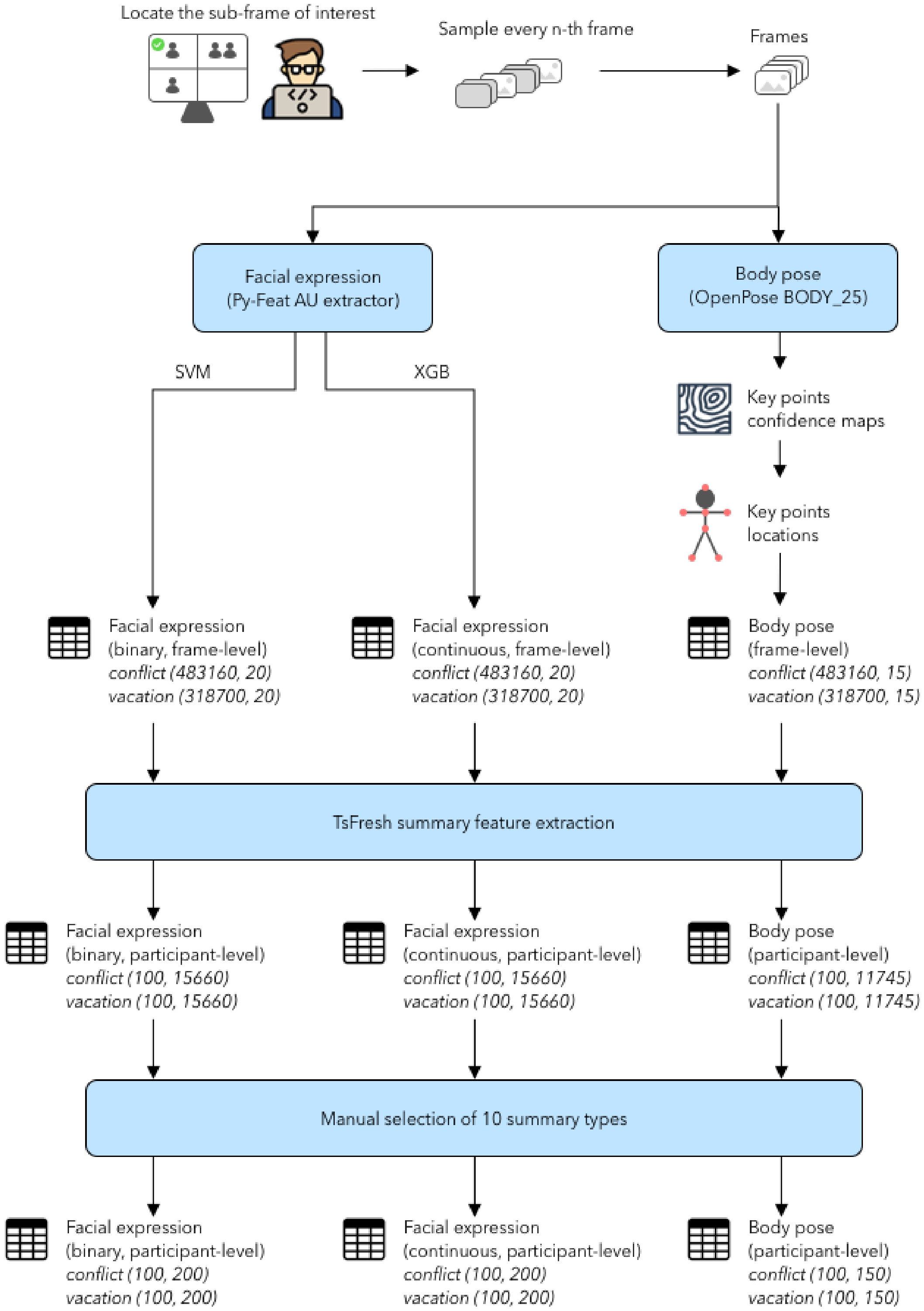


Figure 5. Visual (facial expression and body gesture) features extraction pipeline.

Facial expression detection with Py-Feat

The Facial Action Coding System (FACS) is a systematic mapping of human facial movements. Invented by Carl-Herman Hjortsö in 1978, it was further developed into today's mature form by by Paul Ekman and Wallace V. Friesen in 2002 (Ekman and Friesen, 1978). FACS is a mapping system of facial expressions based on facial Action Units (AU). The activation of an AU signifies that the corresponding muscle or group of muscles deviate from a neutral face expression (see Figure 6). Hence, an expression can be estimated by knowing which AUs are activated.

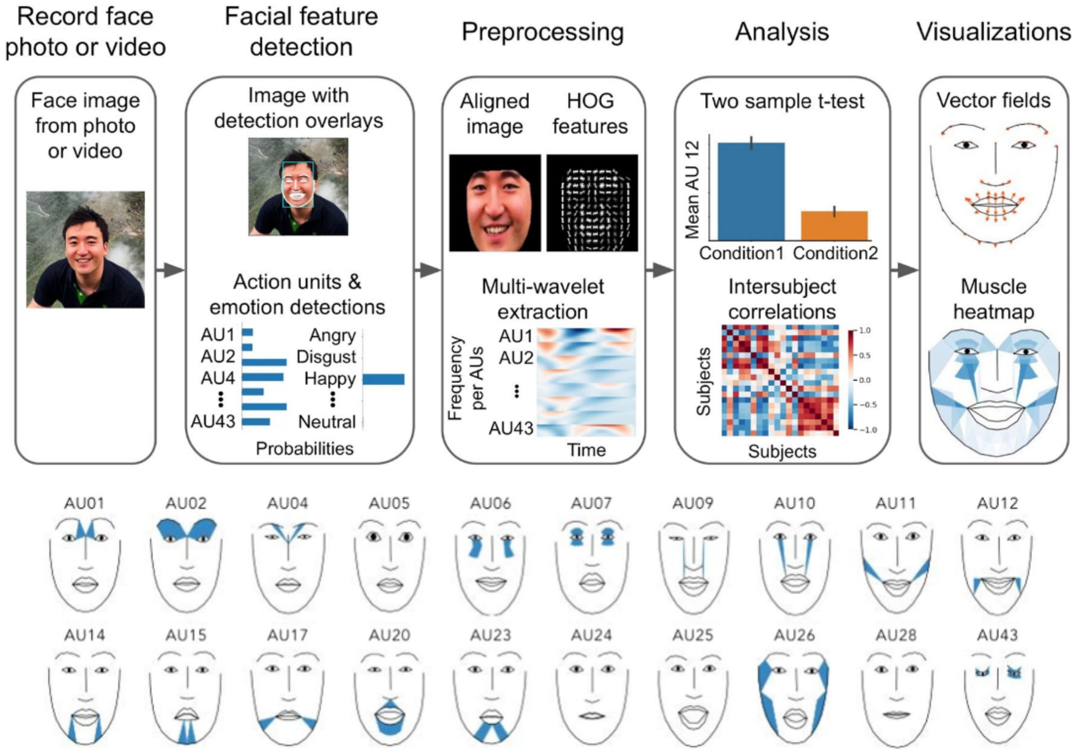


Figure 6. (Above) Py-Feat facial expression analysis pipeline. (Below) Visualization of the resulting facial expressions when an AU is activated (Cheong et al., 2023).

In this work, we leveraged a FACS implementation from an open-source Python toolbox named Py-Feat to estimate the activation probability of 20 AUs on each video frame (listed in Table 3). Although there are 66 known AUs at the time of this writing, most Py-Feat models support a subset of about 25 AUs corresponding specifically to facial muscles. Other AUs are not commonly used to train models as they are associated with eye or head movements.[10]

Py-Feat shares a similar architecture to OpenFace for AU detection (see the pipeline in Figure 6). It first applies a combination of facial landmarks detection and convex hull algorithm to determine the face region. Then it extracts Histogram of Oriented Gradients (HOG) of the face, which captures local information about the intensity gradients and outlines the areas where there are sharp changes in intensity. Finally, the extracted HOG feature is compressed using Principal Component Analysis (PCA) where only the most

[10] https://py-feat.org/pages/au_reference.html

AU number	Description	AU number	Description
01	Inner brow raiser	14	Dimpler
02	Outer brow raiser	15	Lip corner depressor
04	Brow lowerer	17	Chin raiser
05	Upper lid raiser	20	Lip stretcher
06	Cheek raiser	23	Lip tightener
07	Lid tightener	24	Lip pressor
09	Nose wrinkler	25	Lip part
10	Upper lip raiser	26	Jaw drop
11	Nasolabial deepener	28	Lip suck
12	Lip corner puller	43	Eyes closed (muscle relaxation)

Table 3. Action Units and their functions.

relevant information is preserved. This compressed representation is the used to train a model to predict the Action Units (Cheong et al., 2023; Baltrusaitis et al., 2018).

Py-Feat offers two alternative AU detection models, an XGBoost classifier model and an SVM model. The XGBoost model, which is the default, returns a continuous value for each AU, representing the probability of its activation. The alternative SVM model returns a binary value, denoting its presence/activation.[11] We experimented with both alternatives and compared their performances, which will be discussed in Chapter 4.

Summary features extraction using Tsfresh

The frame-level features extracted by Py-Feat had to be transformed into participant-level features before they are usable for model training. For this purpose, we leveraged the Tsfresh[12] library, a Python toolkit based on the FeatuRe Extraction based on Scalable Hypothesis tests (FRESH) algorithm (Christ et al., 2018). Overview of the toolkit's functionalities is shown in Figure 7.

The first functionality Tsfresh offers is feature extraction, where it converts time-bound features into summary features by calculating various statistical characteristics. These include metrics like minimum, maximum, mean, mean absolute change, median, variance, skewness, kurtosis, counts (above/below mean), and so on. From observation, Tsfresh calculates around 783 of such metrics for each feature, which led to a feature vector with thousands of dimensions (15,660 in the case of our face expression feature set).

[11] https://py-feat.org/pages/models.html

[12] https://tsfresh.readthedocs.io/en/latest/index.html

$$H_0^n : \{X_n \text{ and } Y \text{are independent}\} \text{ and } H_1^n : \{X_n \text{ and } Y \text{are dependent}\} \tag{3.3}$$

The second functionality offered is feature selection where Tsfresh deploys a statistical test to find the relevance of a given feature X in predicting the target variable Y (Christ et al., 2016). That is, given *n* extracted features, for $X \in \{X_1, X_2, ..., X_n\}$ it tests the hypotheses in Equation 3.3. This is done for each extracted feature. And, depending on the data type, a different test is used (i.e., Exact Fisher Test of Independence for when both the feature and target variable are binary variables, Kendal Rank Test if both are continuous variables, and the Kolmogorov-Smirnov Test when one is binary and the other is continuous). Finally, using the resulting vector of p-values, the Benjamini-Yekutieli procedure (Benjamini and Yekutieli, 2001) is performed in order to control the rate of falsely rejected null hypothesis (the False Discovery Rate, FDR).

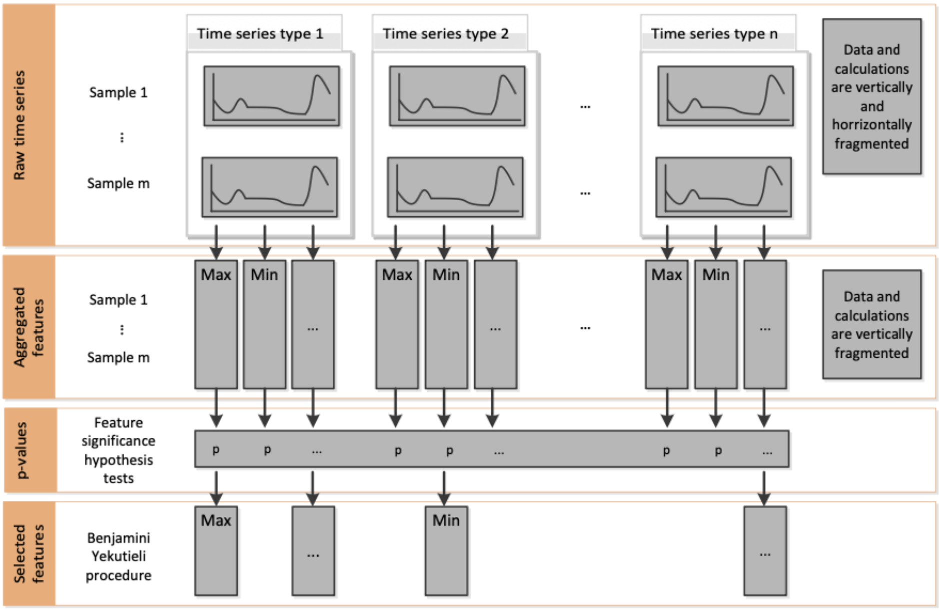

Raw time series
Time series type 1
Time series type 2
Time series type n
Sample 1
⋮
Sample m
...
Data and calculations are vertically and horrizontally fragmented
Aggregated features
Max
Min
...
Data and calculations are vertically fragmented
p-values
Feature significance hypothesis tests
p
Selected features
Benjamini Yekutieli procedure

Figure 7. Tsfresh feature extraction and selection scheme (Christ et al., 2016).

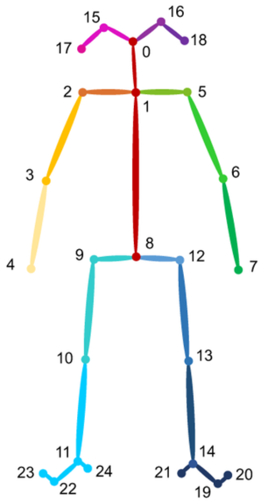
We encountered a number of issues when utilizing Tsfresh's feature selection. Namely, it often selected features that are not easily explainable without considerable mathematical comprehension (e.g., *cwt_coefficients__coeff_0__w_2__widths_(2, 5, 10, 20)* and *fft_coefficient__attr_"real"__coeff_14*). Moreover, Tsfresh's calculation often resulted in very few or even zero features being selected. Subsequent inquiry into literature on Tsfresh revealed some dilemmas with its feature selection mechanism. Several studies on feature extraction strategies found that while Tsfresh produces highly effective features (i.e., leading to higher performance compared to other feature sets), it often comes at the expense of interpretability. The features are difficult to interpret without extensive domain knowledge, and even experts cannot guess at feature definitions from the feature names

(Bosch, 2021). Moreover, the set contains many redundancies, with one study found that 90% of the variance of 779 Tsfresh features was captured with 55 principal components, and around 50% was derived from Fast Fourier Transform (FFT) coefficients (Henderson and Fulcher, 2021).

We believe that in the context of medical application, a transparent and humanely comprehensible mechanism of decision making is critical in order for the AI system to be trusted. Hence, from comparison with other studies (Bilalpur et al., 2023; Radhakrishnan et al., 2023; de Santis et al., 2023), we heuristically selected 10 types of summary statistics: minimum, maximum, mean, median, standard deviation, variance, coefficient of variation, skewness, kurtosis, and mean absolute change. This set was chosen as they are considered common statistical measurements that are highly interpretable without requiring extensive domain knowledge, and the amount was approximated to suppress the feature dimensionality and prevent overfitting. The output is a facial expression feature set with a shape of (100, 200).

Key Points tracking with OpenPose

We utilized the OpenPose library (Cao et al., 2021; Simon et al., 2017; Wei et al., 2016) to extract the gestures of the subject. We specifically used OpenPose's BODY_25 model as it is optimized for the task of estimating body poses (contrary to two other alternative models, namely COCO and MPII).[13]



0 - Nose
1 - Neck
2 - Right shoulder
3 - Right elbow
4 - Right wrist
5 - Left shoulder
6 - Left elbow
7 - Left wrist
8 - Mid hip
9 - Right hip
10 - Right knee
11 - Right ankle
12 - Left hip

13 - Left knee
14 - Left ankle
15 – Right eye
16 – Left eye
17 – Right ear
18 – Left ear
19 - Left big-toe
20 - Left small-toe
21 - Left heel
22 - Right big-toe
23 - Right small-toe
24 - Right heel

Figure 8. OpenPose Key Points locations (Zhang et al., 2023).

As shown in Figure 5, from each video frame, the model produced 25 two-dimensional confidence maps, each corresponds to a Key Point on the human body (important joints and edges such as wrists, elbows, nose, and so on; see Figure 8). The dimension of the confidence map is proportional to the original image, and each cell on the map contains a

[13] https://github.com/CMU-Perceptual-Computing-Lab/openpose/blob/master/doc/05_faq.md

value representing the model's level of confidence that the respective Key Point is located at the cell's coordinate. Therefore, to find the location of a Key Point, we identified the coordinate of the highest confidence level on the point's confidence map and projected it back onto the original image. Only the first 9 points were required for calculating affective body pose features, hence the more detailed points were ignored (such as eye, ear, toe, heel, and so on).

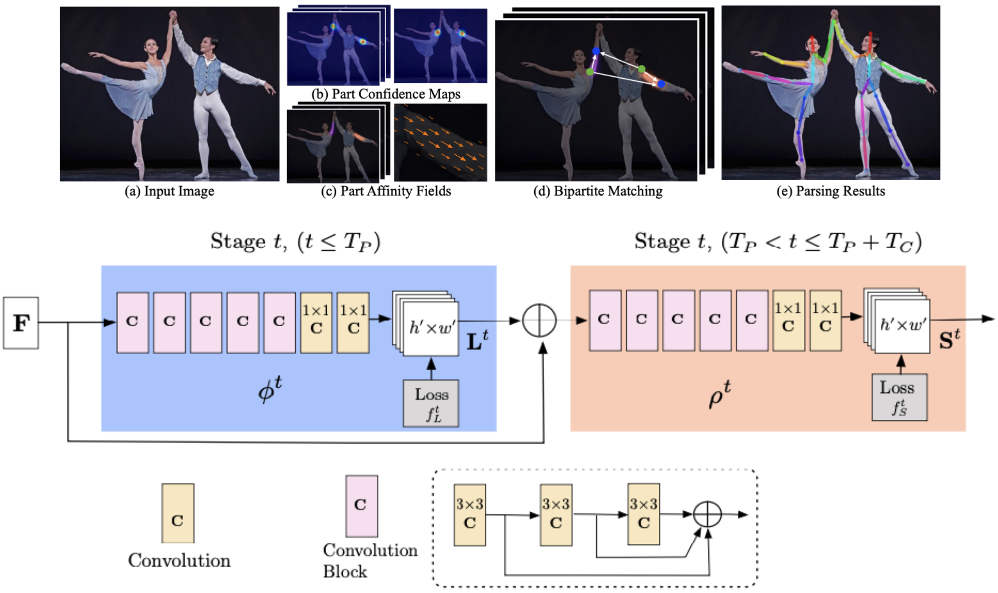


Figure 9. (Above) OpenPose's overall pipeline. (Below) OpenPose's architecture, comprising sets of multi-stage CNN (Cao et al., 2021).

OpenPose's architecture is based on a previous work by Cao et al. (2021). It comprises multiple stages of convolutional neural networks (CNN), largely divided into two stages (Figure 9). A stage is comprised of multiple convolution blocks, each containing 3 layers of convolutions of kernel 3 that are concatenated at their end. An input image is processed by the first 10 layers of the VGG-19 network to produce a set of feature maps F that are the input for the first stage. Then, a series of convolution blocks iteratively predict the Part Affinity Fields (PAFs) which are a series of 2D vector fields that encode the associations between key body parts. After multiple iterations of the first stage, the most updated PAF prediction is concatenated with the original image features (F) and used as input for the second stage. Here, another set of convolution blocks iteratively predict 2D confidence maps, one per body part.

Affective body pose feature calculation

Based on the locations of key points estimated by OpenPose, we derived 15 affective body pose features (listed in Table 4) using the same approach that was used in the study by Jung et al. (2023). These feature set comprises 7 angles, 5 distance ratios, and 3 area ratios (examples depicted in Figure 10).

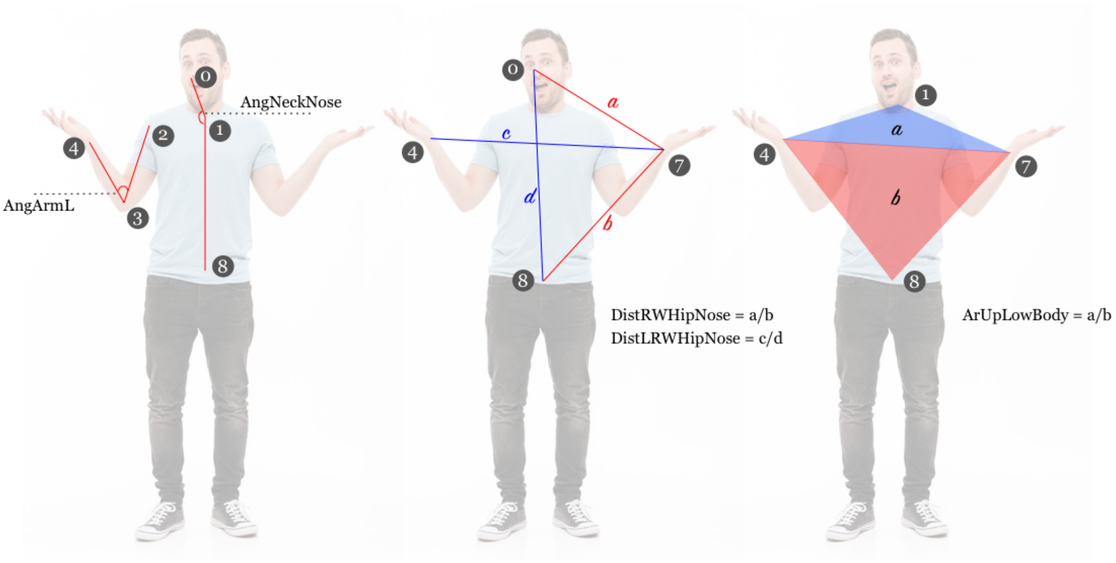


Figure 10. Examples of angular (left), distance ratio (middle), and area ratio (right) affective body pose feature calculations (credit: freepik.com/author/drobotdean).

The selection of these features was informed by prior research on body language and emotion. For instance, Bhattacharya et al. (2021) highlighted the importance of upper body openness in relation to dominance and arousal. Kleinsmith et al. (2011) investigated the connection between emotions and body posture by observing specific joints like head, neck, collar, shoulders, elbows, wrists, torso, hips, and knees. The authors found that torso bending, shoulder rotation, and upper body openness correlate with affective dimensions like arousal, valence, potency, and avoidance. Roether et al. (2009) explored the influence of emotion on joint movement, demonstrating that happiness and anger are associated with larger joint movements, while sadness and fear lead to the opposite effect.

After deriving the affective body pose features at the frame level, we used Tsfresh to generate the summary features, similar to our approach with the facial expression features. From the resulting 11,745 summary features, we then selected 10 statistical characteristics per feature, resulted in a data frame of shape (100, 150).

3.3 Pipeline

To obtain a more comprehensive and reliable assessment of the model's performance, we applied k-fold cross-validation (k = 5) by splitting the data into five subject-independent sub-sets/folds of equal sizes. Before being funneled for model training and evaluation, the data was standardized to have zero mean and unit variance using scikit-learn's Standard Scaler. Subsequently, we trained the models and evaluated the performance across a number of variables: modeling approaches, scenarios (conflict and cooperation), feature modalities, and data folds. Finally, to understand the significant indicators of anxious

Feature Type	Name	Key Points	Description
Angle	AngArmL	2, 3, 4	Upper - lower left arm
	AngArmR	5, 6, 7	Upper - lower right arm
	AngShoNeckL	2, 1, 0	Left shoulder - neck
	AngShoNeckR	5, 1, 0	Right shoulder - neck
	AngShoArmL	1, 2, 3	Left shoulder - upper arm
	AngShoArmR	1, 5, 6	Right shoulder - upper arm
	AngNeckNose	0, 1, 8	Head - torso
Distance ratio	DistLWHipNose	(4, 0) and (4, 8)	Left wrist to head / to mid-hip
	DistLWHipNeck	(4, 1) and (4, 8)	Left wrist to neck / to mid-hip
	DistRWHipNose	(7, 0) and (7, 8)	Right wrist to head / to mid-hip
	DistRWHipNeck	(7, 1) and (7, 8)	Right wrist to neck / to mid-hip
	DistLRWHipNose	(4, 7) and (0, 8)	Between wrists / head to mid-hip
Area ratio	ArUpLowBody	(4, 7, 1) and (4, 7, 8)	Wrists to neck / to mid-hip
	ArWristsNoseHip	(7, 0, 8) and (4, 0, 8)	Left wrist - head - mid-hip / Right wrist - head - mid-hip
	ArWristsNeckHip	(7, 1, 8) and (4, 1, 8)	Left wrist - neck - mid-hip / Right wrist - neck - mid-hip

Table 4. Affective body pose features.

behavior, we used SHAP to analyze the feature relevance of the best-performing model from each scenario and modality.

3.3.1 Data preparation

Each feature set was split into five folds of equal sizes. Considering the class imbalance with the clinically anxious subjects in the minority (approximately 60:40 ratio of healthy or borderline anxious subjects to clinically anxious subjects), the splitting was done in a stratified fashion so that the ratio of the two classes is the same across all folds. Figure11 shows the unimodal model pipeline.

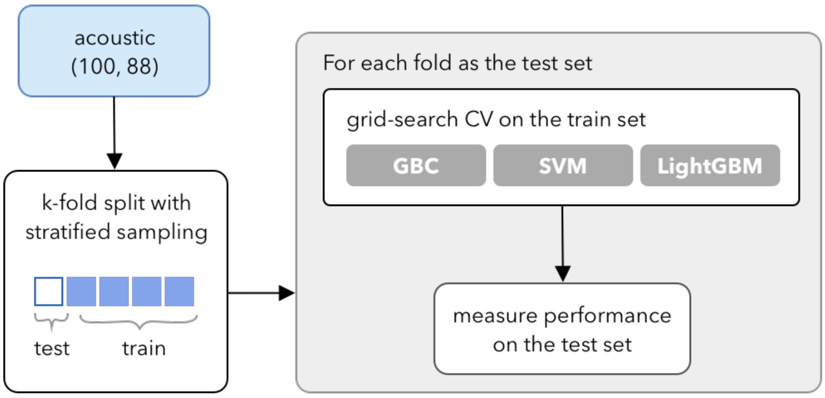


Figure 11. Unimodal model pipeline.

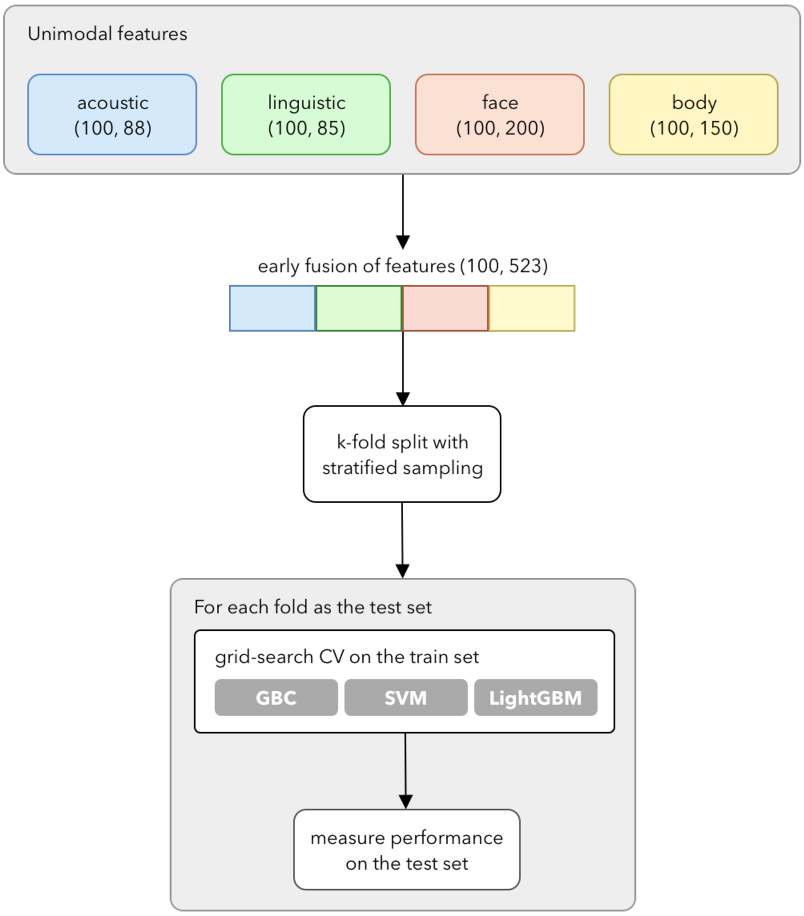


Figure 12. Early fusion model pipeline.

Early and late fusion

In the case of early fusion, feature sets from all modalities were concatenated horizontally, i.e., extending the feature vector of each data point (see Figure 12). Then, the samples were split into five folds using stratified sampling to ensure each fold maintained the same class distribution. A model trained on this combined feature set therefore learned from all modalities, allowing it to consider information from various modalities when making predictions.

In contrast, in the late fusion approach, feature sets from different modalities remained separated and were used to train individual models separately. However, because the predictions from different models were to be fused into a single prediction, it was critical to ensure consistent participant IDs across all classifiers and folds. This means that the participant IDs within the test set of each fold for modality A must be identical to that of the corresponding fold in modality B. This was achieved by seeding the random number generator used by the k-fold stratified sampling function.

Single and combined scenario

The type of interaction was used as an independent variable to examine its impact on the observability of anxiety symptoms. As mentioned in Section 3.1, two types of interactions were reflected through the scenarios: cooperative (planning a vacation together), and conflict (arguing/reprimanding). Accordingly, for each modality and fusion type, three datasets were prepared: one containing only the features from the conflict videos, one from only the cooperative/vacation-planning videos, and one that combined both. We presumed that this approach would allow for a comprehensive analysis of how different interaction atmospheres influence the detectability of anxiety symptoms across various modalities.

To combine the scenarios within a modality, we concatenated the feature frames from the cooperative scenario and the conflict scenario vertically, thus keeping the number of features while doubling the number of samples. This was done to avoid increasing the number of features which could make the model prone to overfitting. However, in retrospect, we believe that this approach was sub-optimal, as the results would later confirm. An alternative approach would be to concatenate feature-wise and applied feature selection and regularization to avoid overfitting.

3.3.2 Model architectures

While there are numerous modeling alternatives to classification problem, we chose to explore three promising options. The first two models are ensembles of decision trees, i.e., Gradient Boosting Classifier (GBC) and Light Gradient Boosting Machine (LightGBM). Tree-based models are seen as highly potential due to their well-established suitability for classification tasks. The third is Support Vector Machine (SVM), which is perhaps one of the most studied modeling algorithm and is widely used as a baseline model due to its relative simplicity and ability to work with small dataset. This section will discuss each approach further.

Support Vector Machine

SVM is a machine learning algorithm that classifies data by calculating the relationships between data points in a higher dimentional space (using a technique called *the kernel trick*) and finding a hyperplane that maximally separates the classes. Not only is SVM able to work with small training data, it also makes a lightweight choice as it stores only the decision boundaries rather than the entire training set. Moreover, SVM is reputed to work well with our dataset's characteristic: a relatively small sample size coupled with a significantly higher number of features compared to samples (especially in the case of early fusion where features from all four modalities are concatenated, bringing the number of features to be around four times the number of samples), also known as the "high dimensionality problem."

We used scikit-learn's implementation of support vector classifier model [14]. The model was optimized through grid-search cross-validation[15] over a set of hyperparameters, the values of which were heuristically approximated from repeated experiments. The hyperparameters optimized for this model were kernel type, degree of the polynomial kernel, kernel coefficient γ, inverse-strength of regularization C, and shrink (whether to try to accelerate training time by shrinking some variables; Chang and Lin (2011)).

Gradient Boosting Classifier

The gradient boosting classifier (GBC) is a variant of the gradient boosting machine (GBM) algorithm applied to classification problem. Since most of our modeling approaches are based on GBM, we will explain the technique more deeply in this section.

GBM is a tree-based algorithm that iteratively builds an ensemble of *weak learners* (often in the form of shallow trees), each attempting to correct the mistakes of the previous learner by training on its errors (termed *pseudo residuals* to distinguish it from linear regression's residual). The algorithm given by Friedman (2002, 2001) is as follow:

```
1. $F_0(\mathbf{x}) = \arg\min_\gamma \sum_{i=1}^{N} \Psi(y_i, \gamma)$
2. For $m = 1$ to $M$ do:
3. $\tilde{y}_{im} = -[\frac{\partial \Psi(y_i, F(\mathbf{x}_i))}{\partial F(\mathbf{x}_i)}]_{F(\mathbf{x})=F_{m-1}(\mathbf{x})}, i = 1, N$
4. $\{R_{lm}\}_1^L = L -$ terminal node $tree(\{\tilde{y}_{im}, \mathbf{x}_i\}_1^N)$
5. $\gamma_{lm} = \arg\min_\gamma \sum_{\mathbf{x}_i \in R_{lm}} \Psi(y_i, F_{m-1}(\mathbf{x}_i) + \gamma)$
6. $F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \nu \cdot \gamma_{lm} 1(\mathbf{x} \in R_{lm})$
7. endFor
```

[14] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[15] https://scikit-learn.org/stable/modules/grid_search.html#exhaustive-grid-search

Ordered explanation of how the algorithm operates is given below:

1. Initialize with a predicted value $F_0(\mathbf{x})$. This is equal to a value ρ that minimizes the sum of losses (also called *pseudo residuals*, as calculated by the loss function L) between each truth value y_i and ρ (i is the index of sample, which goes from 1 to N where N is total number of samples). This value can be found through gradient descent or the sum of the derivatives of the loss function with respect to ρ.
2. Build M number of base learners sequentially.
    (a) For $i \in \{1..N\}$ calculate the pseudo residuals $\tilde{y_i}$. To do this, first take the previous prediction for the sample ($F(x_i) = F_{m-1}(x_i)$), apply it to the loss function L and calculate $\tilde{y_i} = -[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}]$
    (b) Fit a base learner that predicts the residuals.
    (c) Given a set of data points R in a terminal/leaf node l, for each R_{lm} of the new base learner (m), calculate γ_{lm} that minimizes $\sum L(y_i, F_{m-1}(x_i) + \gamma), x_i \in R_{lm}$
    (d) Update the predictions on each data point by following the formula $F_m(x_i) = F_{m-1}(x_i) + \nu \cdot \sum_{l=1}^{m} \gamma_{lm} I(x \in R_{lm})$ where ν is a learning rate that scales γ.
    (e) Repeat (a) to (d) until stopping condition is met (usually in terms of number of trees or reduction in bias).

The derivative operation in step (a) is the *gradient* calculation, referring to the slope of the tangent line on the loss function curve/plane at a specific value of x. The negative sign points in the direction of steepest descent (i.e., gradient descent). Step (a) through (d) is the *boosting* stage where the model sequentially adds a weak learner to 'boost' the performance of the previous learner. For classification problem, a slight modification is applied in step (1) and (a) to convert discrete labels into log-odds and probabilities, hence the residuals become discrepancies between observed and predicted class probabilities.

Thus intuitively, the algorithm iteratively groups the samples based on the difficulty in predicting them (as reflected in the amount of residuals) by fitting a series of weak learners. Then it gradually shrinks the residuals/improve the predictions by minimizing the loss function through gradient descent moderated by learning rate.

One of the benefits of gradient boosting is its built-in regularization effect as a result of using weak learners. A single complex model is prone to overfit the training data and perform poorly on unseen data. However, since each weak learner only captures a small part of the data complexity, it prevents the overall model from becoming overly sensitive to specific features.

Another strength of gradient boosting is its low variance. Because it uses an ensemble of models, the collective prediction averages out the individual errors, leading to lower variance compared to a single complex model. Additionally, its ability to provide feature importance scores makes it more interpretable and aids in feature selection/reduction.

We utilized scikit-learn's implementation of GBC algorithm[16]. The following hyperparameters were optimized: number of estimators/boosting stages, learning rate, maximum depth of an estimator, maximum features to consider during a split, minimum samples required to perform a split, minimum samples required at a leaf node, and the number samples used for fitting an individual base learner.

Light Gradient Boosting Machine

The LightGBM algorithm builds on the previously discussed GBC algorithm by introducing a number of improvements. Besides the capability to leverage parallel processing, built-in regularization, and native support for categorical features, the two major improvements described in (Guolin et al., 2017) are: *Gradient-based One-side Sampling* (GOSS), and *Exclusive Feature Bundling* (EFB).

The idea of GOSS is to reduce the amount of computation required in finding the best split in a node using a histogram of gradients of the instances in that node. Similar to our description for GBC in the previous section, for a gradient to be computed, the model first derives the loss function with respect to each weight variable. Then, a base model with a set of weights is built on the dataset. Next, the gradient is calculated based on the residuals (errors) between the predicted and actual values. The model then gradually finds the optimal weights through gradient descent.

In GOSS, the histogram is built by keeping the instances with the largest gradients (i.e., under-trained instances) and randomly sampling from instances with the smaller gradients. Then, the instances are grouped into bins based on the range of the gradient values. This way, GOSS focuses on instances that contribute more to the loss function and yield more information gain. Moreover, while traditional decision trees evaluate splits at every possible point within the feature range, GOSS reduces the number of potential splits to the number of bins, significantly lowering the computational complexity.

LightGBM also decreases the computational cost by reducing unnecessary computation of zero feature values. Typically, a large number of features results in a sparse high-dimentional feature space. To increase the density, EFB first creates a graph where a vertex represents a feature and a weighted edge represents the amount conflicts between any two

[16] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

features. Conflict is defined as having non-zero values simultaneously. Then, sorting by the degrees, the low conflict features (i.e., vertices with low degree) are merged (bundled) into a new feature. This new feature occupies a value range that is made up of multiple sub-ranges, each belongs to an original feature that composes the bundle. By merging features that do not have non-zero values simultaneously, the new feature effectively encodes the same information as the original features but in a more compact form. Hence, EFB is a nearly lossless approach to not only reduce features but also increase the feature space density.

We employed the Python version of Microsoft's LightGBM library[17]. The optimized hyperparameters were the maximum depth of a base learner, the number of leaves, the minimum number of samples at a leaf node, the learning rate, the number of estimators, and the L1/alpha and L2/lambda regularization terms.

3.3.3 Model training

For each type of modality and interaction scenario, we loaded and prepared the training and testing dataset through concatenation (when applicable) and k-fold stratified sampling, as previously explained. All three models were then fitted on k-1 folds and tested on the remaining fold.

The fitting process involved running a grid-search cross-validation on the training set, optimizing the hyperparameters mentioned in Section 3.3.2. Internally, this process further split the training set into 5 folds and performed cross-validation for each combination of hyperparameter values. After exhaustively testing all hyperparameter combinations, the process returned the optimal parameters configuration as well as an estimator object that had been refitted on the whole dataset. We would then measure the performance of the optimized estimator on the test fold.

The process of fitting and testing on a remaining fold was repeated k-times such that each fold is used as a test fold once. The final performance measures were obtained by calculating the average and standard deviation of each of the performance metrics across all folds.

Late fusion

The late fusion approach is essentially an ensemble of unimodal classifiers. Here, the process of k-fold stratified sampling, model fitting, and predicting on the test fold were executed independently per modality. Their predictions would then be combined (fused) through a voting mechanism using simple majority rule. That is, for each data point, the ensemble would assign the class that was predicted by the highest number of classifiers

[17] https://lightgbm.readthedocs.io/en/stable/Python-Intro.html

(see Figure 13).

The fact that we had an even number of modalities (acoustic, linguistic, facial expression, and body gesture) introduced the possibility of a tie during the fusion phase. In such case, the ensemble was designed to vote for the positive class (the 'anxious' prediction). This decision to prioritize the recall of the positive class was made with the assumption that in the medical context, it is more important for a preliminary diagnostic tool to not miss a possible patient (avoid false negative) than to avoid false positive diagnosis.

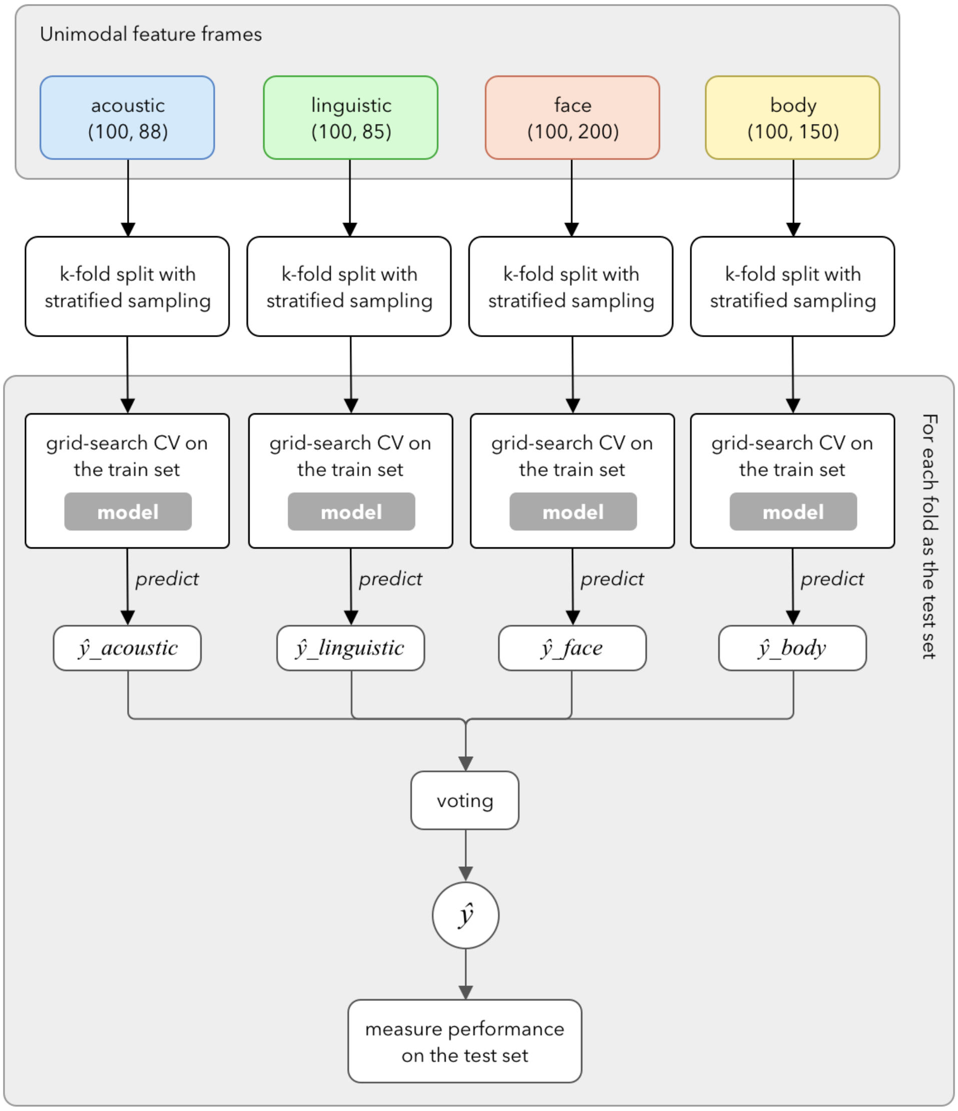


Figure 13. Late fusion model pipeline.

3.3.4 Performance metrics

Performance of the models were measured using a number of performance metrics that we believe to be appropriate for the task at hand. Except *accuracy*, the metrics presented in this section are selected specifically with prevalence and class imbalance in mind, i.e., they are known to be reliable even when applied to a dataset with a skewed class distribution.

Classification accuracy

This is a common way to measure how well a model performs, simply defined as the percentage of correctly classified samples out of all the samples it analyzes. Given a binary classification task and a comparison result between the predicted labels and the ground truth, then accuracy is calculated as Equation 3.4 where *True Positive (TP)* and *True Negative (TN)* are the numbers of samples correctly classified as belonging to the positive and negative classes respectively, while *False Positive (FP)* and *False Negative (FN)* are the numbers of falsely classified observations.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.4}$$

F-1 of the positive class

While accuracy is a popular metric, it is not enough to depict the model's performance, especially in imbalanced datasets where the prevalence must also be considered. To represent both precision and recall in one metric, we calculated their harmonic mean through the F-measure (see Equation 3.5). Harmonic mean is akin to a weighted average with an automatic regularization effect, making it more robust to outliers than arithmetic mean and suitable when working with rates and ratios such as precision and recall. The β variable acts as a weight on the importance of recall in comparison to precision. In our experiments, we used $\beta = 1$ which gave equal importance to both precision and recall. We specifically looked at the positive class because of the hypothetical context of the application of this model (i.e., medical diagnosis) where preliminary identification of potential illness is important.

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall} \tag{3.5}$$

Area under the Receiver Operating Characteristic (ROC AUC)

The Receiver Operating Characteristic (ROC) curve is a graphical plot depicting a model's performance over a series of thresholds that it uses to discriminate between the target classes. The graph consists of two axes, one for True Positive Rate (TPR/sensitivity) and the other for False Positive Rate (FPR or 1 - specificity). Generally, a good model is one that maximizes TPR while minimizes FPR, i.e., maximizes the size of the area under its ROC curve.

Area under the Precision-Recall Curve (PR AUC)

Precision is defined as the ratio of correct predictions to the total number of predictions (Equation 3.6), while recall is the ratio of correct predictions to the actual number of true instances (Equation 3.7). In other words, precision measures how reliable a model's predictions are or the accuracy of the positive predictions, while recall measures the model's ability to identify all relevant instances/the ground truth.

$$Precision = \frac{TP}{TP + FP} \tag{3.6}$$

$$Recall = \frac{TP}{TP + FN} \tag{3.7}$$

Like the ROC curve, the precision-recall curve is a graphical plot that illustrates a model's performance on the precision and recall metrics across various decision thresholds. In scikit-learn, the decision thresholds are acquired from the class probabilities of each prediction (more specifically, the probability of the positive class). Internally, it iterates through each threshold, classifies the data points accordingly, then calculates the precision and recall of the threshold.

A larger area under the precision-recall curve indicates that the model is able to recognize most of the relevant instances while maintaining trustworthiness in its predictions. Conversely, a smaller area under the curve suggests that the model struggles with one or both of these metrics.

Matthew's Correlation Coefficient (MCC)

The Matthews Correlation Coefficient (MCC) is a powerful metric for evaluating the performance of binary classification models. Because it considers the four elements in the confusion matrix (true positive, false positive, true negative, and false negative) in proportion to the size of the classes (see Equation 3.8), MCC is less susceptible to producing overly optimistic results compared to metrics like accuracy or F1-score (Chicco and Jurman, 2020). This makes it particularly robust for handling imbalanced datasets. An MCC score has the range of [-1, 1] where 1 means perfect agreement between the two variables, -1 means perfect disagreement, and 0 means random chance. In our case, the binary variables are the true labels (ground truth) and the predicted labels (model's predictions).

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.8}$$

4. Results

This chapter describes the experimental results in terms of modality type, interaction scenario of the participants, and modeling approach. Overall, the GBC model trained on early-fused multimodal feature set from the conflict scenario achieved the best performance across modalities and scenarios. Performance details are presented in Table 13, 14, and 15.

A number of notable patterns can be observed. First, the best unimodal models that were trained on linguistic and facial expression feature sets performed consistently well across performance metrics. Multimodal features also consistently yielded good performance, and the overall best performance was achieved by a model trained on the early fusion feature set. Second, the Gradient Boosting Classifier (GBC) approach produced most of the best results. Third, all best-performing models were trained on features from the conflict interaction scenario.

4.1 Comparative analysis between modalities

We evaluated the performance of models from each modality using several metrics. Because of class imbalance, we prioritize Matthews Correlation Coefficient (MCC), Area Under the Receiver Operating Characteristic Curve (ROC AUC), and Area Under the Precision-Recall Curve (PR AUC). The results are visualized in Figure 14 and detailed in Table 5.

The performances were also compared to a baseline model that always predicts the majority class (class 0, healthy). However, it should be noted that the baseline model's F1-positive and MCC scores were not usable for benchmarking since they were inherently zero (the model only predicts one class, hence precision and recall of the unpredicted class are zero). Additionally, the baseline PR AUC of 0.71 was unusable since a simple majority-predicting model can produce a high PR AUC in an imbalanced dataset due to the way the curve is drawn. That is, the first point is always assumed at a baseline level where only the positive class is predicted, thus precision = proportion of the positive class and recall = 1.0. The last point of the curve is always at precision = 1.0 and recall = 0.0. Due to these assumptions, the area under the curve of a simple baseline model can reach a moderate size, depending on the proportion of the classes.

Among unimodal models, linguistic and facial expression (with continuous Action Unit values) performed consistently well across most metrics. The acoustic model had the highest PR AUC (0.63), the linguistic model achieved the highest MCC (0.3), and the facial expression model (with continuous AUs) reached the highest accuracy (0.66, exceeding the baseline by 7%).

Modality	**Model**	**Accuracy**	**F1-positive**	**ROC AUC**	**PR AUC**	**MCC**
Acoustic	GBC	0.63	0.43	0.6	0.63	0.21
Linguistic	LightGBM	0.63	0.51	0.6	0.6	0.3
Facial Expression (cont. AUs)	GBC	0.66	0.53	0.63	0.55	0.28
Facial expression (bin. AUs)	SVM	0.64	0.4	0.59	0.52	0.22
Body pose	GBC	0.61	0.45	0.58	0.6	0.2
Early fusion (cont. AUs)	GBC	0.66	0.56	0.64	0.63	0.29
Early fusion (bin. AUs)	LightGBM	0.6	0.49	0.58	0.62	0.17
Late fusion (cont. AUs)	Ensemble of GBC	0.66	0.54	0.64	0.59	0.28
Late fusion (bin. AUs)	Ensemble of LightGBM	0.63	0.5	0.62	0.55	0.25
Baseline		0.59	0	0.5	0.71	0

Table 5. Performance of the best models of each modality. All of these models trained on features from the conflict scenario. Values displayed here are averages across five folds.

Similar trends can be observed with early-fusion and late-fusion models. The late-fusion model (with continuous facial AUs) scored the highest accuracy and ROC AUC (0.64, which was 14% above baseline). Notably, the early-fusion model (also using continuous facial AUs) produced the highest measurements in multiple metrics: accuracy, F1-positive (0.56), ROC AUC, PR AUC, and MCC (despite not the highest, but was only 1% below the highest score).

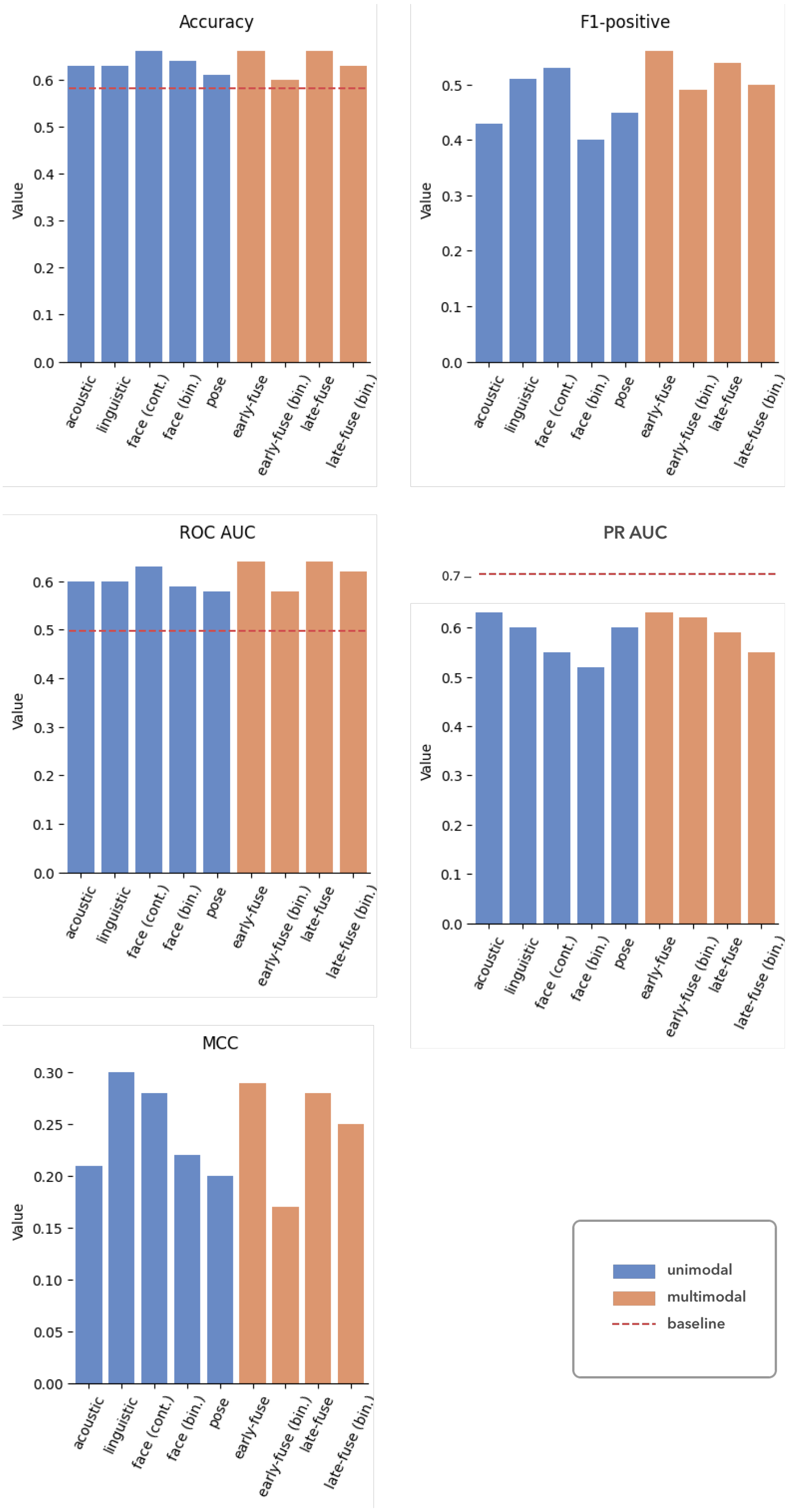


Figure 14. Best performance from various modalities. The models performed slightly above baseline level, suggesting moderate performance. Note: These values are averages across five folds. Also, baselines are unusable for F1-score, MCC, and PR AUC since the model only predicted one class.

Overall, we observed moderate performance from our models. The best accuracy and ROC AUC are only marginally above the baseline. F1-positive scores were between 0.4 and 0.56, PR AUC ranged from 0.52 to 0.63, and MCC ranged from 0.17 to 0.3. These results indicate that the models only slightly surpassed chance level, suggesting room for improvement.

Interestingly, regarding modeling approach, 5 out of the 9 models were GBCs, 3 were LightGBMs, and only 1 was an SVM. Moreover, across all modalities, the best performances were produced by models trained on data from conflict interaction.

McNemar's test was applied to compare the predictions of the best unimodal model (LightGBM model trained on linguistic features) and multimodal model (GBC model trained on early-fused multimodal features). The test hypothesizes that the marginal frequencies of the row and column of the contingency table (see Table 22) are equal, meaning the two binary variables (in this case, the predictions of the two models) are not significantly different from each other. Our result ($p = .03, \alpha = .05$) indicates that there is a statistically significant difference in predictions between the best unimodal and multimodal classifiers. This observation supports the notion that learning multimodal features can improve the model's performance.

4.2 Comparative analysis between scenarios

This section compares the performances of the best model of each scenario (detailed in Table 6 and visualized in Figure 15). Evidently, the feature sets from conflict interaction consistently yielded the highest performance across all metrics and modalities. Cooperation and combined scenarios exhibited comparable performance in most metrics except for F1 metric, in which the cooperation scenario shows a significant drop compared to the other two. Moreover, while the gaps between the first and second best in the other metrics are between 5-7% , MCC observed a 9% gap.

Interestingly, in terms of modeling approach, two out of the three classifiers in Table 6 are GBC models trained on multimodal early-fusion features. Further, when controlled for model type and training features, the cooperation scenario occasionally produced higher results than the conflict scenario. Examples of this are the LightGBM model trained on facial expression data with binary Action Unit values (cooperation: 0.21 MCC, 0.63 accuracy; conflict: 0.02 MCC, 0.54 accuracy) and the late-fusion SVM model trained on multimodal features (cooperation: 0.15 MCC, 0.62 accuracy; conflict: -0.04 MCC, 0.53 accuracy). However, this type of results rarely occurred.

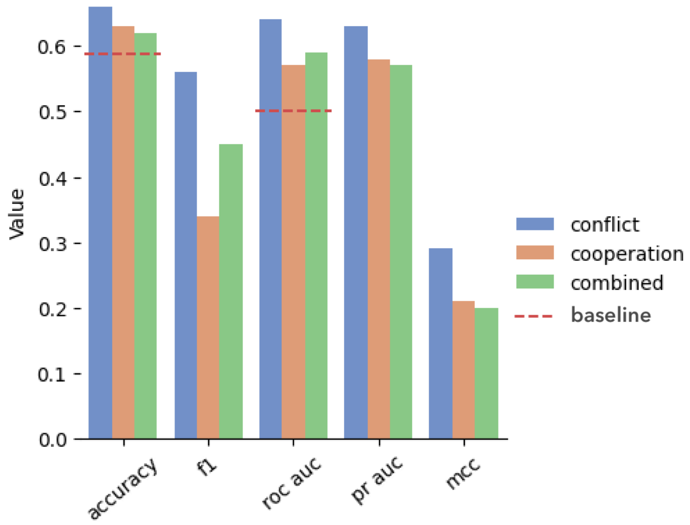


Figure 15. Best performance from conflict, cooperation, and combined scenarios. Anxiety symptoms are better detected during the conflict scenario.

To better understand the behavior of the models (conflict: M_{con}, cooperation: M_{cop}, and combined: M_{both}), we examined their confusion matrices by summing the corresponding entries across folds (details are available in Table 16, 17, and 18). Our analysis revealed that M_{con} excelled in both positive class recall (0.54) and negative class precision (0.7) compared to M_{cop} (0.24 recall, 0.63 precision) and M_{both} (0.37 recall, 0.65 precision). It is apparent that the classifier prioritizes true positives (individuals needing help) and minimizes false negatives (missing those who need support). This prioritization aligns with our hypothetical context (i.e., detecting mental health issues), where failing to identify someone who needs help may have a greater cost than misidentifying a healthy individual.

Conversely, M_{cop} and M_{both} tend to overestimate the number of negative samples, with M_{cop} producing the highest negative class recall (0.9) and predicted the least number of positive samples (16 predictions, compared to 26.5 from M_{both} (normalized) and 37 from M_{con}). This behavior might reflect the peaceful and cooperative interaction exhibited in the model's training data, which led to its tendency to predict the negative class.

The result of McNemar's test ($\alpha = .05$) confirms that there is indeed a statistically significant difference between the output of M_{con} and M_{cop} (see Table 7). However, M_{both} results are in between M_{con} and M_{cop}, hence it is possible that there is no significant difference between M_{both} and M_{cop} and also M_{both} and M_{con}. Whereas it is good to find difference between M_{cop} and M_{con}.

Scenario	Modality/model	Accuracy	F1-positive	ROC AUC	PR AUC	MCC
Conflict	Multimodal early fusion (cont. AUs) / GBC	0.66	0.56	0.64	0.63	0.29
Cooperation	Facial expression (bin. AUs) / LightGBM	0.63	0.34	0.57	0.58	0.21
Combined	Multimodal early fusion (cont. AUs) / GBC	0.62	0.45	0.59	0.57	0.2
Baseline		0.59	0	0.5	0.71	0

Table 6. Performance of the best models of each scenario. Values displayed here are averages across five folds.

Comparison	p-value
Conflict v. cooperation	0.001
Combined v. conflict	1.0
Combined v. cooperation	3.4

Table 7. Results of McNemar's test between scenarios ($\alpha = .05$). The contingency tables are available in Table 19, 20, and 21

4.3 Important predictors

To identify the most important indicators of anxiety symptoms, we employed SHAP (Lundberg and Lee, 2017) analysis on the top performing model of each modality. SHAP is a model-agnostic method of explanation based on game theory. It works by building *explanation models* to approximate the predictions of the original model. Then, it uses *additive feature attribution methods* where a feature's relative importance to a model is measured by comparing the model's prediction with and without the presence of the said feature. Note that the graphs in this section are in different scales, thus comparisons between different graphs should not be made based on the length of the bars.

4.3.1 Acoustic features

Figure 16 shows the most important features used by the model (definitions of the top five are given in Table 8). The first feature, *spectralFlux_sma3_stddevNorm*, refers to the fluctuations/change over time in the distribution of sound energy/amplitude across spectrum of frequencies. While the presence of this feature in most cases contributed towards a negative prediction, there were some cases in which the lower the spectral flux feature value, the higher the probability of presence of anxiety symptoms (SHAP value increases).

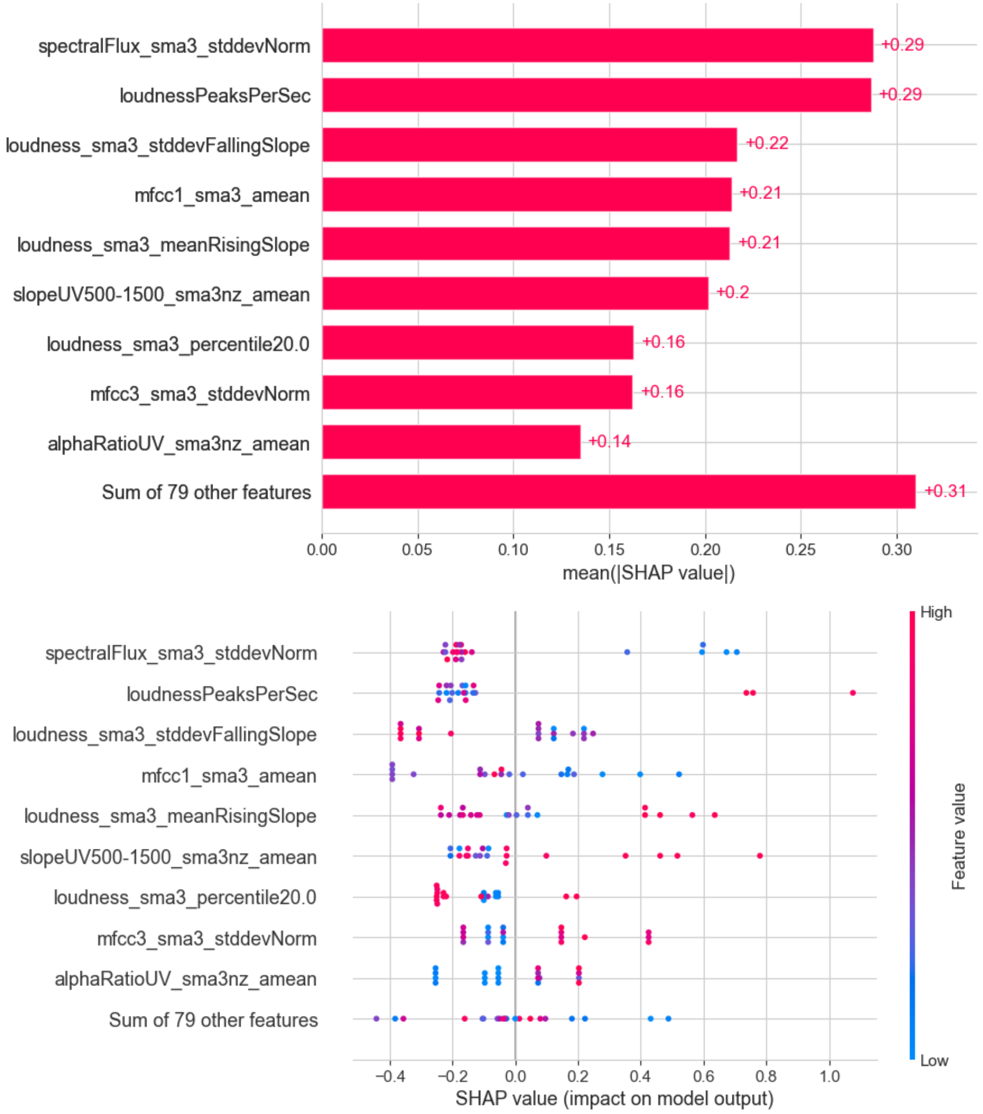


Figure 16. Important features of the GBC model trained on acoustic features from the conflict scenario.

The fourth feature, *mfcc1_sma3_amean*, is related to the first coefficient in the series of Mel-Frequency Cepstral Coefficients (MFCC). Mel-Frequency itself is a transformation technique to warp sound frequencies using the Mel-scale which approximates human auditory perception.[1] The first few coefficients (including MFCC-1 and MFCC-3) tend to capture the overall shape of the spectrum, providing information about the dominant frequencies and the distribution of energy. The impact of this feature varied greatly, with the absolute mean of MFCC-1 inversely related to the prediction of the positive class (i.e., the presence of anxiety symptoms), and standard deviation of MFCC-3 directly related to the positive class.

[1] https://www.sciencedirect.com/topics/computer-science/cepstral-coefficient

Feature name	Description
spectralFlux_sma3_stddevNorm	Normalized standard deviation of the difference of the spectra of two consecutive frames.
loudnessPeaksPerSec	Number of peaks in signal intensity per second.
loudness_sma3_ stddevFallingSlope	Standard deviation of the slope of the falling parts of the loudness signal. Unfortunately we could not find a precise documentation of this feature, however it might relate to the skewness of the spectral distribution calculated in third-order moment.
mfcc1_sma3_amean	Arithmetic mean of the encoded vocal timbre represented using 1 Mel-Frequency Cepstral Coefficients (MFCC).
loudness_sma3_meanRisingSlope	Mean of the slope of the rising parts of the loudness signal.

Table 8. Top 5 features of the model that was trained on acoustic features.

Furthermore, three of the five most important features (*loudnessPeaksPerSec, loudness_sma3_stddevFallingSlope,* and *loudness_sma3_meanRisingSlope*) are related to the loudness of voice (the amplitude of the signal). Although the effects were also mixed, generally, high feature values seemed to indicate the absence of anxiety symptom.

Taken together, this suggests that the distribution of energy across frequencies as well as the size of the amplitudes are the most influential factors to the model's predictions. Looking at the impact direction, it is probable that the suspected anxious children in our sample tended to be quieter and toneless, although it worth noting that we mostly observed mixed effects where a feature's impact varies depending on the values of other features. This is to be expected, as feature interactions and non-linearity can cause different features to be weighed differently between subjects.

Looking at the literature, our finding partially agrees with a previous study by Hinduja et al. (2024). The study found that loudness of voice was the second most important acoustic feature and that it was negatively correlated with the presence of symptoms of obsessive compulsive disorder (OCD). Although the study was about assessment of severity of depression and OCD, it is known that anxiety disorder often comorbid with other mental disorders (such as depression and OCD, as explained in Chapter 2).

4.3.2 Linguistic features

Figure 17 and Table 9 present the most important indicators used by the best-performing model that was trained on the linguistic feature set (conflict scenario). Pronounced clusterings of feature effects were shown by five out of the six most important features. The low usage of 3rd-person plural pronouns (LIWC: *they*), which might indicate a more self-focused speech, contributed to the positive prediction. The low usage of filler sounds indicative of non-fluency (*nonflu*), which could hint at a more reserved or quiet subject, also contributed positively. Furthermore, frequent usage of words emphasizing negative emotions (*negemo*), anxiety (*anx*) and focusing on past events (*focuspast*) tended to correlate positively with the prediction of anxiety symptoms. Additionally, the use of movement-related words (*motion*) also significantly influenced the model's predictions. Although the direction of the impact was generally mixed, there was a slight tendency towards a negative correlation. This could be interpreted as a tendency to be less expressive, similar to the effect of *nonflu*.

Feature name	**Description**
they	The use of 3rd person plural nouns (e.g., they, their, them).
nonflu	Frequent use of filler sounds/words (e.g., oh, uh, um) indicating nonfluency.
negemo	The use of words indicating negative tone (e.g., bad, wrong, too much, hate), previously labeled as $tone\_neg$.
motion	The use of words to perceive movements (e.g., go, come, went, came).
anx	The use of words to indicate anxiety (e.g., worry, fear, afraid, nervous), also labeled as emo_anx.
focuspast	The use of past-oriented words.

Table 9. Top features of the model trained on linguistic features (Boyd et al., 2022).

These findings find some support in the literature. For instance, a previous study by O'Dea et al. (2021) reported that the frequent use of 3rd-person plural pronouns and filler sounds indicating non-fluency were negatively correlated with depression ($\rho = -0.34$ and -0.67, respectively)[2], while the use of 1st-person singular pronouns was positively correlated ($\rho = 0.35$) with the prediction of anxiety (notably, depression and anxiety are often comorbid). Likewise, a study by Al-Mosaiwi and Johnstone (2018) found that individuals suffering from anxiety and depression tend to utter more words from LIWC's *negemo* and *affect* dictionaries, which includes the *anx* dictionary as a subset (Boyd et al., 2022).

Psychological literature tend to agree on the relationship between a subject's mental

[2] https://journals.plos.org/plosone/article/figure?id=10.1371/journal.pone.0251787.t003

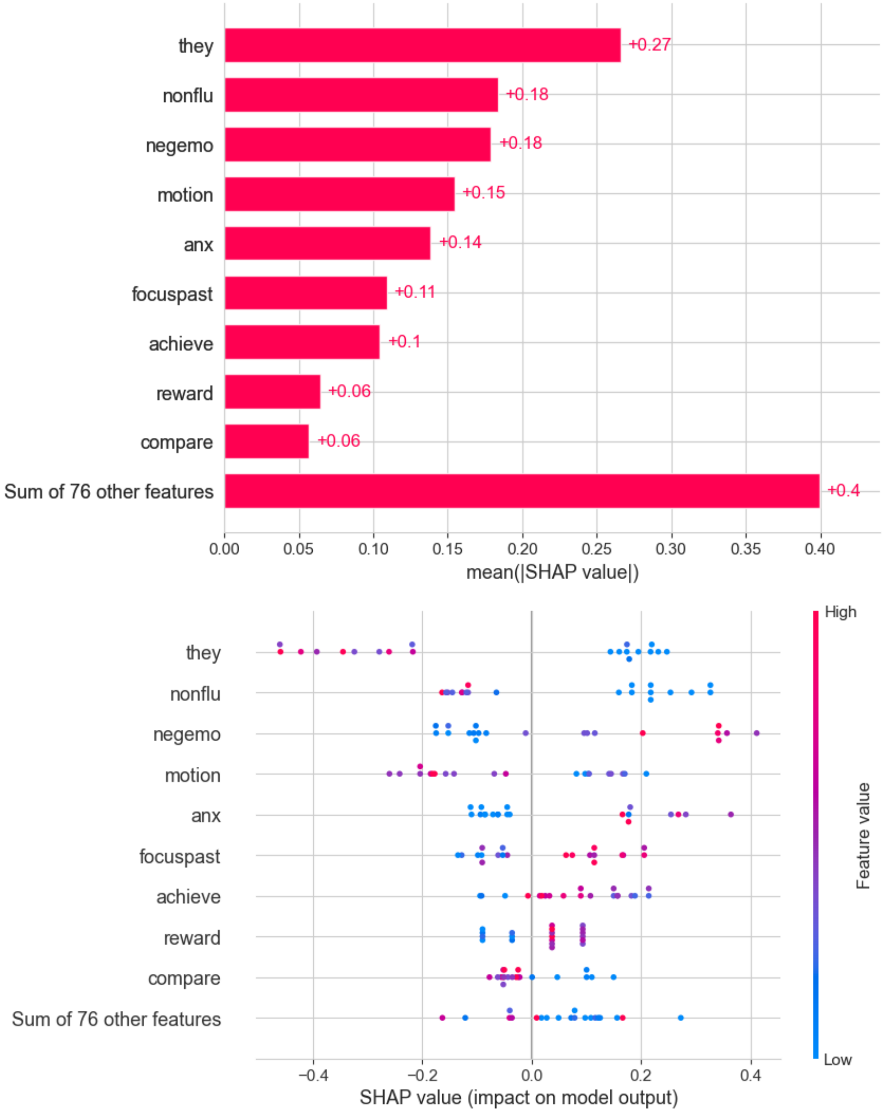


Figure 17. Important features of the LightGBM model trained on verbal features from the conflict scenario.

state/health and their temporal orientation and their is rather well established. For example, Papageorgiou and Wells (1999); Hong (2007) considered anxiety to be linked with worrying about future events, while ruminating is associated with depression about past events. In chapter four of Sirois (2016) the author explores the link between stress, *temporal myopia* (essentially, short-term thinking), and the development of chronic diseases.

Interestingly, although we could not find existing literature linking the use of motion-related words to anxiety, the data show that the two variables were negatively correlated. We speculate that this behavior might stem from the nature of the discussion topic and the fact that the subjects were young children, who may be more inclined to speak in terms of concrete bodily actions and movements rather than abstract concepts (we believe this behavior is rather well-established in the field of psychology and child development, for example see Kaminski et al. (2006)).

To conclude, our analysis suggests that verbal symptoms of anxiety may include placing more attention on the self and less towards others. The subject may also explicitly express his/her anxiety through frequent usage of words related to the negative emotions (examples from Boyd et al. (2022) include: bad, hate, hurt, tired) and anxiety/fear (e.g., worry, fear, afraid, nervous).

4.3.3 Facial expressions features

Figure 18 and Table 10 present the most important facial features used by unimodal facial expression model. Three features showed pronounced positive correlation with the prediction of anxiety (namely, *AU11_mean_abs_change, AU43_standard_deviation,* and *AU43_median*) and four showed the opposite effect (*AU23_maximum, AU10_maximum, AU05_median,* and *AU07_mean*). Interestingly, the two features related to AU 04 (*AU04_skewness, AU04_standard_deviation*) showed mixed effects with inclination towards predicting the negative class.

The negative effect of high *AU23_maximum* (tightening of the lip) and *AU10_maximum* (raising of the upper lip) as well as the negative impact of *AU04_skewness* and *AU04_standard_deviation* (raising and drawing together of eyebrows) may suggest that our less anxious participants were more expressive/displayed more vivid facial expressions. Conversely, the positive correlation of *AU11_mean_abs_change* (nasolabial deepener; lifting and stretching of the upper lip, forming an expression of fear or disgust similar to AU10 only more subtle), *AU43_standard_deviation* and *AU43_median* (closing of the eyelids) suggest that the more anxious participants tended to make subtle expressions of fear through their upper lips and to close their eyes more often.

These findings agree with a significant number of literature. Our first 3 action units (AU 23, 11, 10) are related to the mouth region, while the 4 subsequent units (AU 43, 04, 05, 07) are related to the eyes region. Accordingly, studies have found that anxiety is primarily driven by/related to the fear emotion (Harrigan and O'Connell, 1996; Pediaditis et al., 2015; Giannakakis et al., 2017) which is expressed on the face mostly through lip movements and deformations (such as stretching or tightening) (Metaxas et al., 2004; Pediaditis et al., 2015) as well as increased blink rate and raising/drawing the eyebrows together.

In Facial Action Coding System (FACS), anxiety is most strongly predicted by the activation of AU 01, 02, and 04, and additionally AU 05, 07, 20, 25, and 26 (Arellano et al., 2014). Similarly, the summary statistics of AU 04, 05, and 07 are in our top ten most important indicators. However, while most activations of AU 05 and 07 contributed to the

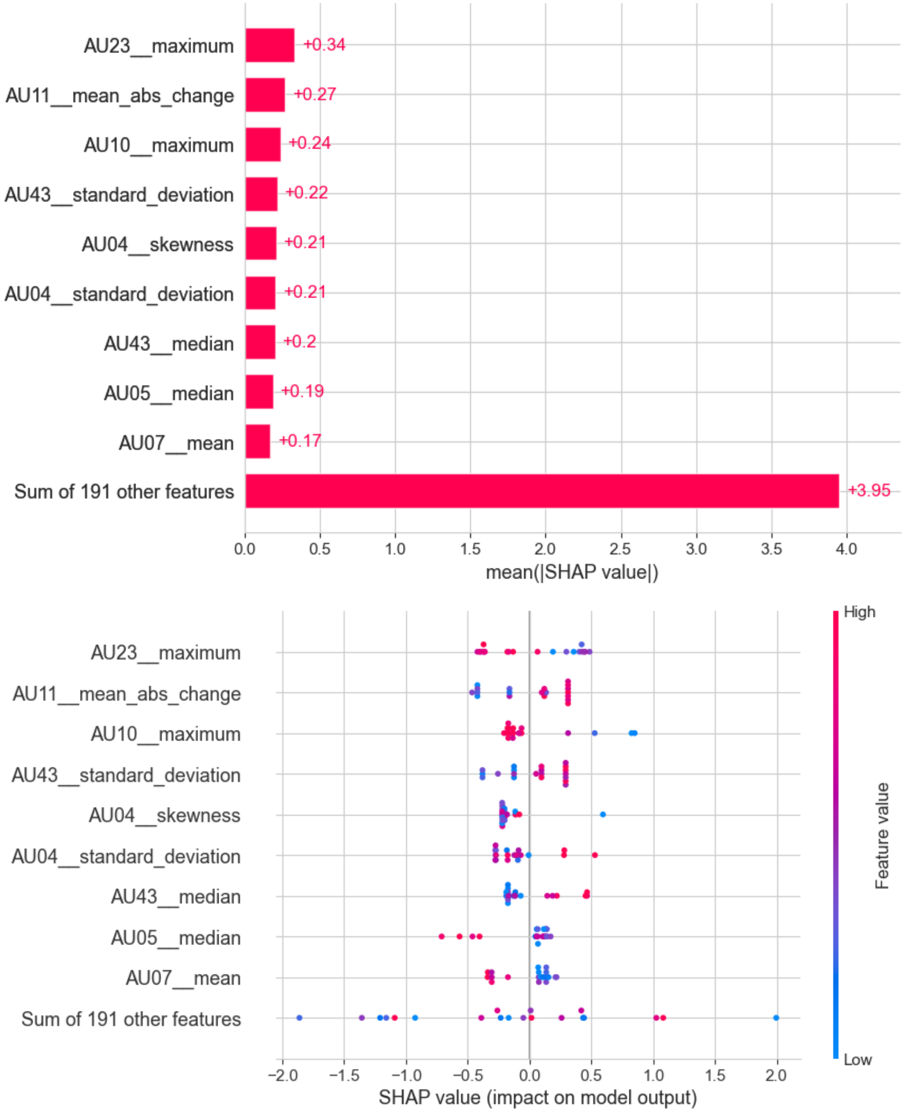


Figure 18. Important features of the GCB model trained on acoustic features from the conflict scenario.

positive prediction, all activations that had the opposite effect were of high probabilities. Also, the activation of AU 04 tended to have a negative correlation with the presence of anxiety symptom. This shows some level of disagreement with some prominent mappings between FACS and emotion.[3][4]

In summary, our observations suggest that anxiety is expressed mostly through the eyes and mouth regions, with the mouth being the most important indicator. Our observation suggests that anxious people display less facial expressions and close their eyes more frequently, likely in an attempt to decrease the amount of incoming stimuli.

[3] https://imotions.com/blog/learning/research-fundamentals/facial-action-coding-system/

[4] https://www.eiagroup.com/resources/facial-expressions/facial-action-coding-system-facs/

Feature name	Description
AU04	Eyebrow lowerer (*Depressor glabellae, Depressor supercilli*, and *Currugator* muscles).
AU05	Eyelids raiser (*Levator palpebrae superioris* muscle).
AU07	Eyelids tightener (*Orbicularis oculi* and *pars palpebralis* muscles).
AU10	Upper lip raiser (*Levator labii superioris* and *Caput infraorbitalis* muscles).
AU11	Nasolabial deepener (*Zygomatic minor* muscle).
AU23	Lip tightener (*Orbicularis oris* muscle).
AU43	Eyelids closing (relaxation of the *Levator palpebrae superioris* muscle).

Table 10. Top 5 features of the model trained on facial expression features.

4.3.4 Affective body pose features

Figure 19 and Table 11 show the most impactful body pose features. Although we were not able to find an direct support linking the exact same feature set to anxiety emotion in the literature, a number of studies have suggested that fidgeting, frequent and regular hand movements, as well as closed body posture and contracted upper body are among the gestures most linked to negative emotions.

Feature name	Description
AngArmL	Angle formed between the upper and lower left arm.
AngShoArmR	Angle formed between the right shoulder and upper arm.
AngShoNeckL	Angle between the left shoulder and neck.
AngNeckNose	Angle between the head and torso.
DistRWHipNose	Distance ratio between the right wrist to nose and to hip.
DistLRWHipNose	Ratio of distances from left to right wrist and from nose to hip.

Table 11. Top 5 features of the model trained on body pose features.

We can see the gestures that our model considered in its predictions. For instance, head movements indicated by the high variance of the angle between the head and torso (*AngNeckNose_variance*), and hand movements indicated by the angle formed by the left elbow (*AngArmL_variance*) as well as the platykurtic distribution (low tailedness/less outliers) of distance ratio between the right wrist to hip and to nose (*DistRWHipNose_kurtosis*)

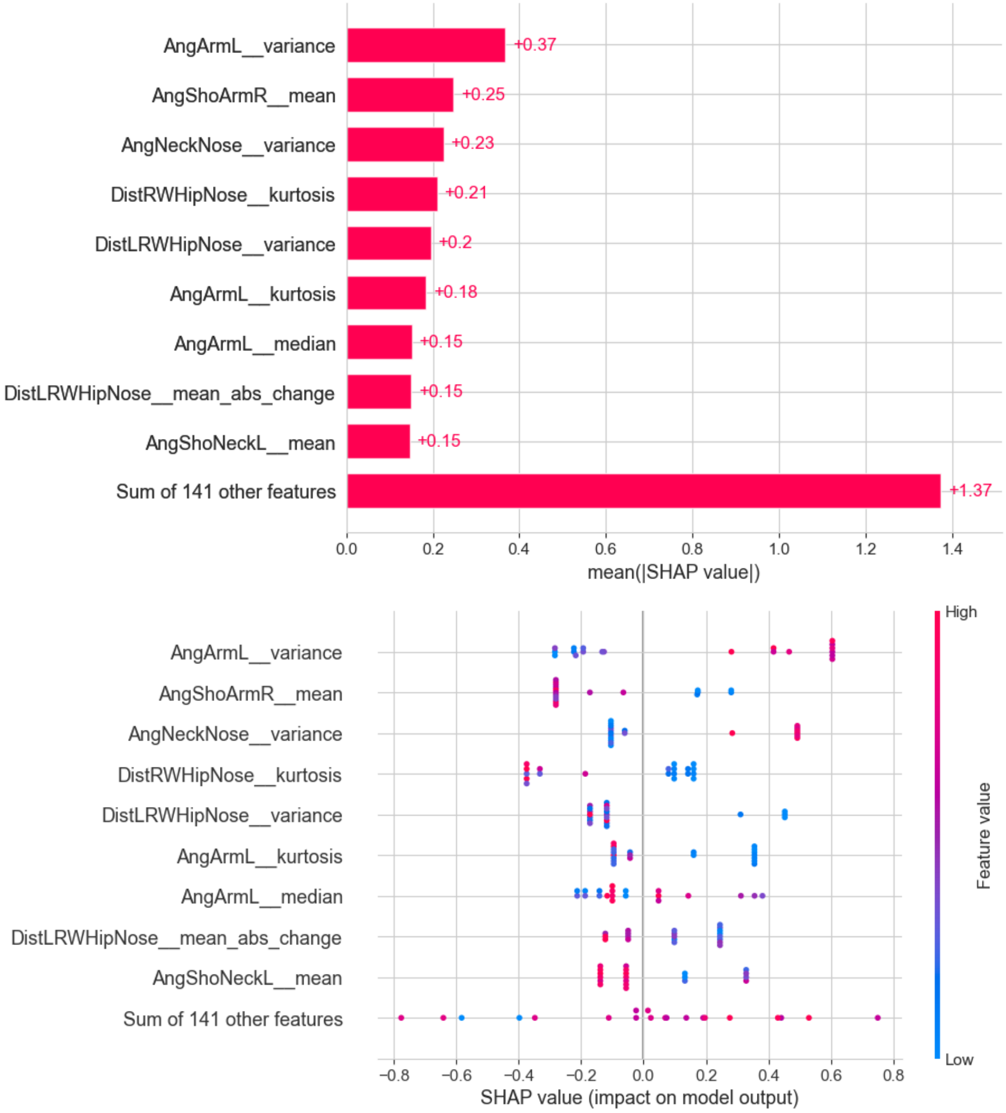


Figure 19. Important features of the GCB model trained on acoustic features from the conflict scenario.

had positively impacted the prediction of the positive class. And although the size and distribution of the left arm's angle (*AngArmL_median* and *AngArmL_kurtosis*) exhibited a mixed effect, there was a tendency for higher median and platykurtic distribution to influence positive prediction. These feature effects suggest that the model's predictions were based on constant head movements and arm movements from the elbow, which aligns with the observations made by Lin et al. (2023) in which constant hand movements and head fidgeting are indicative of anxiety. Roether et al. (2009) also found that the flexion of limbs, head inclination, and body posture are strong predictors of positive affect (e.g., happy, confident) and negative affect (e.g., fearful, anxious, sad).

The model also considered a closed and contracted upper body as an important indicator of anxiety symptom. The small angle between the right shoulder and upper arm

(*AngShoArmR_mean*) and between the left shoulder and neck (*AngShoNeckL_mean*) contributed mostly towards the positive class. The low variance and change in the distance ratio of the wrists to the body's vertical axis (*DistLRWHipNose_variance* and *DistLRWHipNose_mean_abs_change*), suggesting that the posture remained relatively constant (which could also be interpreted as tense) and exhibited less changes in hand positions in the direction of opening/closing of the upper body, also influenced positive prediction. The model's behavior in considering body openness is supported in the literature. For instance, Castillo and Neff (2019); Karg et al. (2013) associate closed upper body posture with the fear emotion.

In conclusion, these observations suggest that non-anxious subjects displayed more open body postures and moved their arms more frequently in the direction of opening/closing the upper body. In contrast, more anxious subjects exhibited more fidgeting movements with their heads and fewer side-to-side arm movements, resulting in a more closed posture.

4.3.5 Combined modality

Figure 20 and Table 12 present the most important predictors used by the highest-performing early-fusion model. Overall, the model seemed to have relied mostly on features related to body gestures, facial expressions and words use.

The high standard deviation of the angle formed by the left arm (indicated by *AngArmL_standard_deviation*) coupled with the platykurtic distribution (low tailedness/less outliers) of the size of the angle (indicated by *AngArmL_kurtosis*) can be interpreted as constant hands movements. Their correlation with anxiety is supported in the work of Lin et al. (2023) where constant hand movements, fidgeting, and more body movements overall are associated with anxiety. Furthermore, the positive skew (mode less than median) of the angle between the right shoulder and arm (indicated by high *AngShoArmR_skewness*) can be interpreted as a closed upper body posture, which is associated with tenseness and fear (Castillo and Neff, 2019; Karg et al., 2013).

In our result, high probability of the activation of the chin raiser muscle (*Mentalis*; as indicated by *AU17__mean_abs_change*) tended to encourage predictions of the positive class. Although we could not find a direct link between the chin muscle and anxiety symptom in the literature, we speculate that the muscle might have altered the shape of the mouth which then indirectly related to anxiety symptom.

The link between ruminations on past events and anxiety and depression is relatively well established, for example in studies such as Hong (2007) and Hilt and Pollak (2012). This supports our finding where the frequent usage of past-oriented words (indicated by

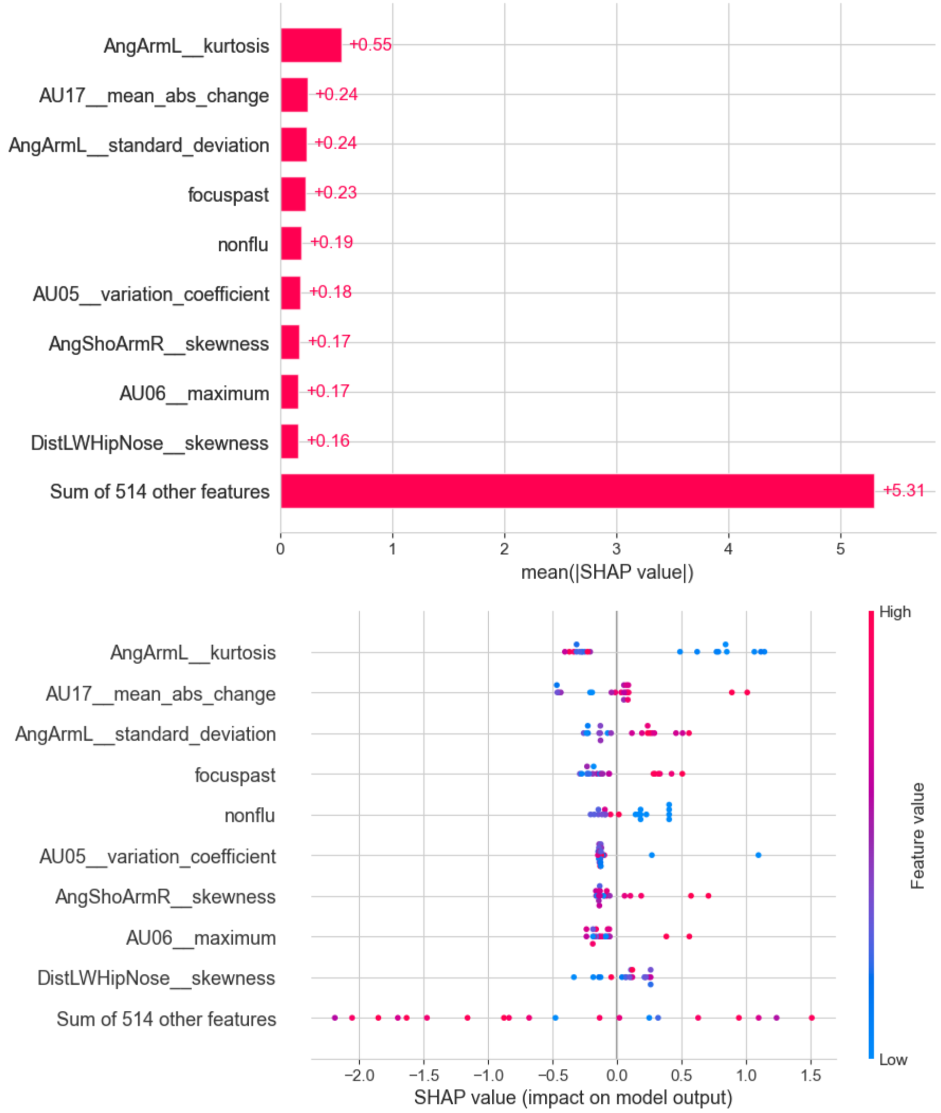


Figure 20. Important features of the early fusion GCB model trained on conflict data.

focuspast) is positively correlated with anxiety prediction.

We could not find a literature that supports the inverse correlation between non-fluency and anxiety. One may speculate that the inverse correlation shown in our result might have stemmed from the participants' tendencies to be quiet and speak less when anxious (similar to the observation in Section 4.3.1). The same behavior was observed in a study by Holleman et al. (2021). Here, children spoke more in the cooperation-scenario whereas parents spoke more in the conflict-scenario.

Another interesting observation is that while some features were deemed most important in unimodal contexts (for example, *they, nonflu*, and *negemo* were most important for the linguistic model), their ranking within the multimodal model decreased. This shift in importance can be attributed to the interplay between features from different modalities

Feature name	**Description**
AngArmL	Angle formed by the upper and lower left arm.
AngShoArmR	Angle between the right shoulder and upper arm.
AU17	Chin raiser (*Mentalis* muscle).
focuspast	Frequent use of past-oriented words (e.g., was, had, were, been).
nonflu	Frequent use of filler sounds/words (e.g., oh, uh, um) indicating nonfluency.

Table 12. Top 5 features of the model trained on dataset from multiple modality.

and the specific workings of our tree-based models. During the boosting process, the training set progressively undergoes splits based on the most informative features. In this case, by a certain iteration, the training data might have already been split in a way that rendered the previously most important linguistic features less explanatory. In other words, splitting based on these features offered less information gain compared to splitting based on features like *focuspast*.

5. Discussion

This chapter discusses the observations derived from the results shown in the previous chapter, utilizing these findings to answer the research questions. It compares unimodal and multimodal features, contrasts the results from conflict and cooperation scenarios, and analyze feature importance to deduce the symptoms of anxiety.

5.1 Better context through multimodality

Several studies support the notion that learning from multimodal feature sets lead to better performance compared to learning from a single modality. Bilalpur et al. (2023), for example, combined 143 features capturing summary statistics of head and face attributes (Face and Head Dynamics/FHD) extracted with AFAR toolbox (Ertugrul et al., 2020, 2019), face Action Units (AU), speech behavior, and verbal features extracted with LIWC. The models trained on this combined feature set achieved higher performance than models that trained with only one of the modalities. Likewise, Hinduja et al. (2024) employed the same feature set with the addition of acoustic features (eGeMAPS and COVAREP). Despite observing a mixed effect on OCD severity prediction, their multimodal models produced the highest performance in depression severity prediction and TEED prediction. A study by Yoon et al. (2022) showcased that their multimodal fusion model—particularly one that used multiply fusion operation—yielded higher precision, recall, and F1 scores than unimodal baseline models, second only to their proposed *Depression Detector* architecture. In a different study by Pranesh (2022), a model that utilizes a series of attention-based fusion techniques on top of pre-trained unimodal models performed better than any of the underlying models.

We posit that the superior performance was the result of the richer and more comprehensive representation produced by multimodal feature sets. Features from one modality might at times be ambiguous when considered alone. However, combining them with features from other modalities can provide additional context and information that can help resolve ambiguities, leading to better predictions. Also, by leveraging multiple sources, the model can potentially compensate for deficiencies in any single modality, making it less susceptible to noise or missing data in any single modality.

5.2 Interaction scenarios and anxiety expressions

The results presented in Section 4.2 points at a difference in expressiveness inherent to the context/theme of the interaction. Our results indicate that anxiety symptoms tend to become more observable in an atmosphere of conflict and turbulence compared to normal or positive ones.

This observed relationship between anxiety and stimuli presented in the environment is similar with findings from a number of previous studies. For instance, an experiment by Mogg et al. (2007) compared the gaze reactions of high- and low-anxious individuals when presented with neutral, angry, fearful, and happy faces. In this study, the authors reported the role of anxiety in producing heightened attention to threat-related cues. High-anxious individuals were more likely to direct gaze at intense negative facial expressions, than low-anxious individuals. However, differences in the duration of attention were noticeable, where fearful expression tended to hold longer attention than angry expression. Likewise, Holmes et al. (2006) found compelling evidence suggesting that neural systems involved in sensitivity to gaze direction in fearful and angry faces can be modulated by levels of anxiety, demonstrating that the eyes are a particularly important source of information for guiding attention in anxious individuals. Notably, the difficulty to distinguish between the performance of high and low anxiety individuals is directly proportional to the level of threat and arousal evoked by the stimulus. Mennin et al. (2007); Watson (2005); Dix and Meunier (2009); Downey and Coyne (1990) also observed associations linking depression and anxiety with more negative reactivity. In particular, anxiety is more related to (and strongly predicted by) heightened intensity, defined as frequently experiencing strong negative affect and having emotional reactions that occur intensely, easily, and quickly.

The age and maturity of the participants might have also played a role in their reactivity to the emotional context of the interaction. This is showcased in a study by van Bommel et al. (2019) that, similar to our work, examined the emotional responses of mothers and their adolescent children during conflict interactions. In the study, it was observed that mothers were more likely to initiate positivity after negativity, while adolescents tended to reciprocate negativity. Interestingly, mothers with internalizing problems, such as depression or anxiety, expressed fewer negative emotions compared to those without such problems. The likely explanation is that they might be less attuned to the emotional cues of a conflict situation and therefore express less negative emotions, leading them to express less negativity outwardly (Rottenberg and Hindash, 2015). This aligns with other research highlighting a potential disconnect between self-reported negative emotions and outwardly displayed negativity in individuals with depressive symptoms (Chaplin, 2006).

5.3 Advantage of the GBC approach

While the overall performance remained moderate, our GBC models consistently outperformed LightGBM and SVC models. This suggests some degree of compatibility between the modeling approach and the properties of our dataset. And while comparative analyses between models and problem contexts often merit empirical observations, we may explore some potential factors that might explain the patterns in our results.

5.3.1 Curse of dimensionality, kernel methods, and tree-based approaches

SVC operates in a high-dimensional implicit feature space generated by the kernel trick. In our case, with a high feature-to-sample ratio (523 features in the concatenated early-fused feature set vs. 100 samples) and potentially excessive transformation (e.g., up to 9th-degree in the case of polynomial kernel), the data might have become too sparsely distributed in the hyperplane, hindering effective classification. In contrast, being a tree-based model, GBC was better equipped to handle non-linear relationships in our dataset that SVC might have struggled with. Thus, for our context, GBC's rather pragmatic strategy of systematically improving based on losses and placing less emphasize on the inherent characteristics of the features was perhaps more effective.

5.3.2 Hyperparameter optimization

SVC's smaller hyperparameter space (5 parameters, compared to 7 for GBC and LightGBM) combined with few value options (3 per parameter) might have limited its optimization potential. On the other hand, GBC employed fewer estimators and shallower trees (maximum 20 estimators and depth of 6) compared to LightGBM (maximum 75 estimators and depth of 10). This potentially helped GBC avoid overfitting the training data, leading to better generalization. Also, GBC allowed for fewer samples per leaf node (6 samples compared to 30 for LightGBM). This might have facilitated finer groupings with less impurity within GBC's leaf nodes whereas LightGBM's leaf nodes may contained greater impurity.

5.3.3 Early vs. late fusion

As to the performance difference between early fusion (0.29 MCC, 0.64 ROC AUC, and 0.66 accuracy) and late fusion GBC models (0.28 MCC, 0.64 ROC AUC, and 0.66 accuracy), we hypothesize that while both were ensembles, the early fusion model learned from a richer feature set incorporating information from all modalities, with each base-learner improving the performance of the previous. In contrast, the late-fusion model consisted of base-learners/sub-models trained on individual modalities with no interconnectedness at the training stage. This might have made it more prone to incomplete information, and explains the slight advantage of early fusion in our case.

5.4 Suitability of the MCC metric

The MCC metric consistently measured lower performance compared to other metrics (see Figure 14 and 15). This is because unlike other metrics, MCC value is high *only* when a model produced a high percentage of true positives (over the total number of positives) *and* true negatives (over the negatives) (Chicco and Jurman, 2020). Table 4 of the previously cited paper demonstrates that in a highly imbalanced data, a simple model predicting the majority class can achieve high accuracy (0.9) and F1-score (0.95) despite entirely missing the minority class. However, the MCC score (-0.03) correctly reflects this poor performance.

Although less optimistic, MCC showed considerable agreement with most of other metrics for relative comparisons between scenarios (also discussed in Section 4.2). This suggests that MCC might be a stricter but potentially more informative metric, especially for datasets with characteristics similar to ours. Moreover, we believe such characteristics (such as small sample size, high dimensionality, and imbalanced class distribution) are more realistic and representative of real-world applications. Additionally, unlike metrics that consider the entire dataset, MCC's measurement focuses on correct predictions per class, making it much less affected by class imbalance.

5.5 Answers to the research questions

Main research question: *To what extent can anxiety symptoms be detected in 9 year-old children using unimodal and multimodal approaches?*

Our experiments suggest the potential of detecting children anxiety symptoms through multiple modalities. The symptoms seem to be more detectable in an atmosphere of disagreement and dispute than a calm or agreeable interaction, and leveraging tree-based classifiers on a multimodal feature set tends to yield better results. With a moderate 0.66 accuracy and 0.29 correlation from our best model (GBC model trained on multimodal, early-fused features from conflict interaction), we believe that there are rooms for improvement before the system reach usage feasibility.

Sub-question 1: *To what extent can unimodal methods detect symptoms of anxiety compared to multimodal approaches?*

We found evidence for a statistically significant advantage for models that used multimodal features over those that learned from unimodal features (as explained in Section 4.1). However, when the performance was averaged across multiple iterations/folds, we saw comparable results despite slightly lower performance from the unimodal model (as shown in Table 5). This might imply that training on unimodal features can still yield a sufficiently good performance. This has practical implications; for example, when there is only sufficient resources to acquire data from one or a few modalities.

To merge the modalities, our results suggested that the early fusion strategy (merging at the feature level) outperformed the late fusion strategy (at the prediction level). As for the unimodal features, linguistic and facial expression seemed to have more predictive power and yielded better results compared to other modalities. Of course, a number of other factors also affect the quality of the training data (such as the data collection procedure, the type of extracted features and their predictive power, the tools used to process the data, and so on).

Sub-question 2: *How do videos recorded during a conflicting and a cooperative task compare to detect symptoms of anxiety?*

The results clearly indicate the superiority of conflict interaction over cooperative or peaceful atmospheres at eliciting anxious behavior. Although the cooperation scenario occasionally outperformed the conflict scenario when controlled for model and feature type, this occurrences were very few. Overall, the highest performing models from each modality and fusion type were all trained on the conflict data.

Sub-question 3: *What are the important verbal and non-verbal indicators of anxiety symptoms?*

SHAP analysis performed on unimodal models suggests that low spectral flux and amplitude (loudness) of voice were the top acoustic predictors of anxiety. Spectral flux refers to the speed at which the power spectrum of a signal changes (Giannoulis et al., 2013). In the content of speech, anxiety is also associated with the frequent use of third-person plural nouns, filler sounds, past-focused words, and words that convey negative emotions. Unsurprisingly, a high usage of words that explicitly refer to anxiousness and fear also strongly indicates the presence of anxiety. Anxiety also manifests itself through facial expression, specifically with high activity of muscles around the mouth (AU 23, 11, 10) and eyes (AU 43, 04). Further, it is also manifested through body language, most prominently with frequent movements of the hands, head, and contraction of the upper body, adopting a closed and guarded posture.

Similarly, when feature sets from different modalities were learned jointly, we found that the top anxiety indicators are movements around the hands and mouth (more specifically the raising of the chin/AU 17), focusing on the past with frequent combination of filler sounds, closing of the eyes, and upper body contraction. This implies an agreement of feature importance. That is to say, the same features deemed highly predictive by the unimodal models were also found to be important by the multimodal models.

6. Limitations & future work

One limitation of this study is the relatively small sample size. Given the complex interplay of multimodal communication, diverse interaction scenarios, and multiple feature engineering strategies, a larger dataset would have provided a robust foundation for model development and likely would have yielded a higher performance. While increasing sample size typically demands substantial resources, leveraging pre-processed datasets specifically engineered for model training could be a viable strategy to mitigate this challenge.

Another limitation may have come from the way the interactions were recorded. The cameras were located at varying angles and distances to the subject, with some videos capturing only part of the subject's body. This might have contributed to inaccuracies when extracting the body pose features. In the same way, the extraction of acoustic features could have been optimized by isolating the subject's voice from environmental noises as well as other participants such as parents and moderators. Although from practical point of view, noises are almost unavoidable in real-world engineering applications and the ability to account for deviations from normal scenarios is a hallmark of a robust system.

Limitations might have also been inherited from the feature extraction tools. For example, despite the high importance assigned to some of our linguistic features, the tool itself (LIWC) is not without its limitations. LIWC is a rather crude instrument (as is admitted on the website[1]) that works by calculating the presence of pre-defined words within a text corpus. This approach has several shortcomings. First, words are analyzed in isolation while context and nuanced meanings are prone to be neglected. Polysemous words (words with different meanings depending on context or user) might be susceptible to misinterpretation. To take an example from our work, we may not know with absolute certainty whether the frequent usage of *focuspast* was due to the participant's mental state or it might had been influenced by the context of the argument between a parent and child (i.e., the discussions likely centered around recounting past events or mistakes rather than future plans, especially considering the age of our participants). Second, the predefined dictionaries may miss relevant terms, including slang, jargon, and newly emerging expressions. It is also not designed to capture non-verbal communication such as the use of non-words (either expressed verbally or written). This is particularly relevant to studies such as ours where non-verbal cues might carry significant meaning. Third, mixed language use that is prevalent in multilingual communities (e.g., Singapore's use of four official languages) or the use of absorbed foreign language (e.g., Dutch words in

[1] https://www.liwc.app/help/howitworks

the Indonesian language or French words in the English language) may pose additional challenges for LIWC. Finally, language is constantly evolving, and LIWC dictionaries may not always keep pace with these changes, potentially overlooking recently developed expressions.

Future work may explore alternative methods in combining samples from multiple scenarios. Our approach to joint scenario analysis treated conflict and cooperation data as separate samples. In the future, one may experiment with concatenating them at an earlier stage, forming an extended feature vector. Subsequently, applying feature selection (e.g., Principal Component Analysis) might be beneficial in order to suppress the feature complexity. This approach might potentially yield a better performance since the datapoints would compactly encode information from multiple scenarios, enabling the model to learn from all the different atmospheres of interactions more efficiently.

Future research in this topic may also experiment with alternative feature selection methods to potentially enhance model performance. While our heuristic approach was informed by previous studies, future works can employ other promising methods such as Principal Component Analysis (PCA). Although the interpretation of principal components is not straightforward, it can be approximated by analyzing the components of the eigenvectors. As different feature importance methods yield varying results, optimizing this step could be a topic of study on its own right (Petelin et al., 2023).

Future research could examine the plethora of both shallow and deep learning approaches, and explore the effects of including deep features in the dataset. A number of studies have shown varying degrees of success in applying deep features. For example, the work of Diep et al. (2022) utilized deep acoustic (extracted with Wav2Vec 2.0) and verbal feature sets (extracted with RoBERTa language tokenization tool) for the task of detecting depression and anxiety. By including deep features, the authors managed to increase the average F1 score by 3% to 4% compared to using hand-crafted features only. Similarly, Schroff et al. (2015) employed 128 deep visual features extracted using FaceNet for training facial recognition models. They managed to achieve around 95% to 99% accuracy on popular face databases. Baird et al. (2020) observed a mixed effect when comparing DEEPSPECTRUM (4,096 dimensions) deep acoustic representations with two other hand-crafted acoustic feature sets (6,373 dimensional ComParE and 88 dimensional eGeMAPS feature sets) for the task of detecting anxiety through vocal characteristics.

Considering the heavy dependence of a young child and the enormous influence of adults on a child's early development, it would be interesting to examine the role of adults in affecting expressions of anxiety. One possibility is to include behavioral features from parents and caretakers as predictors in order to explore the effect of parenting styles and interaction with caretakers on the child's mental well-being. Furthermore, since culture strongly influences the way in which people and institutions react to a child's problems, it is also known that psychopathology vary across cultures (Ryan et al., 2005). For example, shyness and over-sensitivity in Western cultures tend to be associated with peer rejection and social maladjustment. In contrast, for children in Shanghai, such behaviors are associated with leadership, school competence, and academic achievement (Mash and Barkley, 2003). Therefore, the participant's cultural background could be included as a variable in order to investigate its effect on the expressions of mental disorder symptoms and mental well-being in general.

7. Conclusion & final remarks

In this thesis, we studied the feasibility of detecting symptoms of anxiety in children using multimodal indicators. Specifically, we extracted acoustic, linguistic, and visual data from video recordings of parent-and-child dyads interacting in two contexts/scenarios: cooperative (vacation planning) and conflict (arguing). We compared multiple approaches by varying the interaction contexts/scenarios, feature modalities, fusion types, and modeling approaches.

We observed that anxiety symptoms were more pronounced during disagreements. They were exhibited through various cues that agree with the psychological literature. We also found that while training on multimodal datasets was proven beneficial, unimodal feature sets also produced reasonable performance. Moreover, feature extraction techniques also played crucial roles in contributing to the model's performance.

To conclude, our study suggests that anxiety are expressed mostly through the hand movements (and, more broadly, body posture), facial expression (more specifically, the mouth area), and choice of words. It also suggests that training models on multimodal features is more promising than unimodal features. Further, we recommend to collect data from conflictual interactions as they seem to elicit anxious behavior more than cooperative or peaceful interactions. However, these patterns might depend on a number of other factors, such as the tools and techniques and the condition under which the data was collected. While our models managed to produce moderate performance, we identified a number of opportunities for improvements and recommend them for future research in this area.

8. Acknowledgements

I would like to express my sincere gratitude to my supervisors, Dr. Itir Önal Ertuğrul and Mang Ning, M.Sc., for their mentorship and support throughout this research. Their expertise, insights, and ideas were instrumental in bringing this project to fruition. I would also like to thank Prof. Albert Salah for the insightful discussions prior to the start of this work, as well as for his valuable input on my thesis proposal.

I am equally grateful to all my docents from Universitas Katolik Parahyangan, especially Dr. Lionov Wiratma and Dr. Mariskha Adithia. They have imparted not only knowledge but also wisdom to their pupils. Their life stories and messages were the seeds of inspiration that drive my pursuit of higher education. I extend the same gratitude and appreciation to my IELTS teachers, Andy Li, April Saraswati, Qanitan Aryun, and Maya Roxanne. Their class was the spark that set this entire journey in motion.

I have been fortunate to be surrounded by many supportive friends and family. I would like to extend my deepest honor and gratitude to my parents, Vincentius Handrianto and Meiske Walah, and my sibling, Leonardus Andrew, for their unwavering love, support, and encouragement, which have strengthened my resolve to pursue this study. I would also like to thank my fellow students, Caesar Alpha, Areti Psaropoulou, Renate Burema, Alice Vitali, Lisa Tanaka, Thijs van den Berg, and Ruben Weijers, for their helpful discussions and camaraderie. Finally, I express my heartfelt appreciation to my friends Mas Arvie, Mas Ryan & Mba Eka, Mas Zaki & Mba Atina, Samuel Chrisna, Andi Hermansyah, Citra Larasati, Barry Meyer, Ikhwan Rasyidin, Mardhiyyah Anggun, Abdul Hadi, Ati Cahya, Longginus Alyandu, and Nabila Talitha for their unwavering friendship and moral support, which have been invaluable in navigating the challenges of graduate study.

9. Appendix

9.1 Model performance

Performance of models trained on the dataset from the conflict, cooperation, and the combination of both scenarios are shown in Table 13, Table 14, and Table 15.

Modality	Model	Accuracy	F1-positive	ROC AUC	PR AUC	MCC
Acoustic	GBC*	0.63 ± 0.04	0.43 ± 0.16	0.6 ± 0.06	0.63 ± 0.12	0.21 ± 0.11
	SVM	0.62 ± 0.05	0.29 ± 0.11	0.56 ± 0.04	0.46 ± 0.09	0.18 ± 0.14
	LightGBM	0.57 ± 0.04	0.38 ± 0.15	0.54 ± 0.06	0.55 ± 0.12	0.08 ± 0.12
Linguistic	GBC	0.49 ± 0.14	0.32 ± 0.18	0.46 ± 0.14	0.41 ± 0.07	-0.08 ± 0.29
	SVM	0.56 ± 0.04	0.19 ± 0.14	0.5 ± 0.04	0.36 ± 0.05	-0.01 ± 0.09
	LightGBM*	0.63 ± 0.1	0.51 ± 0.11	0.6 ± 0.1	0.6 ± 0.12	$\mathbf{0.3} \pm 0.2$
Facial expression (continuous)	GBC*	$\mathbf{0.66} \pm 0.09$	0.53 ± 0.11	0.63 ± 0.08	0.55 ± 0.07	0.28 ± 0.18
	SVM	0.55 ± 0.13	0.25 ± 0.21	0.5 ± 0.12	0.49 ± 0.13	0 ± 0.31
	LightGBM	0.56 ± 0.09	0.41 ± 0.08	0.53 ± 0.08	0.43 ± 0.07	0.16 ± 0.21
Facial expression (binary)	GBC	0.53 ± 0.05	0.4 ± 0.07	0.51 ± 0.05	0.47 ± 0.09	0.02 ± 0.1
	SVM*	0.64 ± 0.1	0.4 ± 0.23	0.59 ± 0.1	0.52 ± 0.16	0.22 ± 0.28
	LightGBM	0.54 ± 0.04	0.37 ± 0.06	0.51 ± 0.03	0.54 ± 0.1	0.02 ± 0.07
Body pose	GBC*	0.61 ± 0.1	0.45 ± 0.16	0.58 ± 0.1	0.6 ± 0.09	0.2 ± 0.2
	SVM	0.52 ± 0.09	0.41 ± 0.08	0.5 ± 0.07	0.51 ± 0.09	0.02 ± 0.16
	LightGBM	0.6 ± 0.03	0.39 ± 0.1	0.56 ± 0.01	0.56 ± 0.12	0.17 ± 0.07
Multimodal early fusion	GBC*△	$\mathbf{0.66} \pm 0.1$	$\mathbf{0.56} \pm 0.13$	$\mathbf{0.64} \pm 0.1$	0.63 ± 0.18	0.29 ± 0.22
	SVM	0.64 ± 0.09	0.47 ± 0.13	0.61 ± 0.09	$\mathbf{0.64} \pm 0.13$	0.26 ± 0.2
	LightGBM	0.6 ± 0.12	0.48 ± 0.17	0.58 ± 0.13	0.54 ± 0.07	0.16 ± 0.25
Multimodal early fusion (binary facial expression)	GBC	0.59 ± 0.1	0.4 ± 0.1	0.55 ± 0.09	0.55 ± 0.1	0.14 ± 0.2
	SVM	0.52 ± 0.07	0.15 ± 0.14	0.46 ± 0.07	0.43 ± 0.09	-0.1 ± 0.17
	LightGBM*	0.6 ± 0.13	0.49 ± 0.19	0.58 ± 0.14	0.62 ± 0.18	0.17 ± 0.29
Multimodal late fusion	GBC*	$\mathbf{0.66} \pm 0.16$	0.54 ± 0.21	$\mathbf{0.64} \pm 0.16$	0.59 ± 0.14	0.28 ± 0.33
	SVM	0.53 ± 0.07	0.2 ± 0.06	0.47 ± 0.06	0.5 ± 0.11	-0.04 ± 0.18
	LightGBM	0.64 ± 0.1	0.52 ± 0.13	0.62 ± 0.1	0.49 ± 0.12	0.26 ± 0.2
Multimodal late fusion (binary facial expression)	GBC	0.61 ± 0.1	0.49 ± 0.16	0.59 ± 0.12	0.56 ± 0.07	0.19 ± 0.23
	SVM	0.59 ± 0.05	0.33 ± 0.12	0.54 ± 0.04	0.52 ± 0.17	0.12 ± 0.13
	LightGBM*	0.63 ± 0.06	0.5 ± 0.12	0.62 ± 0.06	0.55 ± 0.08	0.25 ± 0.13

Table 13. Model performance on dataset from conflict scenario (averaged over 5 folds). **Bold**: highest value in the respective metric. (*) This model achieved the best performance in the respective modality, across scenarios. (△) From all the models that used the conflict scenario data, the GBC model trained on early-fused multimodal features produced the best overall performance.

Modality	**Model**	**Accuracy**	**F1-positive**	**ROC AUC**	**PR AUC**	**MCC**
Acoustic	GBC	0.55 ± 0.1	0.3 ± 0.13	0.51 ± 0.09	0.46 ± 0.13	0.04 ± 0.25
	SVM	0.6 ± 0.06	0.19 ± 0.12	0.52 ± 0.05	0.54 ± 0.15	0.1 ± 0.2
	LightGBM	0.52 ± 0.05	0.2 ± 0.13	0.47 ± 0.03	0.47 ± 0.1	-0.08 ± 0.08
Linguistic	GBC	0.43 ± 0.12	0.17 ± 0.13	0.39 ± 0.12	0.37 ± 0.09	-0.24 ± 0.23
	SVM	0.58 ± 0.02	0.1 ± 0.13	0.5 ± 0.02	0.44 ± 0.11	0.01 ± 0.05
	LightGBM	0.51 ± 0.14	0 ± 0	0.43 ± 0.11	0.32 ± 0.07	-0.16 ± 0.23
Facial expression (continuous)	GBC	0.55 ± 0.05	0.32 ± 0.13	0.5 ± 0.06	0.45 ± 0.08	0 ± 0.15
	SVM	0.6 ± 0.03	0.27 ± 0.16	0.54 ± 0.03	0.55 ± 0.04	0.11 ± 0.08
	LightGBM	0.55 ± 0.04	0.36 ± 0.08	0.51 ± 0.04	0.48 ± 0.1	0.03 ± 0.09
Facial expression (binary)	GBC	0.58 ± 0.07	0.32 ± 0.13	0.53 ± 0.07	0.47 ± 0.1	0.07 ± 0.16
	SVM	0.58 ± 0.02	0.22 ± 0.16	0.52 ± 0.03	0.44 ± 0.08	0.05 ± 0.08
	LightGBM△	$\mathbf{0.63} \pm 0.04$	0.34 ± 0.12	$\mathbf{0.57} \pm 0.04$	0.58 ± 0.11	$\mathbf{0.21} \pm 0.1$
Body pose	GBC	0.53 ± 0.09	0.25 ± 0.15	0.48 ± 0.08	0.42 ± 0.1	-0.05 ± 0.22
	SVM	0.59 ± 0.04	0.11 ± 0.15	0.51 ± 0.04	0.41 ± 0.07	0.02 ± 0.14
	LightGBM	0.46 ± 0.08	0.25 ± 0.09	0.43 ± 0.07	0.37 ± 0.08	-0.16 ± 0.14
Multimodal early fusion	GBC	0.49 ± 0.08	0.23 ± 0.13	0.44 ± 0.07	0.43 ± 0.1	-0.13 ± 0.17
	SVM	0.55 ± 0.1	0.2 ± 0.16	0.49 ± 0.1	0.41 ± 0.11	-0.01 ± 0.22
	LightGBM	0.59 ± 0.07	$\mathbf{0.38} \pm 0.1$	0.55 ± 0.07	0.52 ± 0.12	0.13 ± 0.17
Multimodal early fusion (binary facial expression)	GBC	0.49 ± 0.07	0.26 ± 0.17	0.45 ± 0.08	0.42 ± 0.09	-0.12 ± 0.18
	SVM	0.61 ± 0.07	0.25 ± 0.17	0.54 ± 0.07	0.52 ± 0.15	0.11 ± 0.16
	LightGBM	0.62 ± 0.07	0.29 ± 0.2	0.56 ± 0.08	$\mathbf{0.59} \pm 0.11$	0.17 ± 0.22
Multimodal late fusion	GBC	0.52 ± 0.08	0.26 ± 0.12	0.47 ± 0.07	0.43 ± 0.08	-0.05 ± 0.17
	SVM	0.62 ± 0.04	0.19 ± 0.16	0.55 ± 0.05	0.56 ± 0.02	0.15 ± 0.16
	LightGBM	0.45 ± 0.1	0.16 ± 0.13	0.4 ± 0.1	0.44 ± 0.08	-0.23 ± 0.22
Multimodal late fusion (binary facial expression)	GBC	0.53 ± 0.07	0.28 ± 0.14	0.49 ± 0.05	0.43 ± 0.08	-0.04 ± 0.15
	SVM	0.59 ± 0.02	0.08 ± 0.1	0.51 ± 0.01	0.38 ± 0.04	0.03 ± 0.03
	LightGBM	0.52 ± 0.09	0.17 ± 0.15	0.46 ± 0.09	0.51 ± 0.1	-0.11 ± 0.24

Table 14. Model performance on dataset from cooperation scenario. (△) The overall best performance was achieved by the LightGBM model trained on the facial expression features (with binary AUs).

Modality	Model	Accuracy	F1-positive	ROC AUC	PR AUC	MCC
Acoustic	GBC	0.55 ± 0.05	0.38 ± 0.06	0.52 ± 0.04	0.49 ± 0.06	0.05 ± 0.09
	SVM	0.55 ± 0.05	0.22 ± 0.1	0.5 ± 0.03	0.43 ± 0.04	0.02 ± 0.12
	LightGBM	0.61 ± 0.06	0.42 ± 0.09	0.57 ± 0.06	0.55 ± 0.1	0.18 ± 0.13
Linguistic	GBC	0.53 ± 0.02	0.23 ± 0.09	0.47 ± 0.02	0.4 ± 0.01	-0.07 ± 0.06
	SVM	0.51 ± 0.03	0.17 ± 0.12	0.46 ± 0.04	0.4 ± 0.07	-0.12 ± 0.1
	LightGBM	0.5 ± 0.1	0.07 ± 0.06	0.44 ± 0.09	0.36 ± 0.06	-0.13 ± 0.23
Facial expression (continuous)	GBC	0.57 ± 0.04	0.39 ± 0.08	0.53 ± 0.04	0.49 ± 0.08	0.07 ± 0.09
	SVM	0.55 ± 0.11	0.41 ± 0.11	0.52 ± 0.1	0.54 ± 0.14	0.06 ± 0.22
	LightGBM	0.55 ± 0.05	0.39 ± 0.05	0.52 ± 0.03	0.44 ± 0.06	0.05 ± 0.07
Facial expression (binary)	GBC	0.54 ± 0.08	0.36 ± 0.2	0.51 ± 0.1	0.46 ± 0.1	0.01 ± 0.23
	SVM	0.61 ± 0.1	$\mathbf{0.5} \pm 0.11$	$\mathbf{0.59} \pm 0.09$	0.53 ± 0.11	0.19 ± 0.2
	LightGBM	0.58 ± 0.02	0.43 ± 0.07	0.56 ± 0.02	0.49 ± 0.06	0.13 ± 0.04
Body pose	GBC	0.58 ± 0.05	0.34 ± 0.07	0.53 ± 0.04	0.48 ± 0.08	0.07 ± 0.11
	SVM	0.49 ± 0.08	0.3 ± 0.13	0.47 ± 0.07	0.44 ± 0.07	-0.03 ± 0.18
	LightGBM	0.53 ± 0.12	0.39 ± 0.17	0.51 ± 0.13	0.46 ± 0.11	0.01 ± 0.26
Multimodal early fusion	GBC△	$\mathbf{0.62} \pm 0.08$	0.45 ± 0.1	$\mathbf{0.59} \pm 0.07$	$\mathbf{0.57} \pm 0.1$	$\mathbf{0.2} \pm 0.17$
	SVM	0.58 ± 0.05	0.4 ± 0.08	0.55 ± 0.05	0.51 ± 0.09	0.1 ± 0.11
	LightGBM	0.57 ± 0.14	0.41 ± 0.13	0.54 ± 0.13	0.51 ± 0.16	0.09 ± 0.28
Multimodal early fusion (binary facial expression)	GBC	0.56 ± 0.1	0.38 ± 0.08	0.53 ± 0.09	0.48 ± 0.1	0.09 ± 0.19
	SVM	0.59 ± 0.06	0.4 ± 0.11	0.55 ± 0.06	0.53 ± 0.11	0.11 ± 0.13
	LightGBM	0.58 ± 0.11	0.43 ± 0.07	0.55 ± 0.1	0.46 ± 0.11	0.14 ± 0.22
Multimodal late fusion	GBC	0.51 ± 0.06	0.3 ± 0.13	0.47 ± 0.07	0.49 ± 0.08	-0.06 ± 0.16
	SVM	0.5 ± 0.11	0.31 ± 0.13	0.47 ± 0.1	0.55 ± 0.14	-0.06 ± 0.22
	LightGBM	0.54 ± 0.1	0.34 ± 0.12	0.51 ± 0.1	0.45 ± 0.06	0.02 ± 0.23
Multimodal late fusion (binary facial expression)	GBC	0.55 ± 0.02	0.39 ± 0.05	0.49 ± 0.07	0.51 ± 0.07	0.04 ± 0.05
	SVM	0.55 ± 0.09	0.36 ± 0.12	0.52 ± 0.09	0.53 ± 0.11	0.04 ± 0.18
	LightGBM	0.56 ± 0.12	0.39 ± 0.15	0.53 ± 0.12	0.49 ± 0.03	0.08 ± 0.27

Table 15. Model performance on dataset from both conflict and cooperation scenario. (△) The GBC model trained on early-fused multimodal features produced the best performance.

9.2 Confusion matrices

Below are confusion matrices of the best-performing models trained on features from the conflict (Table 16), cooperation (Table 17), and the combination of both scenarios (Table 18).

Predicted / Actual	**Positive**	**Negative**	**Sum**
Positive	22	15	37
Negative	19	44	63
Sum	41	59	
Accuracy	0.66		
Precision	0.59	0.7	
Recall	0.54	0.75	
F1	0.56	0.72	

Table 16. Confusion matrix of the best model trained on the conflict scenario

Predicted / Actual	**Positive**	**Negative**	**Sum**
Positive	10	6	16
Negative	31	53	84
Sum	41	59	
Accuracy	0.63		
Precision	0.63	0.63	
Recall	0.24	0.90	
F1	0.35	0.74	

Table 17. Confusion matrix of the best model trained on the cooperation scenario

Predicted / Actual	Positive	Negative	Sum
Positive	30	23	53
Negative	52	95	147
Sum	82	118	
Accuracy	0.63		
Precision	0.57	0.65	
Recall	0.37	0.81	
F1	0.45	0.72	

Table 18. Confusion matrix of the best model trained on both the conflict and cooperation scenario. Note that in the combined scenario, datasets from the two scenarios are concatenated, doubling the number of samples. Thus normalization is required when making side-by-side comparison between the number of predictions.

9.3 Contingency tables

Below are contingency tables used in McNemar's tests to compare the performances between various scenarios.

	Conflict-positive	Conflict-negative
Cooperation-positive	7	9
Cooperation-negative	30	54

Table 19. Contingency table between conflict and cooperation scenario.

	Conflict-positive	Conflict-negative
Combined-positive	35	3
Combined-negative	4	58

Table 20. Contingency table between conflict and combined scenario.

	Cooperative-positive	**Cooperative-negative**
Combined-positive	12	27
Combined-negative	4	57

Table 21. Contingency table between cooperative and combined scenario.

	Unimodal-positive	**Unimodal-negative**
Multimodal-positive	20	19
Multimodal-negative	7	54

Table 22. Contingency table between the predictions of the best unimodal model and multimodal model.

9.4 Hyperparameters

Optimal hyperparameter settings of the best performing model of each modality are shown in Table 23.

Model	**Parameter**
Unimodal acoustic (GBC)	learning rate 0.25, max. depth 4, max. features 0.8, min. samples leaf 6, min. samples split 2, num. estimators 5, subsampling rate 0.5
Unimodal linguistic (LightGBM)	lambda L1 0.1, lambda L2 0.1, learning rate 0.1, max. depth 4, min. data in leaf 15, num. estimators 15, num. leaves 15
Unimodal face cont. AUs (GBC)	learning rate 0.25, max. depth 6, max. features 0.9, min. samples leaf 2, min. samples split 6, num. estimators 10, subsampling rate 0.5
Unimodal face bin. AUs (SVM)	C 3, degree 2, gamma auto, kernel polynomial, shrinking True
Unimodal body pose (GBC)	learning rate 0.2, max. depth 2, max. features 0.8, min. samples leaf 6, min. samples split 2, num. estimators 10, subsampling rate 0.5
Multimodal early fusion cont. AUs (GBC)	learning rate 0.25, max. depth 4, max. features 0.6, min. samples leaf 6, min. samples split 2, num. estimators 20, subsampling rate 0.85
Multimodal early fusion bin. AUs (LightGBM)	lambda L1 0.2, lambda L2 0.1, learning rate 0.2, max. depth 4, min. data in leaf 15, num. estimators 30, num. leaves 15

Table 23. Hyperparameters of the best performing model of each modality.

Optimal hyperparameter settings of the best performing model of each scenario are shown in Table 24.

Model	**Parameter**
Conflict - Multimodal early fusion cont. AUs (GBC)	learning rate 0.25, max. depth 4, max. features 0.6, min. samples leaf 6, min. samples split 2, num. estimators 20, subsampling rate 0.85
Cooperation - Unimodal face bin. AUs (LightGBM)	lambda L1 0.05, lambda L2 0.05, learning rate 0.1, max. depth 4, min. data in leaf 30, num. estimators 15, num. leaves 15
Combined - Multimodal eearly fusion cont. AUs (GBC)	learning rate 0.2, max. depth 4, max. features 0.6, min. samples leaf 2, min. samples split 2, num. estimators 20, subsampling rate 0.85

Table 24. Hyperparameters of the best performing model of each scenario.

Bibliography

Adams, A., Mahmoud, M., Baltrušaitis, T., and Robinson, P. (2015). Decoupling facial expressions and head motions in complex emotions. In *2015 International Conference On Affective Computing and Intelligent Interaction (ACII)*, pages 274–280, Xi'an, China.

Ahn, W.-Y. and Busemeyer, J. R. (2016). Challenges and promises for translating computational tools into clinical practice. *Current Opinion in Behavioral Sciences*, 11:1–7.

Al-Mosaiwi, M. and Johnstone, T. (2018). In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation. *Clinical Psychological Science*, 6(4):529–542.

Arellano, D., Perales, F. J., and Varona, J. (2014). Mood and Its Mapping onto Facial Expressions. In Perales, F. J. and Santos-Victor, J., editors, *Articulated Motion and Deformable Objects*, pages 31–40. Springer International Publishing.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020)*, pages 12449–12460, Vancouver, Canada.

Baird, A., Cummins, N., Schnieder, S., Krajewski, J., and Schuller, B. (2020). An evaluation of the effect of anxiety on speech - computational prediction of anxiety from sustained vowels. In Meng, H., Xu, B., and Zheng, T., editors, *Proceedings: Interspeech 2020, 25-29 October 2020*, pages 4951–4955, Shanghai, China.

Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L.-P. (2018). OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66, Xi'an, China.

Barreto, A., Zhai, J., and Adjouadi, M. (2007). Non-intrusive physiological monitoring for automated stress detection in human-computer interaction. In Lew, M., Sebe, N., Huang, T. S., and Bakker, E. M., editors, *Human–Computer Interaction*, pages 29–38, Berlin, Heidelberg. Springer Berlin Heidelberg.

Bateson, M., Brilot, B., and Nettle, D. (2011). Anxiety: An Evolutionary Approach. *The Canadian Journal of Psychiatry*, 56(12):707–715.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165 – 1188.

Bernstein, G. A., Massie, E. D., Thuras, P. D., Perwien, A. R., Borchardt, C. M., and Crosby, R. D. (1997). Somatic Symptoms in Anxious-Depressed School Refusers. *Journal of the American Academy of Child and Adolescent Psychiatry*, 36(5):661–668.

Bhattacharya, U., Rewkowski, N., Banerjee, A., Guhan, P., Bera, A., and Manocha, D. (2021). Text2Gestures: A Transformer-Based Network for Generating Emotive Body Gestures for Virtual Agents. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces (IEEE VR)*, pages 1–10, Lisboa, Portugal. IEEE.

Bilalpur, M., Hinduja, S., Cariola, L. A., Sheeber, L. B., Alien, N., Jeni, L. A., Morency, L.-P., and Cohn, J. F. (2023). Multimodal Feature Selection for Detecting Mothers' Depression in Dyadic Interactions with their Adolescent Offspring. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8, Waikoloa Beach, HI, USA.

Bosch, N. (2021). AutoML Feature Engineering for Student Modeling Yields High Accuracy, but Limited Interpretability. *Journal of Educational Data Mining*, 13(2):55–79.

Boyd, R. L., Ashokkumar, A., Seraj, S., and Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22*. Austin, TX: University of Texas at Austin.

Bredin, H. (2023). pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. Interspeech 2023*, pages 1983–1987, Dublin, Ireland.

Bredin, H. and Laurent, A. (2021). End-to-end speaker segmentation for overlap-aware resegmentation. In *Proc. Interspeech 2021*, pages 3111–3115, Brno, Czech Republic.

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 *2020 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2020)*, pages 7124–7128, Barcelona, Spain.

Buimer, E. E. L., Brouwer, R. M., Mandl, R. C. W., Pas, P., Schnack, H. G., and Pol, H. E. H. (2022). Adverse childhood experiences and fronto-subcortical structures in the developing brain. *Frontiers in Psychiatry*, 13:955871.

Cao, Z., Hidalgo, G., Simon, T., Wei, S., and Sheikh, Y. (2021). OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 43(01):172–186.

Castillo, G. and Neff, M. (2019). What do we express without knowing? Emotion in gesture. In *AAMAS '19, Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 702–710, Richland, SC.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–27.

Chaplin, T. M. (2006). Anger, happiness, and sadness: Associations with depressive symptoms in late adolescence. *Journal of Youth and Adolescence*, 35:977–986.

Cheong, J. H., Jolly, E., Xie, T., Byrne, S., Kenney, M., and Chang, L. J. (2023). Py-Feat: Python Facial Expression Analysis Toolbox. *Affective Science*, 4:781–796.

Chicco, D. and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(6).

Christ, M., Braun, N., Neuffer, J., and Kempa-Liehr, A. W. (2018). Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). *Neurocomputing*, 307:72–77.

Christ, M., Kempa-Liehr, A. W., and Feindt, M. (2016). Distributed and parallel time series feature extraction for industrial big data applications. In *ACML Workshop on Learning on Big Data (WLBD)*, pages 1–17, Hamilton, New Zealand.

Cook, M. (1969). Anxiety, speech disturbances and speech rate. *British Journal of Social and Clinical Psychology*, 8(1):13–21.

de Santis, R. B., Gontijo, T. S., and Costa, M. A. (2023). A Data-Driven Framework for Small Hydroelectric Plant Prognosis Using Tsfresh and Machine Learning Survival Models. *Sensors*, 23(1):12.

Degottex, G., Kane, J., Drugman, T., Raitio, T., and Scherer, S. (2014). COVAREP — A collaborative voice analysis repository for speech technologies. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 960–964, Florence, Italy.

Dibeklioğlu, H., Hammal, Z., and Cohn, J. F. (2018). Dynamic Multimodal Measurement of Depression Severity Using Deep Autoencoding. *IEEE Journal of Biomedical and Health Informatics*, 22(2):525–536.

Diep, B., Stanojevic, M., and Novikova, J. (2022). Multi-modal deep learning system for depression and anxiety detection. In *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, New Orleans, United States.

Dinges, D., Rider, R., Dorrian, J., McGlinchey, E., Rogers, N., Cizman, Z., Goldstein, S., Vogler, C., Venkataraman, S., and Metaxas, D. (2005). Optical computer recognition of facial expressions associated with stress induced by performance demands. *Aviation, Space, and Environmental Medicine*, 76(6 Suppl):B172–B182.

Dix, T. and Meunier, L. N. (2009). Depressive symptoms and parenting competence: An analysis of 13 regulatory processes. *Developmental Review*, 29:45–68.

Downey, G. and Coyne, J. C. (1990). Children of depressed parents: An integrative review. *Psychological Bulletin*, 108:50–76.

Egger, H. L. and Angold, A. (2006). Common emotional and behavioral disorders in preschool children: presentation, nosology, and epidemiology. *Journal of Child Psychology and Psychiatry*, 47(3-4):313–337.

Ekman, P. and Friesen, W. V. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA.

Ertugrul, I. O., Cohn, J. F., Jeni, L. A., Zhang, Z., Yin, L., and Ji, Q. (2020). Crossing Domains for AU Coding: Perspectives, Approaches, and Measures. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2(2):158–171.

Ertugrul, I. O., Jeni, L. A., Ding, W., and Cohn, J. F. (2019). AFAR: A Deep Learning Based Tool for Automated Facial Affect Recognition. In *2019 14th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2019)*, page 1, Lille, France.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., and Truong, K. P. (2015). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Fast, L. and Funder, D. (2008). Personality as manifest in word use: Correlations with self- report, acquaintance report, and behavior. *Journal of Personality and Social Psychology*, 94(2):334–346.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189 – 1232.

Friedman, J. H. (2002). Stochastic Gradient Boosting. *Computational Statistics & Data Analysis*, 38(4):367–378.

Gao, H., Yuce, A., and Thiran, J.-P. (2014). Detecting emotional stress from facial expressions for driving safety. In *2014 IEEE International Conference on Image Processing, ICIP 2014*, Paris, France.

Giannakakis, G., Pediaditis, M., Manousos, D., Kazantzaki, E., Chiarugi, F., Simos, P. G., Marias, K., and Tsiknakis, M. (2017). Stress and anxiety detection using facial cues from videos. *Biomedical Signal Processing and Control*, 31:89–101.

Giannoulis, D., Massberg, M., and Reiss, J. D. (2013). Automating Dynamic Range Compression. *Journal of the Audio Engineering Society*, 61(10):716–726.

Goberman, A., Hughes, S., and Haydock, T. (2011). Acoustic characteristics of public speaking: Anxiety and practice effects. *Speech Communication*, 53(6):867–876.

Gorman, J. M. (1996). Comorbid depression and anxiety spectrum disorders. *Depression and anxiety*, 4(4):160–168.

Guolin, K., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems (NIPS 2017)*, volume 30, pages 3149–3157. Curran Associates, Inc.

Hadar, U., Steiner, T., Grant, E., and Rose, F. (1983). Head movement correlates of juncture and stress at sentence level. *Language and Speech*, 26(Pt 2):117–129.

Han, J., Qian, K., Song, M., Yang, Z., Ren, Z., Liu, S., Liu, J., Zheng, H., Ji, W., Koike, T., Li, X., Zhang, Z., Yamamoto, Y., and Schuller, B. W. (2020). An Early Study on Intelligent Analysis of Speech under COVID-19: Severity, Sleep Quality, Fatigue, and Anxiety. *Interspeech 2020.*

Hargrave, J. (2008). Do you speak body language?: Mastering the art of nonverbal communication key in interrogations. *The Forensic Examiner*, 17(3):17–22.

Harrigan, J. A. and O'Connell, D. M. (1996). How do you look when feeling anxious? Facial displays of anxiety. *Personality and Individual Differences*, 21(2):205–212.

Henderson, T. and Fulcher, B. D. (2021). An Empirical Evaluation of Time-Series Feature Sets. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pages 1032–1038, Auckland, New Zealand.

Hilt, L. M. and Pollak, S. D. (2012). Getting out of rumination: Comparison of three brief interventions in a sample of youth. *Journal of Abnormal Child Psychology*, 40(7):1157–1165.

Hinds, A. L., Woody, E. Z., Drandic, A., Schmidt, L. A., Ameringen, M. V., Coroneos, M., and Szechtman, H. (2010). The psychology of potential threat: Properties of the security motivation system. *Biological Psychology*, 85(2):331–337.

Hinduja, S., Darzi, A., Ertugrul, I. O., Provenza, N., Gadot, R., Storch, E. A., Sheth, S. A., Goodman, W. K., and Cohn, J. F. (2024). Multimodal prediction of obsessive-compulsive disorder, comorbid depression, and energy of deep brain stimulation. *IEEE Transactions on Affective Computing*, pages 1–16.

Holleman, G. A., Hooge, I. T. C., Huijding, J., Deković, M., Kemner, C., and Hessels, R. S. (2021). Gaze and speech behavior in parent–child interactions: The role of conflict and cooperation. *Current Psychology*, 42:12129–12150.

Holmes, A., Richards, A., and Green, S. (2006). Anxiety and sensitivity to eye gaze in emotional faces. *Brain and Cognition*, 60(3):282–294.

Holtzman, N. S., Tackman, A. M., Carey, A. L., Brucks, M. S., Küfner, A. C. P., große Deters, F., Back, M. D., Donnellan, M. B., Pennebaker, J. W., A.Sherman, R., and Mehl, M. R. (2019). Linguistic Markers of Grandiose Narcissism: A LIWC Analysis of 15 Samples. *Journal of Language and Social Psychology*, 38(5-6):773–786.

Hong, R. Y. (2007). Worry and rumination: Differential associations with anxious and depressive symptoms and coping behavior. *Behavior Research and Therapy*, 45(2):277–290.

Ireland, M. E. and Mehl, M. R. (2014). Natural language use as a marker of personality. In Holtgraves, T. M., editor, *The Oxford handbook of language and social psychology*, pages 201–218. Oxford University Press.

Jeni, L. A., Cohn, J. F., and Kanade, T. (2015). Dense 3D Face Alignment from 2D Videos in Real-Time. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–8.

Jung, M. M., van Vlierden, M., Liebregts, W., and Önal Ertuğrul, I. (2023). Do Body Expressions Leave Good Impressions? - Predicting Investment Decisions based on Pitcher's Body Expressions. In *ICMI '23 Companion: Companion Publication of the 25th International Conference on Multimodal Interaction*, pages 36–40. ACM.

Kaminski, J. A., Sloutsky, V. M., and Heckler, A. F. (2006). Do Children Need Concrete Instantiations to Learn an Abstract Concept? In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 28, pages 411–416.

Kane, A. A. and van Swol, L. M. (2023). Using Linguistic Inquiry and Word Count Software to Analyze Group Interaction Language Data. *Group Dynamics: Theory, Research, and Practice*, 27(3):188–201.

Karg, M., Samadani, A.-A., Gorbet, R., Kühnlenz, K., Hoey, J., and Kulić, D. (2013). Body Movements for Affective Expression: A Survey of Automatic Recognition and Generation. *IEEE Transactions on Affective Computing*, 4(4):341–359.

Kasap, S. and Tanhan, F. (2019). The Effect of Body Posture on Foreign Language Anxiety. *Sakarya Üniversitesi Eğitim Fakültesi Dergisi*, (37):46–65.

Kasl, S. and Mahl, G. (1965). Relationship of disturbances and hesitations in spontaneous speech to anxiety. *Journal of Personality and Social Psychology*, 1(5):425–433.

Kleinsmith, A., Bianchi-Berthouze, N., and Steed, A. (2011). Automatic Recognition of Non-Acted Affective Postures. *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, 41(4):1027–1038.

Knepley, M. J., Kendall, P. C., and Carper, M. M. (2019). An Analysis of the Child Behavior Checklist Anxiety Problems Scale's Predictive Capabilities. *Journal of Psychopathology and Behavioral Assessment*, 41(2):249–256.

Konnopka, A. and König, H. (2020). Economic Burden of Anxiety Disorders: A Systematic Review and Meta-Analysis. *PharmacoEconomics*, 38(1):25 – 37.

Kwon, N., Hossain, S., Blaylock, N., O'Connell, H., Hachen, N., and Gwin, J. (2022). Detecting Anxiety and Depression from Phone Conversations using x-vectors. In *Proc. Workshop on Speech, Music and Mind*, pages 1–5.

Laukka, P., Linnman, C., Åhs, F., Pissiota, A., Örjan Frans, Faria, V., Åsa Michelgård, Appel, L., Fredrikson, M., and Furmark, T. (2008). In a Nervous Voice: Acoustic Analysis and Perception of Anxiety in Social Phobics' Speech. *Journal of Nonverbal Behavior*, 32:195–214.

Lebrun-Harris, L. A., Ghandour, R. M., Kogan, M. D., and Warren, M. D. (2022). Five-Year Trends in US Children's Health and Well-being, 2016-2020. *JAMA Pediatrics*, 176(7):e220056.

Liao, W., Zhang, W., Zhu, Z., and Ji, Q. (2005). A Real-Time Human Stress Monitoring System Using Dynamic Bayesian Network. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, pages 70–70.

Lin, W., Orton, I., Li, Q., Pavarini, G., and Mahmoud, M. (2023). Looking At The Body: Automatic Analysis of Body Gestures and Self-Adaptors in Psychological Distress. *IEEE Transactions on Affective Computing*, 14(2):1175–1187.

Lundberg, S. M. and Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Marks, I. M. and Nesse, R. M. (1994). Fear and Fitness: An Evolutionary Analysis of Anxiety Disorders. *Ethology and Sociobiology*, 15:247–261.

Mash, E. J. and Barkley, R. A. (2003). *Child psychopathology (2nd ed.).* The Guilford Press.

Masi, G., Millepeidi, S., Mucci, M., Poli, P., Bertini, N., and Milantoni, L. (2004). Generalized Anxiety Disorder in Referred Children and Adolescents. *Journal of the American Academy of Child and Adolescent Psychiatry*, 43(6):752–760.

Mauss, I. B., Wilhelm, F. H., and Gross, J. J. (2004). Is there less to social anxiety than meets the eye?Emotion experience, expression, and bodily responding. *Cognition and Emotion*, 18(5):631–662.

Mazefsky, C. A., Anderson, R., Conner, C. M., and Minshew, N. (2011). Child Behavior Checklist Scores for School-Aged Children with Autism: Preliminary Evidence of Patterns Suggesting the Need for Referral. *Journal of psychopathology and behavioral assessment*, 33(1):31–37.

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 Journal of Biomedical and Health Informatics*, 23(6):2294–2301.

Mennin, D. S., Holaway, R. M., Fresco, D. M., Moore, M. T., and Heimberg, R. G. (2007). Delineating components of emotion and its dysregulation in anxiety and mood psychopathology. *Behavior Therapy*, 38(3):284–302.

Metaxas, D. N., Venkataraman, S., and Vogler, C. (2004). Image-Based Stress Recognition Using a Model-Based Dynamic Face Tracking System. In Bubak, M., van Albada, G. D., Sloot, P. M. A., and Dongarra, J., editors, *Computational Science - ICCS 2004. Lecture Notes in Computer Science*, volume 3038, pages 813–821, Berlin, Heidelberg. Springer Berlin Heidelberg.

Mogg, K., Garner, M., and Bradley, B. P. (2007). Anxiety and orienting of gaze to angry and fearful faces. *Biological Psychology*, 76(3):163–169.

O'Dea, B., Boonstra, T. W., Larsen, M. E., Nguyen, T., Venkatesh, S., and Christensen, H. (2021). The relationship between linguistic expression in blog content and symptoms of depression, anxiety, and suicidal thoughts: A longitudinal study. *PLOS ONE*, 16(5):e0251787.

Onland-Moret, N. C., Buizer-Voskamp, J. E., Albers, M. E., Brouwer, R. M., Buimer, E. E., Hessels, R. S., de Heus, R., Huijding, J., Junge, C. M., Mandl, R. C., Pas, P., Vink, M., van der Wal, J. J., Hulshoff Pol, H. E., and Kemner, C. (2020). The YOUth study:

Rationale, design, and study procedures. *Developmental Cognitive Neuroscience*, 46:100868.

Papageorgiou, C. and Wells, A. (1999). Process and meta-cognitive dimensions of depressive and anxious thoughts and relationships with emotional intensity. *Clinical Psychology and Psychotherapy*, 6(2):156–162.

Paulus, F. W., Backes, A., Sander, C. S., Weber, M., and von Gontard, A. (2015). Anxiety Disorders and Behavioral Inhibition in Preschool Children: A Population-Based Study. *Child Psychiatry & Human Development*, 46(1):150–157.

Pearlstein, S. L., Taylor, C. T., and Stein, M. B. (2019). Facial Affect and Interpersonal Affiliation: Displays of Emotion During Relationship Formation in Social Anxiety Disorder. *Clinical Psychological Science*, 7(4):826–839.

Pease, B. and Pease, A. (2008). *The Definitive Book of Body Language: The Hidden Meaning Behind People's Gestures and Expressions*. Random House Publishing Group.

Pediaditis, M., Giannakakis, G., Chiarugi, F., Manousos, D., Pampouchidou, A., Christinaki, E., Iatraki, G., Kazantzaki, E., Simos, P. G., Marias, K., and Tsiknakis, M. (2015). Extraction of Facial Features as Indicators of Stress and Anxiety. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 3711–3714, Milan, Italy.

Petelin, G., Cenikj, G., and Eftimov, T. (2023). Towards understanding the importance of time-series features in automated algorithm performance prediction. *Expert Systems with Applications*, 213:119023.

Plaquet, A. and Bredin, H. (2023). Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*. ISCA.

Pope, B., Blass, T., Siegman, A., and Raher, J. (1970). Anxiety and depression in speech. *Journal of Consulting and Clinical Psychology*, 35(1, Pt. 1):128–133.

Pranesh, R. R. (2022). Exploring Multimodal Features and Fusion Strategies for Analyzing Disaster Tweets. In *Proceedings of the 2022 COLING Workshop: The 8th Workshop on Noisy User-generated Text (W-NUT 2022)*, pages 62–68, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning, ICML '23*, pages 28492 – 28518.

Radhakrishnan, B. L., Ezra, K., and Jebadurai, I. J. (2023). Feature Extraction From Single-Channel EEG Using Tsfresh and Stacked Ensemble Approach for Sleep Stage Classification. *International Journal of e-Collaboration*, 19:1–20.

Read, K. L., Settipani, C. A., Peterman, J., Kendall, P. C., Compton, S., Piacentini, J., McCracken, J., Bergman, L., Walkup, J., Sakolsky, D., Birmaher, B., Albano, A. M., Rynn, M., Ginsburg, G., Keeton, C., Gosch, E., Suveg, C., Sherrill, J., and March, J. (2015). Predicting Anxiety Diagnoses and Severity with the CBCL-A: Improvement Relative to Other CBCL Scales? *Journal of Psychopathology and Behavioral Assessment*, 37(1):100–111.

Rockhill, C., Kodish, I., DiBattisto, C., Macias, M., Varley, C., and Ryan, S. (2010). Anxiety Disorders in Children and Adolescents. *Current Problems in Pediatric and Adolescent Health Care*, 40(4):66–99.

Roether, C. L., Omlor, L., Christensen, A., and Giese, M. A. (2009). Critical features for the perception of emotion from gait. *Journal of Vision*, 9(6):15.

Rottenberg, J. and Hindash, A. C. (2015). Emerging evidence for emotion context insensitivity in depression. *Current Opinion in Psychology*, 4:1–5.

Ruberg, S., Menon, S., and Demanuele, C. (2023). Resolving the Credibility Crisis: Recommendations for Improving Predictive Algorithms for Clinical Utility. *Harvard Data Science Review*, 5(4).

Ryan, R. M., La Guardia, J. G., Solky-Butzel, J., Chirkov, V., and Kim, Y. (2005). On the interpersonal regulation of emotions: Emotional reliance across gender, relationships, and cultures. *Personal Relationships*, 12:145–163.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). FaceNet: A Unified Embedding for Face Recognition and Clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., and Kim, S. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH 2013)*, pages 148–152.

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M. E. P., and Ungar, L. H. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLOS ONE*, 8(9):1–16.

Setz, C., Arnrich, B., Schumm, J., La Marca, R., Tröster, G., and Ehlert, U. (2010). Discriminating Stress From Cognitive Load Using a Wearable EDA Device. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):410–417.

Simon, T., Joo, H., Matthews, I., and Sheikh, Y. (2017). Hand Keypoint Detection in Single Images using Multiview Bootstrapping. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4645–4653, Los Alamitos, CA, USA. IEEE Computer Society.

Sirois, F. M. (2016). Chapter 4 - Procrastination, Stress, and Chronic Health Conditions: A Temporal Perspective. In Sirois, F. M. and Pychyl, T. A., editors, *Procrastination, Health, and Well-Being*, pages 67–92. Academic Press, San Diego.

Szechtman, H. and Woody, E. Z. (2004). Obsessive-compulsive disorder as a disturbance of security motivation. *Psychological Review*, 111:111–127.

Tartter, V. (1980). Happy talk: Perceptual and acoustic effects of smiling on speech. *Perception & Psychophysics*, 27(1):24–27.

van Bommel, D. M. H., van der Giessen, D., van der Graaff, J., Meeus, W. H. J., and Branje, S. J. T. (2019). Mother–Adolescent Conflict Interaction Sequences: The Role of Maternal Internalizing Problems. *Journal of Research on Adolescence: the official journal of the Society for Research on Adolescence*, 29(4):1001–1018.

Vaughan, J., Coddington, J. A., Ahmed, A. H., and Ertel, M. (2017). Separation Anxiety Disorder in School-Age Children: What Health Care Providers Should Know. *Journal of Pediatric Health Care*, 31(4):433–440.

Verhulst, F., Ende, J. V. D., and Koot, H. (1996). *Handleiding voor de CBCL/4-18 (Manual for the CBCL/4-18)*. Erasmus University/Department of Child and Adolescent Psychiatry, Sophia Children's Hospital/AZR/EUR, Rotterdam, The Netherlands.

Watson, D. (2005). Rethinking the mood and anxiety disorders: a quantitative hierarchical model for DSM-V. *Journal of abnormal psychology*, 114(4):522–536.

Wei, S.-E., Ramakrishna, V., Kanade, T., and Sheikh, Y. (2016). Convolutional Pose Machines. In *CVPR*.

Woody, E. Z., Lewis, V., Snider, L., Grant, H., Kamath, M., and Szechtman, H. (2005). Induction of compulsive-like washing by blocking the feeling of knowing: an experimental test of the security-motivation hypothesis of Obsessive-Compulsive Disorder. *Behavioral and Brain Functions*, 1:11.

Woody, E. Z. and Szechtman, H. (2005). Motivation, Time Course, and Heterogeneity in Obsessive-Compulsive Disorder: Response to Taylor, McKay, and Abramowitz (2005). *Psychological Review*, 112(3):658–661.

Yarkoni, T. (2010). Personality in 100,000 Words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, 44(3):363–373.

Yoon, J., Kang, C., Kim, S., and Han, J. (2022). D-vlog: Multimodal Vlog Dataset for Depression Detection. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, volume 36, pages 12226–12234.

Yu, Y., Li, Q., and Liu, X. (2023). Automatic anxiety recognition method based on microblog text analysis. *Froentiers in Public Health*, 11.

Zhai, J. and Barreto, A. (2006). Stress Recognition Using Non-invasive Technology. In *Proceedings of the 19th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2006)*, pages 395–401, Melbourne Beach, Florida, USA.

Zhang, M., Zhou, Y., Xu, X., Ren, Z., Zhang, Y., Liu, S., and Luo, W. (2023). Multi-view emotional expressions dataset using 2D pose estimation. *Scientific Data*, 10(1).

Zhao, N., Jiao, D., Bai, S., and Zhu, T. (2016). Evaluating the Validity of Simplified Chinese Version of LIWC in Detecting Psychological Expressions in Short Texts on Social Network Services. *PLOS ONE*, 11(6):1–15.