Towards Improving Transparency in Multimodal Affect Perception

Torsten Wörtwein

CMU-LTI-23-015

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Thesis Committee:
Louis-Philippe Morency (Chair)
Jeffrey F. Cohn (University of Pittsburgh)
Florian Metze (Carnegie Mellon University)
Emily Mower Provost (University of Michigan)

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in Language and Information Technologies

Keywords: Multimodal, Transparency, Data, Reliability, Model Mechanics, Personalization, Modality Importance, Cross-Modal Interactions, Unique and Redundant Contributions

Abstract

Affective computing has emerged as a core research topic in artificial intelligence (AI) with broad applications in healthcare. For example, affective technologies can be used as a decision-support tool that quantifies behaviors related to emotions and affective states, which helps clinicians in their assessment of mood disorders such as depression. For these AI applications to be used and trusted, we need to focus on improving their *transparency*. Transparency is the degree to which users have information about a model's internal mechanics and the reliability of its output. We expand this definition to include information about the data used to train a model, as data influences what a model learns. These three components of transparency (data, reliability, and internal mechanics) are studied in this thesis in two main research thrusts geared towards improving transparency for machine learning practitioners.

In our first thrust on general transparency, we explore three challenges related to data and reliability transparency, where two focus on data transparency and one on reliability transparency. The first challenge, referred to as *population-level data transparency*, is analyzing data patterns across people to understand which patterns a model will likely learn. For this, we used statistical approaches to analyze patterns between how people speak and the symptom severity of psychosis. The second challenge, referred to as *reliability transparency*, estimates how accurate a model's output is to enable better risk management, as the output might not always be correct. We created approaches to efficiently estimate the reliability of a primary model using a secondary model that learns when the primary model makes mistakes. The third challenge, referred to as *personalized data transparency*, is separating person-specific patterns from patterns shared across people and analyzing them. We efficiently integrated neural networks with mixed effect models, a statistical modeling approach that can separate these two types of patterns.

In our second research thrust, we focus on the third component of transparency, internal mechanics. More specifically, we focus on the mechanics of *multimodal* models as affect is expressed through multiple modalities, such as visually smiling and audibly laughing. The first challenge, referred to as *modality importance transparency*, quantifies how much a model focuses on modalities to derive its output, which is a proxy for how important each modality is. We created a model that not only quantifies the modality importance but also reflects how informative humans perceive each modality. The second challenge, referred to as *multimodal interaction transparency*, quantifies interactions between three modalities, including both bimodal and trimodal interactions. Our approach separated unimodal, bimodal, and trimodal interactions by prioritizing simpler interactions over more complicated interactions, e.g., unimodal prioritized over bimodal. The third challenge, referred to as *modality contribution transparency*, factorizes a modality's unique contributions, what can be explained by only one modality, from what can redundantly be contributed by multiple modalities. Our approach used correlational measures to define these contributions and the learned factorization correlated with human judgments.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Affective computing is an interdisciplinary field at the intersection of computer science, psychology, and cognitive science. It aims to create computational models that perceive, understand, and express affective states such as emotions [37, 130, 162]. Perceiving affective states will make today's artificial intelligence (AI) applications more emotionally and socially aware. It can enable future AI applications assisting in healthcare [24, 39, 168], for example, providing information about perceived mood disorder symptoms based on what a person says [46], their visual appearance [27, 200], or even their smartphone usage [112]. People trusting affective AI applications is integral in making these future AI applications a reality as people tend to trust them less and might, therefore, not use them [56, 60]. To increase trust in affective AI applications in the future, we focus on a core dimension of it, transparency [58]. Transparency is the degree to which users have information about a model's internal mechanics and its reliability [72]. We extend this definition to include information about the data used to train a model, as data influences what a model learns. As visualized in Figure 1.1, we focus on two main research thrusts for transparency in affect perception. In the first thrust, we focus on challenges covering data and reliability, while model mechanics are studied later in the second thrust with a focus on multimodal models for affect perception. Both thrusts focus on improving transparency of affect perception applications to hopefully make them, in the future, more trustworthy. In this thesis, we focus on

improving transparency for machine learning practitioners who have a technical understanding of how computational models function.

The first thrust on general transparency covers data and reliability, split into three challenges: uncovering patterns in the data at the population-level (population-level data transparency), determining how reliable the model's output is (reliability transparency), and uncovering person-specific patterns in the data (personalized data transparency). The first challenge is gaining insights into patterns in the data to better understand the data before training affect perception models (**population-level data transparency**). We start by focusing on population-level patterns, e.g., patterns shared by people, as we typically want a model to perform well across many people. Analyzing these patterns can provide insights into what a model is likely to learn [113]. The second challenge is being transparent about how reliable the model's output is (**reliability transparency**). While we like the model's output always to be correct, it might be incorrect in some cases. Knowing how reliable a model's output is, enables better risk management, which can especially be important in critical areas such as healthcare [5, 213]. A model's unreliable output can be disregarded or might still be used if it is within an acceptable margin of error. The third challenge extends population-level data transparency by separating population-level and person-specific patterns (**personalized data transparency**). Many patterns, for example, behaviors and regular events, are consistently expressed across people (population-level behaviors), such as people being less stressed on weekends, but some patterns are specific to individual people, such as a person's weekly events [97, 160]. To better understand these two types of patterns, we need to account for individual differences in an interpretable manner to improve transparency. Personalized models that have both population-level and person-specific parameters have the potential to learn both types of patterns in an interpretable manner [41].

Since affective states are expressed through behaviors in multiple modalities [25, 152], such as through smiling in the visual modality, laughing in the acoustic modality, and saying "like it" in the language modality, our second thrust on multimodal transparency focuses on model

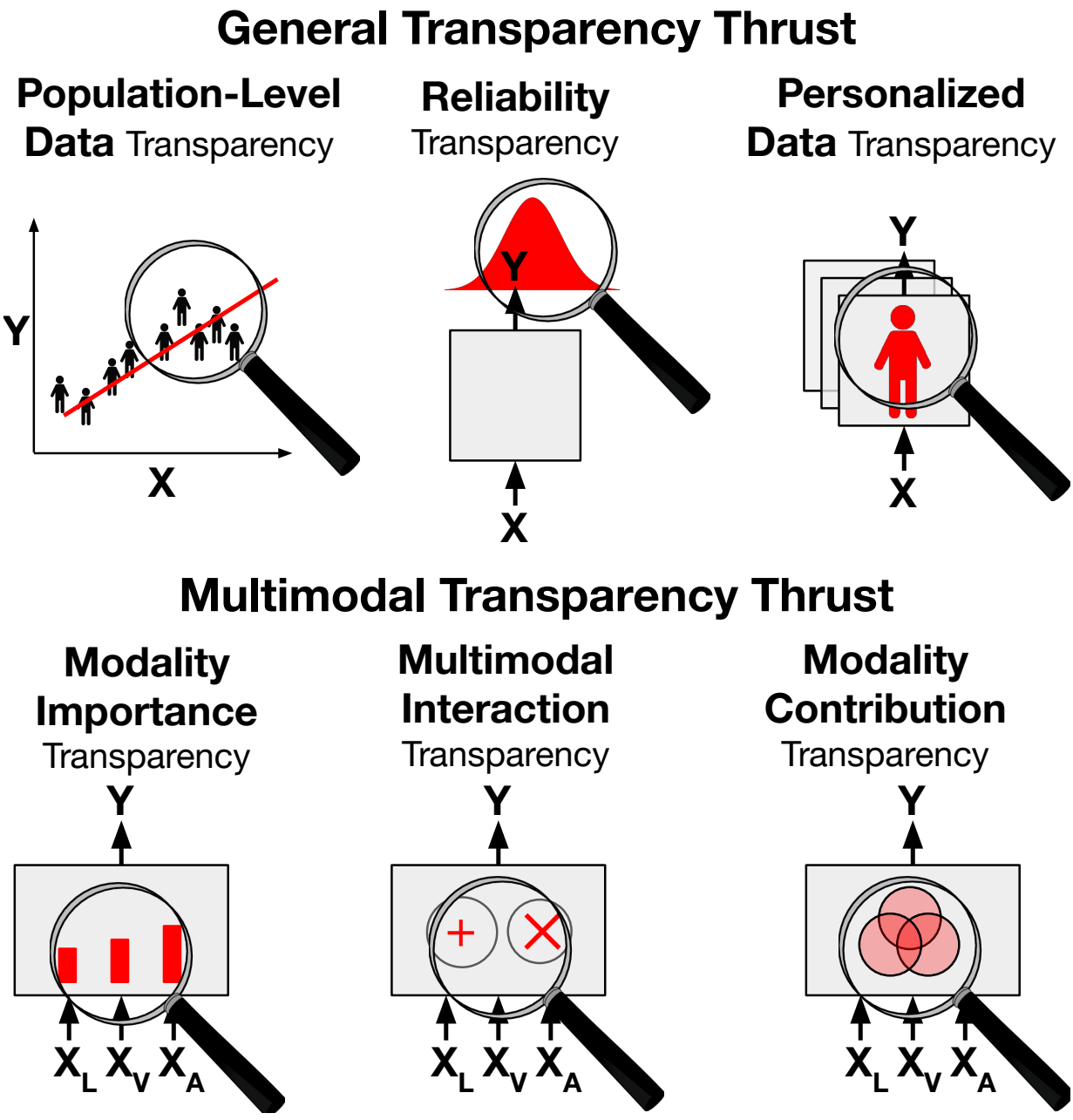


Figure 1.1: The general transparency thrust focuses on data and reliability, while the multimodal transparency thrust focuses on multimodal model mechanics.

mechanics in the multimodal setting when a model uses multiple modalities. We focus on quantifying the importance of modalities (modality importance transparency), analyzing which interactions a model learns between modalities (multimodal interaction transparency), and analyzing what a modality uniquely contributes to the model's output and what is redundantly contributed by multiple modalities (modality contribution transparency).

First, knowing how much a model focuses on a modality can inform machine learning practitioners how important the modalities of the current input are for the model's output (**modality importance transparency**). For example, the vision modality might be more important when a person smiles. Beyond quantifying modality importance, if a model's modality importance is similar to how humans perceive the importance of modalities, the model might perform better and be perceived as more trustworthy in the future [58]. Second, information from one modality can interact with information from another modality (**multimodal interaction transparency**). If someone says "this is so ridiculous", they might be perceived as very negative when also frowning or as very positive when smiling. In this example, the language modality amplifies the directionality expressed in the vision modality. The challenge we address is defining types of multimodal interactions and quantifying how much they contribute to the model's output. Finally, a modality might contribute information that is provided by no other modality (unique contribution) or information that is also provided by another modality (redundant contribution; **modality contribution transparency**). For example, smiling and laughing provide similar information indicating a positive state. Our goal is to inform machine learning practitioners in an interpretable manner about the unique and redundant modality contributions.

1.1 General Transparency Thrust

In the thrust on general transparency, we concentrate on three challenges focused primarily on data and reliability transparency. We analyze population-level data patterns to understand patterns consistent across people (population-level data transparency), measure a model's reliability

to indicate how accurate a model's output is (reliability transparency), and quantify personalized data patterns by analyzing which adjustments models learn for different people (personalized data transparency).

1.1.1 Population-Level Data Transparency

The goal of data transparency is to gain insights into the relationship and structure of the data. Analyzing these relationships also gives insights into what a model will likely learn [113]. We propose to start by reviewing statistical approaches on population-level data transparency to analyze general trends across an entire dataset, e.g., trends consistent across people. This challenge has been studied in statistics, especially in the linear context. We propose to take advantage of these approaches and empirically apply them to a specific case where transparency is important, namely to decision-support technologies in healthcare. Decision-support technologies aim to quantify symptom-relevant behavior to assist clinicians in their assessment.

Population-level data transparency is especially important for small datasets, which is often the case for clinical datasets, as an input might, by coincidence, have a strong relation to the ground truth. In Chapter 2, we explore the relation between acoustic features and the severity of symptoms of psychosis, such as schizophrenia. The long-term goal is to create decision-support systems that help clinicians assess symptoms while conducting semi-structured interviews with their clients. **RQ 1.1:** *Which acoustic behaviors have a relationship to severities of psychosis symptoms that are consistent across people?*

1.1.2 Reliability Transparency

In critical applications, such as healthcare, it is important to know not only the model's output but also how close the output is to the ground truth. This is often referred to as reliability transparency. While we would like for the model's output to always be correct, the output will sometimes be wrong. Having an estimate of how reliable the model output is can enable better

risk management. For example, the model output might be relied upon only when it is within an acceptable margin of error.

An example of a reliability estimate for regression models are intervals around the model's output where their width indicates how reliable the output is. When the interval is small, the output is more reliable, whereas when the interval is larger, the model's output is less reliable. In Chapter 3, our goal is to create a reliability estimate for the regression task of estimating facial action units. Facial action units describe the contraction and relaxation of facial muscles and play a crucial role in inferring someone's emotions which is important for assessing mood-related symptoms [47]. The first challenge of creating a reliability estimate for facial action units is defining the properties of the reliability estimate, for example, as small as possible intervals that still contain the ground truth at a targeted rate. The second challenge is learning this reliability estimate. **RQ 1.2:** *How can we derive small prediction intervals containing the ground truth at a target rate?*

1.1.3 Personalized Data Transparency

We go beyond our initial study of population-level data patterns by separating behavioral patterns into a) patterns that are consistent across people (population-level patterns) and b) patterns that are specific to some individuals. This can be achieved computationally by having person-generic parameters (for the population-level patterns) and person-specific parameters to represent person-specific patterns. Our secondary goal is to learn those parameters while keeping some interpretability to analyze similarities and differences between people. For example, does an affect recognition model learn different person-specific parameters for people who experience symptoms of mood disorders compared to people who do not and what do those differences represent? Such person-specific models are important when person-specific differences are prominent, for example, while self-assessment ratings might be consistent within a person, they might not be consistent across people [145], or the input features might be very similar within a person but

very different across people [160], for example, people use their phones very differently.

One popular statistical framework with both person-generic and person-specific parameters is the framework of mixed effect models [41]. In Chapter 4, we present our work combining mixed effect models with neural networks. This allows us to efficiently train personalized models that perform well while separating complex person-generic and person-specific parameters. We demonstrate that it is possible to analyze the learned person-specific parameters of our proposed model to identify differences in temporal transition patterns between consecutive affective states for people experiencing mood disorder symptoms and for people who do not. **RQ 1.3:** *How can we learn a model quantifying person-generic and person-specific patterns?*

1.2 Multimodal Transparency Thrust

The second thrust focuses on challenges providing more information about internal model mechanics. In particular, we focus on three model mechanics challenges of multimodal models. We focus on quantifying the importance of modalities in a human-compatible manner (modality importance transparency), analyzing which multimodal interactions a model learns between modalities (multimodal interaction transparency), and factorizing the contributions from each modality in unique and redundant contributions (modality contribution transparency).

1.2.1 Modality Importance Transparency

The first challenge of multimodal transparency is quantifying how important each modality is for the model's output. For example, does the model focus mostly on language or how a person speaks? Going beyond quantifying modality importance, we are interested in models whose modality importance aligns with how important humans perceive the modalities. Quantifying modality importance not only improves transparency but human-like modality importance might even, in the future, elicit more trust.

A first step of modality importance transparency is to study how humans judge the informativeness of modalities for a specific task, in our case, for affect perception. The second step is guiding the model's modality importance using the human modality judgments while not degrading the model's performance. The resulting model brings us closer to modality transparency by learning human-like modality importance. This research is presented in Chapter 5 at the example of four affective states that have shown a relation to a future onset of depression [167]. **RQ 2.1:** *How can we learn a model that quantifies its focus on modalities in a human-inspired manner?*

1.2.2 Multimodal Interaction Transparency

When simultaneously observing multiple modalities, affective states are often inferred through interactions between these modalities. For example, when a person says "this is ridiculous" the person might be in a very negative or very positive state. If we simultaneously also observe that the person is smiling, the person might be in a very positive state. We want to quantify which interactions a model learns between modalities and how much they influence the model's output.

To study multimodal interaction transparency, the primary challenge is to mathematically characterize and separate the different types of interactions between modalities, e.g., bimodal and trimodal interactions. In Chapter 6, we explore and evaluate this challenge on six datasets covering multiple affective phenomena. The secondary challenge is conducting human studies to test whether the learned interactions are similar to the interactions humans use. **RQ 2.2:** *How can we factorize a model output to quantify which unimodal, bimodal, and trimodal interactions are necessary for the output?*

1.2.3 Modality Contribution Transparency

Modalities often contribute unique information that no other modality contains. In other cases, modalities can have redundant information that is also conveyed through other modalities. An example of redundant information might be visually smiling while also audibly laughing. The

goal of this technical challenge is learning a model that factorizes its output to separate contributions that are unique to a modality and that are redundant between modalities. Our proposed approach can help machine learning practitioners inspect what a modality uniquely contributes and what overlaps between modalities.

The first sub-challenge when factorizing unique and redundant contributions is defining them mathematically so that we can also test whether we achieve the desired factorization. The second sub-challenge is that we might need to incentivize a model to derive redundant contributions from multiple modalities, as intuitively, a model might not need to derive the same information twice. In Chapter 7, we propose a model that defines unique and redundant contributions using correlational measures and that is incentivized to learn redundant contributions. Our proposed model explicitly factorizes these contributions, making it easier for machine learning practitioners to inspect how these two types of contributions influence the model's output. Finally, we compare the outcome of human judgments with the learned factorization of our model. **RQ 2.3:** *How can we factorize what is uniquely contributed by a modality despite modalities providing redundant information for a task?*

1.3 Contributions

1. **Population-Level Data Transparency (Chapter 2) [207]**

   (a) We analyzed how acoustic features, such as voice quality consistency, variation of speech rate and volume, vowel space, and a parameter of glottal flow, relate to the severity of symptoms of psychosis as measured by the Brief Psychiatric Rating Scale (BPRS) [136]. We used a nonparametric rank correlation measure between features and the rating scale to interpret their relationship. Further, we showed that specific acoustic descriptors could track a patient's state from admission to discharge.

   (b) The machine learning experiments using Support Vector Machines demonstrated that

a significant amount of variance in the BPRS items could be predicted.

(c) Towards a fully automatic system, we evaluated our model's performance when replacing the manual speaker segmentation with an automatic speaker segmentation.

2. **Reliability Transparency (Chapter 3) [206]**

   (a) We operationalized reliability in regression as the absolute error between a model's prediction and the ground truth. Our two proposed approaches used a secondary model to predict the reliability of a primary predictive model. Our first approach leveraged the assumption that similar observations are likely to have similar reliability and predicts reliability with a non-parametric method. Our second approach is a secondary model directly predicting the primary predictive model's reliability.

   (b) We observed that approaches that directly predict the reliability generally perform better than approaches that indirectly estimate reliability.

   (c) Using our reliability estimate, we derived intervals using the framework of inductive conformal prediction. The intervals based on our proposed reliability methods are smaller than other reliability methods.

3. **Personalized Data Transparency (Chapter 4) [210]**

   (a) We efficiently integrated mixed effect models in neural networks to learn person-generic and person-specific parameters.

   (b) Our approach demonstrated performance increases across six datasets in a within-person testing setup. Our model performed well, even in the extreme case of very little data per person.

   (c) We showcased that the learned person-specific parameters can be interpreted and analyzed. We learned person-specific transition patterns between affective states and observed significant differences in those patterns between people affected by depression and people not affected by depression.

4. **Modality Importance Transparency (Chapter 5) [208]**

   (a) We performed a human study to understand better which subset of modalities people find informative when confirming four affective states. Our studies showed that humans can reliably annotate which modalities are informative.

   (b) We observed that models guided by the human annotations significantly improved interpretability on unseen data, i.e., the models attend to modalities similarly to how humans rate the modality informativeness, while at the same time showing a slight increase in predictive performance.

   (c) Replacing the learned modality influence with the human annotations yields a significant performance improvement in guided models demonstrating that users could potentially improve the predictive performance by correcting the model.

5. **Multimodal Interaction Transparency (Chapter 6) [209]**

   (a) We proposed Multimodal Residual Optimization (MRO) to separate unimodal, bimodal, and trimodal interactions in a multimodal model. Empirically, we observed that MRO successfully separates unimodal, bimodal, and trimodal interactions while not degrading predictive performance.

   (b) Since MRO-optimized models prioritize simpler interactions over more complex interactions, they can be used to debug whether more complex interactions are required for a task on a specific dataset.

   (c) We complemented our empirical results with a human perception study and observed that MRO's learned multimodal interactions align with human judgments.

6. **Modality Contribution Transparency (Chapter 7)**

   (a) We proposed a covariance-based optimization method to learn factorized models that express their predictions as the sum of unique and pairwise redundant contributions. The unique contributions represent what is specific to a modality, and the pairwise

redundant contributions represent the commonalities between the two modalities.

(b) On eight affective datasets, we observed that our approach maintained performance while also achieving its factorization. We further, observed that our approach is more robust to missing modalities as it derives redundant contributions from multiple modalities.

(c) We demonstrated that the learned factorization has relationships to human judgments on three datasets, indicating that our approach has the potential to improve interpretability.

Chapter 2

Population-Level Data Transparency

Before training a model, we want to understand general patterns of the data in an interpretable way (**population-level data transparency**). This analysis is especially important for clinical datasets for two reasons. First, transparency is more important in critical domains such as healthcare. Second, many clinical datasets' major goal is to better understand a condition. In this chapter, we explore the relationship between acoustic features and the severity of symptoms of psychosis to make the patterns likely learned by a model more transparent. The work in this chapter was published at the Interspeech conference [207][1].

Psychosis disorders, such as bipolar and schizophrenia, severely impact social functions [86, 129]. Psychosis affects how we speak [59] and how we express ourselves with facial expressions [184, 193]. Thus, medical assessments have included speech-related descriptors for a long time [119, 185]. Such differences in speech might be difficult for humans to assess objectively but can be captured by computational acoustic descriptors. This brings the opportunity to support clinicians in assessing symptoms and allows for better decision-making.

Work on this topic is limited, and many computational acoustic descriptors have not been studied with clinical patients suffering from psychosis, including articulation rate, vowel space [164],

[1]The published paper is titled "Computational Analysis of Acoustic Descriptors in Psychotic Patients" and is available at `https://doi.org/10.21437/Interspeech.2017-466`.

speech volume, and glottal flow parameters. While an early measure of voice tenseness was associated with schizophrenia [181], more recent robust measures of voice quality such as peak slope [79] have not been investigated. Psychosis is characterized by positive symptoms, exaggerations of normal functions (e.g., grandiosity and hallucination), and negative symptoms, declines of normal functions (e.g., emotional withdrawal and motor retardation) [83, 133]. While speech-related behaviors of negative symptoms have been studied through work on depression and PTSD, positive symptoms have not been paid much attention in computational studies.

In this chapter, we perform a computational study of acoustic descriptors to better understand psychosis and its positive symptoms. This study is performed on a dataset of semi-structured interviews between clinicians and clinical patients experiencing symptoms of psychosis. We investigate the following questions.

Q1: What are the acoustic descriptors related to overall psychosis severity? What are the acoustic descriptors related to specific positive symptoms?

Q2: Can we estimate positive symptoms and overall severity of psychosis with acoustic descriptors?

2.1 Related Work

While the research community has studied how individuals suffering from psychosis perceive, e.g., speech [103] and emotions [92], there is less research on whether they express themselves differently through speech and language.

In a within-patient study of patients diagnosed with schizophrenia, a decrease of the fundamental frequency (pitch) and a better pronunciation of vowels, i.e., the first and second formants were closer to a reference pronunciation, were observed at discharge compared to admission [181]. In the same study, a tendency toward a more tense voice was observed for patients with schizophrenia, while the opposite has been seen in depressed patients at discharge [181].

Compared to this within-patient study, this chapter investigates more acoustic descriptors, including a more robust estimator of voice tenseness, between patients. An exhibition of inadequate speech behaviors, e.g., in volume, rate, and pitch variation, was found in children diagnosed with schizophrenia compared to a control group of the same size and demographic [59]. The same study observed that children with schizophrenia were not identifiable by a single speech behavior but often deviated more from the norm on many speech-related behaviors, e.g., speaking too loudly or quietly. In contrast to this chapter, all speech- and language-related behaviors were manually assessed.

More recently, second formant variation was linked to the severity of negative symptoms [83], e.g., blunted affect and emotional withdrawal, among patients with schizophrenia [30]. Besides the second formant, the first formant showed a similar but not statistically significant trend. In other disorders, more severe negative symptoms have been linked to a smaller vowel space for self-reported PTSD compared to a control group [164], i.e., similar first and second formants for different vowels, and to a more tense voice in self-reported depression [163]. Even though these two studies are based on many individuals, the individuals are not clinically diagnosed and are not hospitalized, i.e., we expect milder symptoms.

We contrast with previous research by investigating robust computational measures of acoustic descriptors, which have not been studied previously in psychosis. Besides establishing relationships between acoustic descriptors and symptoms, we also investigate the automatic estimation of psychosis severity, focusing on positive symptoms.

2.2 Methods

Our dataset consists of audio and video recordings of 29 semi-structured interviews between clinicians and 20 unique individuals who are experiencing symptoms of psychosis and are hospitalized in an inpatient service at a psychiatric hospital. The semi-structured interview protocol was designed to reflect the daily clinical encounters between patients and their clinicians. This

dataset is a significant expansion of a previously-published dataset used to study facial expressions of patients suffering from psychosis [193]. Our new dataset includes multiple interviews with the same patient from admission to discharge to analyze temporal changes. While most patients are diagnosed with schizophrenia, some are diagnosed with bipolar or mania. The average duration of these interviews is 8.33 minutes (SD=4.22). Of the 29 interviews, 17 interviews are with male and 12 interviews are with female patients. They are recorded using head-mounted microphones and two webcams facing the patient's and clinician's upper body.

After each interview, the patients are assessed using the 24-item version of the Brief Psychiatric Rating Scale (BPRS) [119]. It was designed to measure the severity of relatively independent symptoms often found in psychiatric disorders [136]. The BPRS total score (M=42.4, SD=13.6) is the sum of all BPRS items, which are scored on an ordinal scale from 1 (not present) to 7 (extremely severe). Therefore, BPRS total ranges theoretically from 24 to 168. In our analysis, we focus on the total score as well as on positive items [133]. Further, we omit BPRS items that do not vary in our patient population (SD$<$1). This leaves the following six positive BPRS items: *grandiosity*, *elevated mood*, *hallucination*, *unusual thoughts*, *excitement*, and *motor hyperactivity*.

2.2.1 Speaker Diarization

A first step when computing acoustic descriptors is speaker diarization. Our experimental setup includes head-mounted microphones designed to reduce cross-over speech. Even in these good recording setups, we hear the other person talking. This chapter explores manual annotations and an automatic diarization for speaker diarization. The experiments with manual diarization allow studying computational acoustic indicators in the ideal case. Experiments using automatic diarization allow us to get closer to our goal of building decision-support tools.

Our automatic speaker diarization is based on the time delay of arrival (TDOA). Since we have only two speakers, each wearing a head-mounted microphone, we can distinguish speakers

by TDOA between the two audio signals as estimated by the generalized cross-correlation with phase transform [90]. The patient and clinician, who are spatially separated by the recording setup, are approximately 3 meters apart. Therefore, we expect a delay of 8ms between the audio signals. TDOA might not be reliably estimated when the other microphone does not pick up an audio signal. Therefore, we rely on TDOA only if a voice is detected [43] in both audio signals. If TDOA is less than 5ms, and if a voice is detected in both signals, we assume that both patient and clinician are speaking. If speech is detected in one signal only, we assume that the corresponding person is speaking. A recording problem during six interviews made it impossible to recover the audio signal from the clinician's microphone. For this reason, experiments with automatic diarization are performed on 23 interviews.

We calculate the automatic approach's diarization error rate (DER) over all interviews based on the manual annotations. It is suggested to use a 250ms no-score collar around the annotated segment boundaries [134]. However, this would remove a significant amount of our annotations. Without this collar, we reach a DER of 20.10%, which is still comparable to DERs in similar settings [64] with the collar.

2.2.2 Computational Acoustic Descriptors

As mentioned in Section 2.1, limited prior work has investigated computational acoustic descriptors in interviews with patients suffering from psychosis. We use descriptors inspired by work on depression and PTSD behavior analyses [163, 164]. Our descriptors include the first Mel-frequency cepstral coefficient ($MFCC_0$) as a measure of volume, vowel space [164], formants (F_1 and F_2), fundamental frequency, voice quality descriptors from COVAREP [35], and articulation rate from Praat [33].

Since many descriptors can only be estimated for voicing, we remove parts of the audio signals where the patient is not voicing [42]. On average, we have 3.14 minutes (SD=2.08) of voicing for the patients. Then, we compute descriptive statistics of our descriptors, i.e., me-

dian and interquartile range (IQR). These two statistics are used because they are robust against outliers, which might occur due to the diarization.

Articulation rate is the ratio of the number of syllables and the phonation time over all speech segments according to the diarization. The variation of articulation rate is the IQR of the articulation rates per speech segment. We do not calculate the median for fundamental frequency or formants because they have only been shown to be indicative in within-patient studies [181]. Speech volume (median of MFCC_0) is not used because it has been shown to be sensitive to the recording environment. This leads to 12 acoustic descriptors for computational analysis. All descriptors are mean-centered and normalized by their standard deviation.

2.2.3 Automatic Estimation of BPRS Items

In Q2 we want to estimate BPRS items. Since we have not too many interviews, we choose linear support vector regression (SVR) to estimate BPRS items. Experiments are performed in a speaker-independent fashion using the leave-one-patient-out method. Hyperparameters of the linear SVR, including descriptor selection, are determined automatically using a nested leave-one-patient-out validation on the training set.

For each training partition, we find a suitable subset of descriptors by conducting a greedy forward selection on the minimizing criteria $-corr(Y, \hat{Y})$ (Pearson's linear correlation), where $\hat{Y}$ are the estimated scores and Y the corresponding ground truth scores. The maximum number of descriptors is restricted to five descriptors to prevent over-fitting. During the descriptor selection, we validate the SVR's penalty parameter C (between 0.001 and 100) with Bayesian optimization [172], which uses a Gaussian process to model $-corr(Y, \hat{Y})$ of the nested leave-one-patient-out validation.

We use two evaluation metrics in our experiments: Pearson's correlation coefficient (r) and the mean absolute error (MAE). Before these two metrics are calculated, estimations are clipped to valid BPRS scores. For comparison, we calculate the MAE of a naive mean estimation

(MAE_{naive}) as a baseline, where the mean is calculated on each training fold.

2.3 Results and Discussion

In this section, we present our experiments to study the previously introduced research questions: (Q1) correlation analysis of acoustic descriptor with positive BPRS items and BPRS total score, and (Q2) models to estimate BPRS items based on computational descriptors.

2.3.1 Acoustic Descriptors of Psychosis Symptoms (Q1)

We investigate acoustic descriptors related to the BPRS total score and positive BPRS items. Since the relationship between acoustic descriptors and BPRS might be non-linear, we use Spearman's rank correlation coefficient ρ. All descriptors are based on manual diarization to use all 29 interviews. Our significant correlation results ($p < 0.05$) are summarized in Figure 2.1 and Table 2.1. In the next paragraphs, we discuss these results.

Speech volume: IQR of $MFCC_0$, a measure related to the variation of speech volume, correlates positively with BPRS total score. In line with our result, it has been observed that patients diagnosed with schizophrenia deviate more from the norm for speech volume and other descriptors [59]. Individual positive items show no correlation with a variation in speech volume.

Articulation rate: We do not observe any correlations between articulation rate and BPRS. However, IQR of articulation rate correlates negatively with the BPRS total score and many positive items. For children with schizophrenia, it was observed that they have a more excessive variation in speech rate [59]. We found the opposite to be the case for our dataset.

Glottal flow: Quasi-open-quotient (QOQ) [63] measures the ratio of the opening time of the vocal folds. The Median and IQR of QOQ correlate negatively with the BPRS total score. Larger BPRS total scores tend to be related to a smaller QOQ range and a shorter opening time of the vocal folds. The range of QOQ is often reduced for people with functional dysphonias [63], in com-

Table 2.1: Significant correlations of acoustic descriptors for all positive BPRS items ($p < 0.05$).

Positive symptom	Acoustic descriptor	ρ
Hallucinations	Median PS	0.43
	IQR F_1	-0.40
	IQR F_2	-0.37
Unusual thoughts	IQR PS	0.56
	Median PS	0.52
	Vowel space	0.41
	IQR F_2	-0.45
	IQR F_1	-0.55
Elevated mood	Median PS	0.40
	IQR F_2	-0.38
	IQR F_1	-0.47
	IQR articulation rate	-0.62
Grandiosity	IQR articulation rate	-0.40
	IQR F_1	-0.44
	IQR F_2	-0.44
Excitement	Vowel space	0.40
	IQR articulation rate	-0.52
Motor hyperactivity	Vowel space	0.45
	Median QOQ	-0.39
	IQR articulation rate	-0.46

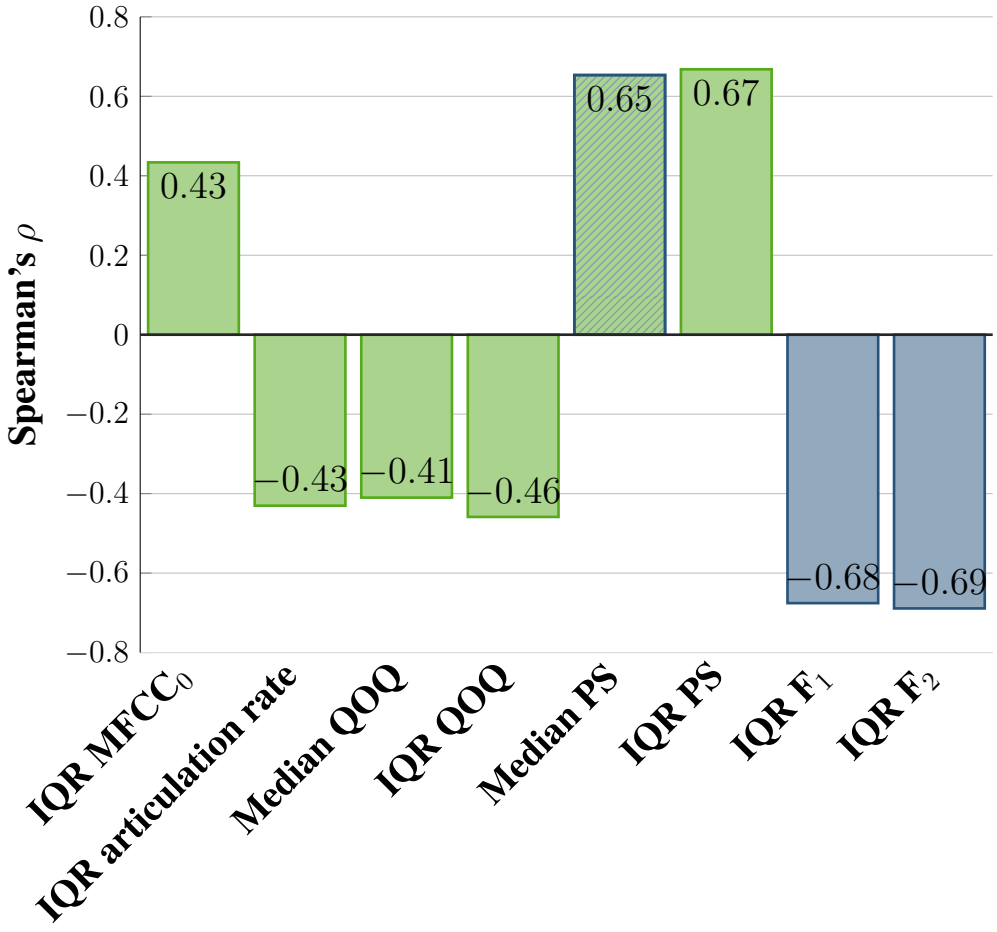


Figure 2.1: Correlations between acoustic descriptors and BPRS total score on our dataset. Descriptors colored in blue have also been studied in previous work. Median PS is marked differently because an early descriptor of tenseness [100] has been investigated, but the opposite result was observed [181].

bination with a low QOQ speaking loudly requires more effort and sounds more "stalled" [63].

Voice quality: Peak slope (PS) [79], a voice quality descriptor related to the breathy-modal-tense spectrum, correlates positively with many BPRS items. This indicates a more tense voice for more severe symptoms. A more tense voice was associated with clinical [181] and self-reported [163] depression, but the opposite was reported for patients with schizophrenia [181]. While the contradicting study observed less tense voice based on an early tenseness measure [100],

this change was not statistically significant. IQR of PS also correlates positively with many BPRS items. A variation in breathy-modal-tense voice seems to be as indicative as the actual voice quality. A positive correlation of IQR of PS indicates that more severe symptoms tend to be associated with less consistent voice quality. The consistency of voice quality has to our knowledge never been studied computationally in any clinical study.

Vowel space: It was found that vowel space correlates negatively with self-reported depression [164]. Depression is mainly characterized by negative symptoms. For positive items, e.g., *excitement* and *motor hyperactivity*, we observe a positive correlation with vowel space. This indicates that it is important to analyze positive and negative symptoms separately since effects could average out, i.e., we do not observe a correlation with overall symptoms.

Formants: IQR of the first two formants correlates negatively with almost all BPRS items, i.e., a smaller range correlates with more severe symptoms. This has previously been observed for the variation of the second formant and indicated for the first formant for negative symptoms [30]. The first and second formant are mainly influenced by the tongue's position and the jaw's extension. It could be argued that patients suffering from psychosis with more severe symptoms do not move their tongue [30] and mouth as much.

We further investigate within-patient differences from admission to discharge for one patient. While some acoustic descriptors, such as pitch in in-between studies, might seem indicative due to, e.g., a gender bias in a dataset, they might not be indicative in within-studies. Therefore, we would like to see previously unstudied descriptors behave similarly over time as BPRS total within a patient from admission to discharge. We plot the two peak slope statistics because they have the strongest correlations with BPRS total. As shown in Figure 2.2, peak slope's statistics behave similarly to BPRS total for this patient.

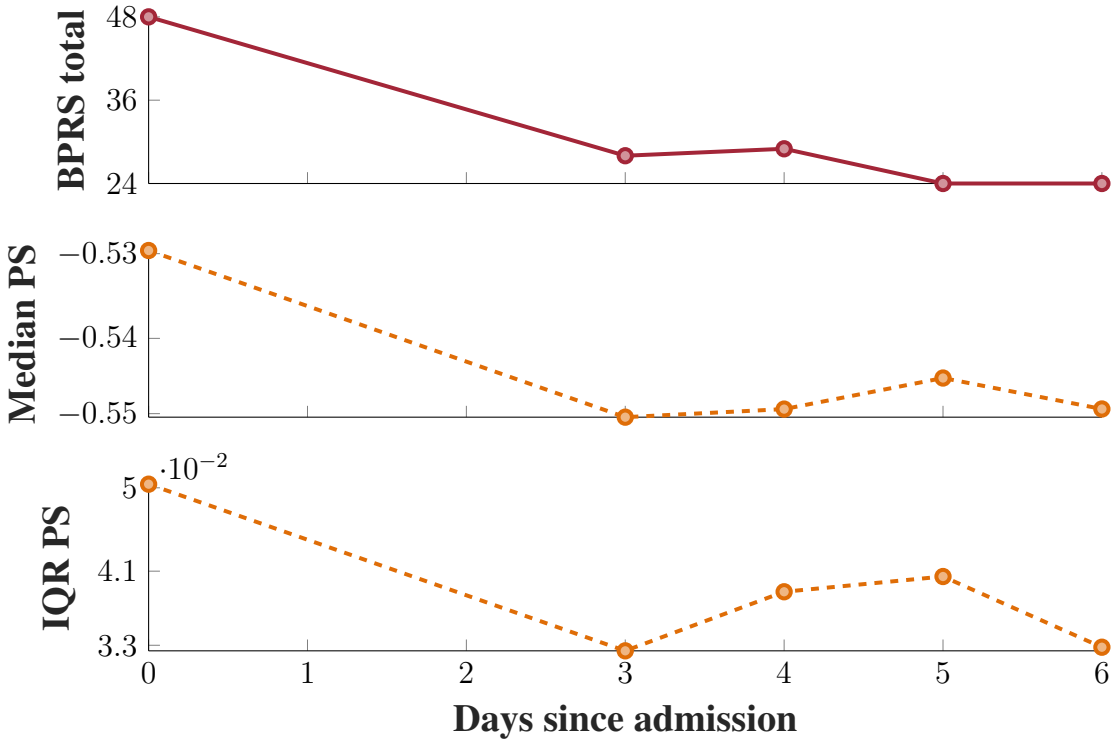


Figure 2.2: BPRS total (solid line) and acoustic descriptors (dashed lines) from admission to discharge for one patient.

2.3.2 Automatic Estimation of BPRS Items (Q2)

Table 2.2 summarizes the estimation results for all positive BPRS items as well as the BPRS total score. Except for *hallucinations* and *unusual thoughts*, we can estimate the BPRS items well (high correlation and lower MAE than baseline). BPRS specifies to assess these two items based on only what individuals say, not how they speak. While we observe correlations with these two items, see Table 2.1, a linear SVR is, in our case, unable to estimate these two items well. All well-estimable BPRS items have some relation to acoustic descriptors. While *motor hyperactivity* might not appear at first glance to be related to acoustic descriptors, the BPRS manual specifically advises taking rapid speech into account.

The performance comparison between descriptors based on the two diarization systems is summarized in Table 2.3. Even though all models are trained on descriptors based on manual di-

Table 2.2: Estimated generalization error of BPRS items on 29 interviews. * and ** mark significantly smaller absolute errors ($p < 0.05$ and $p < 0.01$, Wilcoxon signed-rank test).

BPRS item	r_{our}	MAE_{our}	$\text{MAE}_{\text{naive}}$
Hallucinations	0.27	1.00^{*}	1.34
Unusual thoughts	0.00	1.12	1.01
Elevated mood	0.73	0.64^{*}	1.12
Grandiosity	0.81	0.36^{**}	0.67
Excitement	0.71	0.60	0.91
Motor hyperactivity	0.74	0.55^{**}	0.94
Total	0.57	9.73	12.47

arization, we see a very similar performance with descriptors derived from automatic diarization. This indicates that even though our automatic diarization is imperfect, we can robustly estimate our descriptors.

2.4 Conclusion

We discovered several acoustic descriptors related to symptoms of psychosis while studying our first research question (Q1). Among them are a less consistent voice quality (larger variation in tenseness), a larger variation in speech volume, less variation in the opening time of vocal folds, a larger vowel space, and a smaller variation in speech rate. We also observe a smaller variation in formants for the overall severity and positive symptoms, which has been found in prior work for negative symptoms [30]. For tenseness of voice, our observations are in line with studies of depression and PTSD [163, 164]: a more tense voice is associated with more severe symptoms. We see small differences between positive symptoms and the overall severity. Namely, no variation in speech volume, not as prominent tense voice, and not as prominent reduced variation

Table 2.3: Performance comparison between descriptors based on manual (m) and automatic (a) diarization of 23 interviews. * and ** mark significantly smaller absolute errors ($p < 0.05$ and $p < 0.01$, Wilcoxon signed-rank test).

BPRS item	r_{our}		MAE_{our}		$\text{MAE}_{\text{naive}}$
	m	a	m	a	
Hallucinations	0.23	0.29	1.15	1.84	1.44
Unusual thoughts	0.12	0.19	1.09	1.17	1.10
Elevated mood	0.78	0.68	0.62^{*}	0.80^{*}	1.21
Grandiosity	0.82	0.80	0.41^{**}	0.46^{**}	0.77
Excitement	0.76	0.72	0.53^{*}	0.69	0.94
Motor hyperact.	0.73	0.67	0.51^{**}	0.84	0.92
Total	0.47	0.49	11.02	10.35	12.56

in formants for positive symptoms. Compared to previous work on depression [164], which observed a smaller vowel space for more severe negative symptoms, we observe the opposite for the severity of positive symptoms. This emphasizes that positive and negative symptoms should be studied separately.

Based on these computational acoustic descriptors, we trained models to estimate BPRS (Q2). As we observed a strong relationship between the acoustic features and the BPRS items, it is very likely that the trained models also learned those relationships. Since we achieved good performance even with automatic diarization, we can estimate BPRS items without needing manual diarization.

After understanding general data patterns in this chapter, Chapter 3 focuses on how reliable the model output is to enable better risk management as the output might not always be correct. The current chapter is also extended in Chapter 4 to cover more complex patterns at the person-level.

Chapter 3

Reliability Transparency

After exploring population-level patterns that a model might learn in Chapter 2, this chapter focuses on how reliable the model output is for each observation (reliability transparency). Unless a model can perfectly solve a task, knowing to which degree we should rely on a model's output is important, especially in critical such as health care. The work in this chapter was published at the Automatic Face and Gesture Recognition conference [206][1].

Facial action unit (AU) intensity estimation is central to many critical technologies, including assistive technologies in health care, driver fitness evaluations in automotive applications, and screenings in hiring agencies. For many of these applications, transparency is also important. We need models that predict not only a primary output but also a secondary quantity describing the reliability of its prediction. This is important for regression tasks as typically only the primary output, a model's best guess, is provided without quantifying its reliability.

Reliability in machine learning models primarily originates from two sources: 1) aleatoric uncertainty, in which observations can be noisy, and 2) epistemic uncertainty, in which the model might not be well-estimated or might have an improper structure [84]. For real-world applications, we need a reliability measure that combines both types of uncertainty. We can operational-

[1]The published paper is titled "Simple and Effective Approaches for Uncertainty Prediction in Facial Action Unit Intensity Regression" and is available at `https://doi.org/10.1109/FG47880.2020.00045`.

ize this reliability for regression tasks as the absolute error between a model's prediction and the ground truth. While prediction reliability has been studied in different fields [81], reliability in facial AU intensity prediction has not been studied.

Most existing approaches for reliability prediction rely on only epistemic uncertainty. This chapter studies approaches that capture epistemic and aleatoric uncertainties by predicting the absolute error. We describe two such approaches and compare them to a variety of established approaches[2]. Both approaches have a secondary model that predicts the absolute error of the primary model. Our first approach assumes that reliability is a weighted combination of known absolute errors from similar reference observations. This assumption has been previously demonstrated to work well [15, 78] when using a k-nearest neighbor approach. Our second approach uses a multi-layer perceptron (MLP) to predict the reliability. Such an approach has proven to work well in the past where a single perceptron predicts the reliability [139, 140]. With these two approaches, we can capture the prediction reliability, whether it is caused by epistemic uncertainty, aleatoric uncertainty, or by a combination of both. Finally, we demonstrate that our estimated reliabilities enable smaller prediction intervals than those derived from our baseline approaches. The prediction intervals are derived from the estimated reliablities using the framework of inductive conformal prediction and contain the ground truth at a targeted rate [141].

3.1 Problem Statement

The goal is to have a primary model that estimates the facial action unit intensity $\hat{y}$ given an input feature vector $\mathbf{X}$ and a secondary model that estimates the primary model's reliability $\hat{\epsilon}$. The reliability estimate $\hat{\epsilon}$ should correlate with the absolute error $|y - \hat{y}|$. To provide a confident reliability estimate for a machine learning practitioner, it should be possible to use $\hat{\epsilon}$ to derive prediction intervals around $\hat{y}$ that contain the ground truth y at a targeted rate, for example, 95% of the times.

[2]The code is available at `https://github.com/twoertwein/UncertaintyRegression`

3.2 Simple and Effective Reliability Prediction

We propose two approaches to estimate the reliability of a primary's model prediction $\hat{y}$ by training a secondary model to predict $|y - \hat{y}|$. For better comparison, the primary model is the same between the two proposed approaches and, when possible, also the same as for the baseline approaches. The primary model is a Multi-Layer Perceptron (MLP) that predicts the facial action unit intensity $\text{MLP}(\mathbf{X}) = \hat{y}$. Our two proposed approaches use an embedding learned by this primary model, namely the representation of the second last layer concatenated with the model's output. We refer to this embedding as $\mathbf{e}$. We use this embedding $\mathbf{e}$ instead of the high-dimensional input features $\mathbf{X}$ to reduce the secondary model's complexity and because it led to better performance in preliminary experiments.

3.2.1 DWAR: Similarity-based Error Prediction

The intuition of our first approach is that similar observations have a similar reliability. Our first approach adopts the non-parametric deep weighted averaging classifier [21] for regression (DWAR) as our secondary model. The DWAR model consists of an MLP to transform $\mathbf{e}$ in a low-dimensional embedding ($\mathbf{h}$) in which it usees an RBF kernel to define the similarity w between a new observation and the reference data (training data) where reference observation i is represented by its embedding $\mathbf{h_i}$.

$$\mathbf{h} = \text{MLP}(\mathbf{e}) \tag{3.1}$$

$$w(\mathbf{h}, \mathbf{h_i}) = \exp\left(-\|\mathbf{h} - \mathbf{h_i}\|^2\right) \tag{3.2}$$

The predicted reliability $\hat{\epsilon}$ of a new observation is the similarity-weighted average over the reference absolute errors[3].

$$\hat{\epsilon} = \frac{\sum_{i=1}^{N} |y_i - \hat{y}_i| \; w(\mathbf{h}, \mathbf{h_i})}{\sum_{i=1}^{N} w(\mathbf{h}, \mathbf{h_i})} \tag{3.3}$$

[3]Using the validation data should result in less biased errors, but we use the validation set for the prediction interval evaluation, and therefore cannot use the validation set to estimate the reliability.

The parameters of the transformation MLP are learned by minimizing $||y - \hat{y}| - \hat{\epsilon}|$. During training, only the current batch is used as reference data, leading to a time complexity of $\mathcal{O}(n^2)$ per batch of size n. At test time, the entire training dataset (size N) is used, resulting in $\mathcal{O}(N)$ time complexity for a single prediction.

3.2.2 U-MLP: Direct Error Prediction

Our second approach predicts reliability using an MLP (Uncertainty MLP, abbreviated as U-MLP). Meaning U-MLP is defined as $\text{MLP}(\mathbf{e}) = \hat{\epsilon}$ and we minimize $||y - \hat{y}| - \hat{\epsilon}|$.

3.3 Baselines

We compare DWAR and U-MLP to epistemic and aleatoric baselines. As aleatoric uncertainty is difficult to measure without the influence of epistemic uncertainty, we name approaches that directly predict reliability "supervised approaches".

3.3.1 Epistemic Baselines

Ensemble: The variance of ensembles is an established approach to quantify the prediction reliability [68]. Ensembles often consist of k models of the same type trained on bootstrapped data. While this approach does not make assumptions about the type of error distribution and can be used with any model, it can be computationally expensive to train k models instead of one model. We use an ensemble consisting of $k = 10$ MLP models to represent this baseline.

Dropout: By using dropout at inference time in a neural network, Bayesian inference can be approximated without the high computational costs associated with training Bayesian models [55]. To estimate reliability, we keep dropout at the second-to-last layer of the primary model activated and consider the variance over 1,000 inference runs for each observation to approximate Bayesian inference [55] (requiring 999 additional matrix multiplications). Since no additional

computations are required during training, this approach is more practical than the ensemble approach.

Gaussian Process (GP): Gaussian processes estimate a best guess (the mean) and a reliability (variance). While Gaussian processes typically use an RBF kernel on the input features to define the similarity between observations, a GP can use a neural network to learn an embedding for its kernel. For example, an MLP can learn a low-dimensional embedding and then use this embedding as the input to the RBF kernel. The neural network can jointly be learned together with the GP by minimizing the GP's marginal likelihood. GPs typically require a time complexity of N^2, which is prohibitive for large facial action unit datasets of size N. Sparse Gaussian processes [180] overcome this issue by representing the N data points by a smaller set of M so-called inducting points. The time complexity during training of the sparse GP is $\mathcal{O}(NM^2)$, where M is the number of inducing points ($M = 2,000$ in all our experiments). We use an MLP to learn the embedding for the RBF kernel. All parameters of this model—MLP parameters, scale parameter of the RBF kernel, inducing points of the GP, and the GP's observation noise parameter, which is shared between all inducing points—are trained end-to-end, optimizing the sparse GP's marginal likelihood.

3.3.2 Supervised Baselines

Multi-task MLP: Training a multi-task MLP for two tasks, one for the AU intensity estimation ($\hat{y}$) and one for the estimated absolute error ($\hat{\epsilon}$), could improve prediction of the absolute error at the cost of worsened AU intensity prediction. We optimize the following combined loss for the two tasks.

$$(y - \hat{y})^2 + 0.5(|y - \hat{y}| - \hat{\epsilon})^2 \tag{3.4}$$

Attenuation: A loss-function-agnostic approach deriving reliability is attenuating the original loss function by allowing the model to estimate its prediction variance (σ^2) for a prediction

($\hat{y}$) [84].

$$\frac{\text{loss}(y, \hat{y})}{\sigma^2} + \log \sigma^2 \tag{3.5}$$

The smaller the reliability estimation of a model (larger σ^2), the less confidence it has in its prediction. We use $-\sigma^2$ as the estimate for $\hat{\epsilon}$.

3.4 Experimental Setup

3.4.1 Datasets

We focus on two facial action unit (AU) datasets, and for comparison, we run the same experiments on a (subset of) MNIST to evaluate whether the approaches simply exploit skewed labels[4].

BP4D+: This dataset [228] (version 0.2) consists of videos of 140 subjects that have been annotated for facial AU intensities during emotion-eliciting tasks (AU 6, 10, 12, 14, and 17). We use subject-stratified hold-out sets for training (containing 60% of subjects), validation (20% of subjects), and testing (20% of subjects). Stratification is used to ensure a similar average AU intensity for each set.

DISFA: This dataset contains AU-annotated videos of 27 subjects viewing an emotion-eliciting video [123] (AU 1, 2, 4, 5, 6, 9, 12, 15, 17, 20, 25, and 26). We use the same method as for BP4D+ to determine the dataset split. The input representation of each image for these two datasets is the same as for OpenFace 2.0's AU intensity estimation [7]: face-aligned HOG features and the rigid and non-rigid shape parameters.

MNIST: We use MNIST [102] to validate the reliability approaches on a different domain. To make MNIST similar to BP4D+ and DISFA, we reframe MNIST as a regression task instead of a classification task and only consider numbers between 0 and 5 (the same range as facial AU intensities). To study the impact of skewed datasets on reliability, we use the 0-5 MNIST and

[4]Facial AU datasets are known to be highly skewed towards little activation. A reliability estimation approach might simply learn to associate a high activation estimation with a low reliability estimation.

also generate a skewed 0-5 MNIST that reflects the same skew as AU 6 (cheek raiser) has on the BP4D+ dataset.

To evaluate whether the reliability estimation generalizes to different datasets, we train a model on BP4D+ and evaluate its performance on the DISFA test set (without any adaptation, denoted as BP4D+ $\rightarrow$ DISFA+).

3.4.2 Implementation Details

All MLP models have the following hyper-parameters validated using the concordance correlation coefficient (CCC) on the validation set: the number of layers, the learning rate of Adam (the learning rate optimizer we use), and the dropout rate. Early stopping is based on the CCC on the validation set for the last 50 epochs. The maximum number of epochs is 500. Aside from the GP, all models are optimized for the mean squared error.

DWAR, U-MLP, and Dropout share the same primary model.

3.4.3 Metrics

We report the performance of the AU intensity predictions using the subject-averaged intra-class coefficient (ICC). For the predicted reliability, we use the following two metrics.

Spearman Correlation coefficient (ρ): We use the Spearman correlation coefficient to measure the monotonic relationship between the estimated reliability and the absolute error. This metric has the advantage that it does not penalize approaches that do not predict the intensities well but can estimate their own error well.

Prediction Interval Width (|PI|): We construct prediction intervals using the normalized inductive conformal prediction [141], where the normalization coefficient ($\hat{\sigma}$) is the predicted reliability.

$$\alpha = \hat{\sigma}\text{Perc}_{95}\left(\left\{\frac{|y_i - \hat{y}_i|}{\hat{\sigma}_i}\Big| i \in \text{calibration set}\right\}\right) \tag{3.6}$$

$$P(y \in [\hat{y} - \alpha, \hat{y} + \alpha]) \geq 0.95\ , \tag{3.7}$$

	DISFA	BP4D+	BP4D+ $\rightarrow$ DISFA[1]	DISFA[1]
DWAR/U-MLP/Dropout	0.502	0.653	0.520	0.545
Ensemble	0.341	0.664	0.495	0.535
GP	0.460	0.662	0.491	0.467
Multi-Task	0.477	0.643	0.450	0.549
Attenuation	0.477	0.646	0.479	0.556

Table 3.1: ICC (higher is better) of the primary models averaged over all AUs. [1] averaged over the common AUs between DISFA and BP4D+ (AU 6, 12, and 17).

where Perc_{95} is the 95th-percentile. The constructed intervals have an asymptotic coverage rate (the probability of containing the true intensity) of 95%, assuming that the calibration set (validation set) and the test set are both independent and identically distributed. We report the median interval width as a measure of the efficiency of the prediction intervals [142]. Smaller intervals at the same coverage rate are potentially more useful. This metric is affected by the accuracy of the AU intensity prediction.

Theoretically, a higher correlation (ρ) should result in a smaller interval width and vice-versa when the same primary model is used. In practice, this is not always the case as many outliers, i.e., more than 5% in the reliability prediction can negatively impact the intervals.

We test for statistical differences between our two described approaches at the subject level and against all baseline approaches. These tests are conducted with subject-clustered percentile bootstrapping[5]. We do not conduct these tests for DISFA because we have only five subjects in the test set. We use bootstrapping without clustering for 0-5 (Skewed) MNIST.

[5]We calculate the metric of interest for each cluster (each subject), and then bootstrap the difference between the approaches (5000 re-samplings and a 95%-confidence interval).

	0-5 MNIST		0-5 Skewed MNIST		DISFA		BP4D+		BP4D+ → DISFA	
	$\rho \uparrow$	$\lvert\text{PI}\rvert \downarrow$	$\rho \uparrow$	$\lvert\text{PI}\rvert \downarrow$	$\rho \uparrow$	$\lvert\text{PI}\rvert \downarrow$	$\rho \uparrow$	$\lvert\text{PI}\rvert \downarrow$	$\rho \uparrow$	$\lvert\text{PI}\rvert \downarrow$
DWAR	0.923^{U}	0.085^{U}	0.900^{U}	0.001^{U}	0.646	0.303	0.681	1.819	0.754	1.486
U-MLP	0.932	0.042	0.970	0.001	0.835	0.502	0.690	1.211	0.754	0.902
Ensemble	0.710^{UD}	0.360^{UD}	0.891^{U}	0.001^{UD}	0.506	0.907	0.591	1.839	0.572	1.621
Dropout	0.537^{UD}	0.119^{UD}	0.887^{U}	0.001^{UD}	0.795	0.351	0.576	1.923	0.614	1.543
GP	0.023^{UD}	0.365^{UD}	0.766^{UD}	0.235^{UD}	0.369	0.739	0.213	2.132	0.365	1.580
Multi-Task	0.851^{UD}	0.121^{UD}	0.785^{UD}	0.303^{UD}	0.654	0.881	0.620	2.065	0.689	1.800
Attenuation	0.617^{UD}	0.301^{UD}	0.834^{UD}	0.411^{UD}	0.576	1.162	0.632	2.091	0.735	1.672

Table 3.2: Averaged reliability metrics over AUs. MNIST reliability metrics are not averaged. For MNIST, marked results indicate a significantly worse performance compared to U-MLP (U) / DWAR (D).

3.5 Results and Discussion

Our initial experiment evaluates whether estimating reliability degrades the performance of AU intensity estimation, which would influence the prediction interval width ($|\text{PI}|$). Table 3.1 shows that almost all models (with the exception of the ensemble) predict AUs with comparable performance. The main experimental results are shown in Table 3.2. Table 3.4 provides AU-specific results for BP4D+, including the statistical test outcomes, and Table 3.5 demonstrates that the constructed prediction intervals reach their targeted coverage rate.

Reliability under Skew: To study the effects of label skew, we report results on 0-5 MNIST and 0-5 Skewed MNIST in Table 3.2. Since reliability predictions are better on 0-5 Skewed MNIST for almost all approaches, these approaches at least partly exploit the label skew. The skew may also explain the different prediction interval widths between DISFA and BP4D+: DISFA is much more skewed and has much smaller prediction interval widths.

Cross-dataset evaluation: Testing the BP4D+ models on DISFA provides two particularly

	0-5 MNIST		0-5 Skewed MNIST		DISFA		BP4D+		BP4D+ $\rightarrow$ DISFA	
	$\rho \uparrow$	$\lvert\text{PI}\rvert \downarrow$	$\rho \uparrow$	$\lvert\text{PI}\rvert \downarrow$	$\rho \uparrow$	$\lvert\text{PI}\rvert \downarrow$	$\rho \uparrow$	$\lvert\text{PI}\rvert \downarrow$	$\rho \uparrow$	$\lvert\text{PI}\rvert \downarrow$
DWAR	0.923	0.085	0.900	0.001	0.646	0.303	0.681	1.819	0.754	1.486
U-MLP	0.932	0.042	0.970	0.001	0.835	0.502	0.690	1.211	0.754	0.902
Ensemble	0.710	0.360	0.891	0.001	0.506	0.907	0.591	1.839	0.572	1.621
Dropout [1]	0.537	0.119	0.887	0.001	0.795	0.351	0.576	1.923	0.614	1.543
GP-VFE [2]	0.023	0.365	0.766	0.235	0.369	0.739	0.213	2.132	0.365	1.580
Multi-Task	0.851	0.121	0.785	0.303	0.654	0.881	0.620	2.065	0.689	1.800
Attenuation [3]	0.617	0.301	0.834	0.411	0.576	1.162	0.632	2.091	0.735	1.672

Table 3.3: Averaged reliability metrics over AUs. MNIST reliability metrics are not averaged. For MNIST, marked results indicate a significantly worse performance compared to U-MLP (U) / DWAR (D).

interesting results. The first result is a high correlation between the absolute error and the estimated reliability for the BP4D+ $\rightarrow$ DISFA evaluation (shown in Table 3.2). The second result is that the coverage rates for the prediction intervals, as reported in Table 3.5, are closely centered around the targeted 95%, even though the BP4D+ validation set is used for the calibration set. This could indicate that the evaluated reliability approaches generalize to data from slightly different conditions.

Epistemic vs. Supervised Approaches: The best performing approaches for each dataset and metric are supervised approaches. We hypothesize that supervised approaches perform better because they can use both the aleatoric and epistemic uncertainty to estimate the prediction reliability. In contrast, epistemic approaches can only capture epistemic uncertainty.

DWAR: This non-parametric approach achieves high correlations for almost all evaluations but tends to perform less well for the prediction interval width: it performs significantly worse on it than other approaches for severely-skewed AUs, e.g., AU 14 and AU 17. Like the weighted av-

	AU 6		AU 10		AU 12		AU 14		AU 17	
	$\rho \uparrow$	$\lvert\text{PI}\rvert \downarrow$	$\rho \uparrow$	$\lvert\text{PI}\rvert \downarrow$	$\rho \uparrow$	$\lvert\text{PI}\rvert \downarrow$	$\rho \uparrow$	$\lvert\text{PI}\rvert \downarrow$	$\rho \uparrow$	$\lvert\text{PI}\rvert \downarrow$
DWAR	0.723	2.693^{U}	0.543	2.551^{U}	0.653	2.569^{U}	0.749^{U}	1.202	0.738^{U}	0.080^{U}
U-MLP	0.653^{D}	1.541	0.520	1.596	0.631^{D}	1.485	0.783	1.423	0.862	0.012
Ensemble	0.635^{D}	2.750^{U}	0.391^{D}	2.571^{U}	0.509^{UD}	2.607^{U}	0.761	1.189_{D}	0.657^{UD}	$0.077^{\text{U}}_{\text{D}}$
Dropout	0.507^{UD}	$2.505^{\text{U}}_{\text{D}}$	0.418^{UD}	3.356^{UD}	0.389^{UD}	2.617^{UD}	0.702^{UD}	$1.107^{\text{U}}_{\text{D}}$	0.862_{D}	$0.029^{\text{U}}_{\text{D}}$
GP	0.339^{UD}	2.524^{U}	-0.035^{UD}	3.130^{UD}	-0.132^{UD}	3.231^{UD}	0.297^{UD}	1.366^{U}	0.598^{UD}	0.407^{UD}
Multi-Task	0.557^{UD}	2.930^{U}	0.392^{UD}	2.888^{UD}	0.576^{D}	2.698^{UD}	0.799_{D}	1.285	$0.774^{\text{U}}_{\text{D}}$	0.523^{UD}
Attenuation	0.632^{D}	2.479^{U}	0.464^{D}	3.039^{UD}	0.602^{D}	2.701^{UD}	0.771	1.258	0.692^{UD}	0.976^{UD}

Table 3.4: Statistical tests on BP4D+. Results marked in superscript/subscript indicate a significantly worse/better performance compared to U-MLP (U) / DWAR (D).

erage, DWAR is confined to the previously-observed range of errors in the training set. This may artificially truncate its correlation and result in large prediction intervals. This approach seems to work very well, even across datasets. Compared to the U-MLP, it requires more computational efforts but also provides transparency. A user can inspect the nearest neighbors, which influence the prediction the most.

U-MLP: U-MLP works very well across all datasets and metrics and never performs significantly worse than any other approach (see Table 3.4). It produces remarkably efficient prediction intervals across all datasets, e.g., +/- 0.6 on average for AU intensities on BP4D+, whereas other approaches need around +/- 0.9. In a few cases, it is outperformed by DWAR and dropout but is otherwise always the best-performing approach across both families.

Epistemic Baselines: The MLP ensemble and dropout are the best-performing epistemic baselines. The sparse GP poorly estimates the variance of some AUs (and the 0-5 MNIST). We hypothesize that this occurs because the marginal likelihood of this specific sparse GP is known to have many local minima [11]. Despite this drawback, this specific sparse GP has been shown

	Mean Coverage Rate		
	DISFA	BP4D+	BP4D+ $\rightarrow$ DISFA
DWAR	0.953	0.934	0.964
U-MLP	0.961	0.935	0.955
Ensemble	0.962	0.938	0.937
Dropout	0.956	0.935	0.960
GP	0.944	0.941	0.953
Multi-Task	0.951	0.944	0.951
Attenuation	0.957	0.934	0.953

Table 3.5: Observed coverage rate (ratio of the true value being in the interval) for the prediction intervals averaged over AUs.

to estimate the variance better than other sparse GPs [11].

Unlike the MLP ensemble, dropout variance has no overhead during training and is still computationally feasible at test time: there are only $n - 1$ additional matrix multiplications for the last layer. It is also already used in many situations, making it a convenient approach that can be implemented easily without re-training or training an additional model to derive reliability.

Supervised Baselines: The motivation behind a multi-task MLP model and the loss attenuation was to attain good error estimation despite a decrease in AU intensity prediction performance. The results suggest that estimating the error separately (as done in DWAR and U-MLP) outperforms these two baselines. However, it is important to note that both baselines have less computational overhead during training and testing than the U-MLP, only requiring back-propagation for an additional variable and one additional dot product at test time.

3.6 Conclusion

We evaluated the performance of two supervised approaches to estimating reliability and compared them to several established approaches. Some of these approaches require the use of slightly different architectures (GP, multi-task MLP, and loss attenuation), some require secondary models (U-MLP, DWAR, and ensemble), and some generally do not require any changes for existing users (dropout). The results suggest that epistemic approaches achieve a worse performance than supervised approaches, perhaps because they do not capture aleatoric uncertainty. The best-performing and simplest approach is the prediction of absolute error with a secondary MLP model (U-MLP). However, a notable result is that dropout provides decent reliability estimation while requiring the fewest changes during training.

This chapter laid the foundation to enable users of AI applications to know when to rely on a model's output and to which degree. Chapter 4 on personalized data transparency extends the Chapter 2 on population-level data transparency by focusing on how similar people are and what differentiates them.

Chapter 4

Personalized Data Transparency

In this chapter, we extend the initial study of population-level patterns from Chapter 2 to models that separate population-level patterns and individual differences by having person-generic and person-specific parameters. This not only allows us to quantify the differences a model learns for each person but also allows us to test whether those differences relate to additional information about that person. For example, does an affective model learn different temporal transition patterns between affective states for people who experience symptoms of depression? The work in this chapter was published at the International Conference on Multimodal Interaction [210][1].

Personalized prediction is a machine learning approach that predicts a person's future observations based on their past labeled observations. This type of model is typically used for sequential tasks that would be difficult without knowledge of the person, such as predicting daily mood from only smartphone data or predicting affective state sequences where transitions between states might be influenced by depression [146, 151]. As illustrated in Figure 4.1, a personalized model benefits by combining two types of trends (a) person-generic trends shared across people, such as being happier on weekends, and (b) unique person-specific trends, such as stressful weekly meetings or weekly socializing with friends. Person-specific trends can be

[1]The published paper is titled "Neural Mixed Effects for Nonlinear Personalized Predictions" and is available at https://doi.org/10.1145/3577190.3614115.

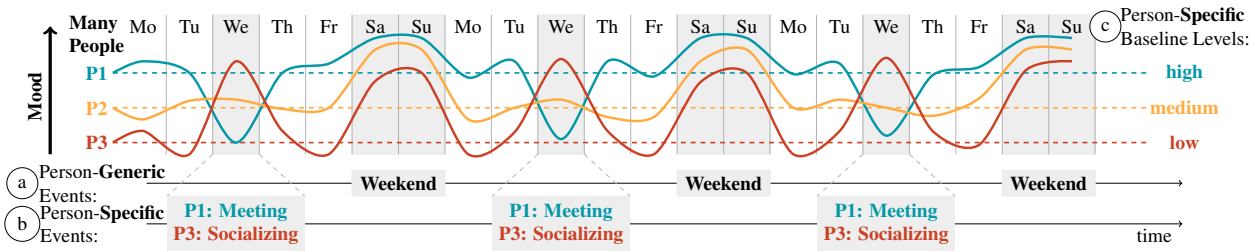
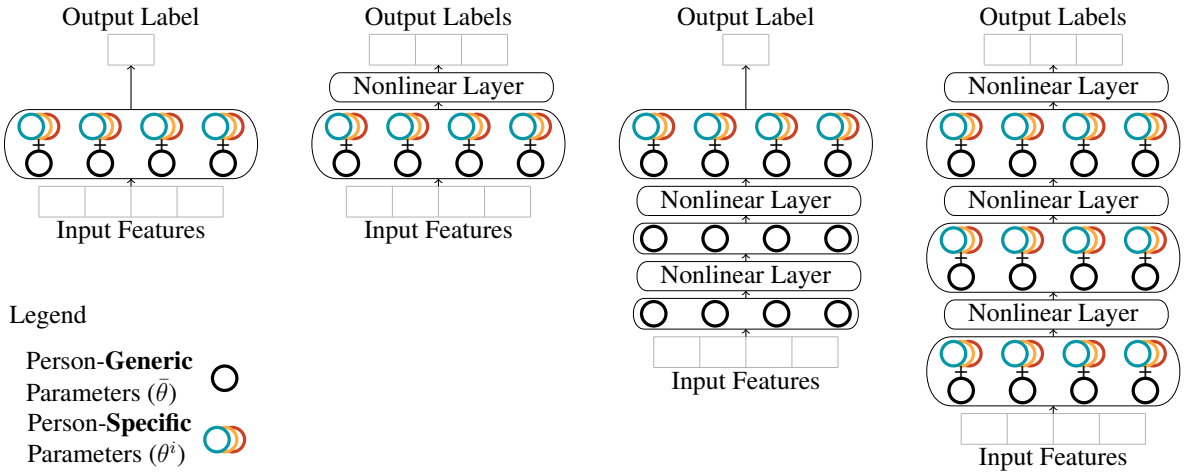


Figure 4.1: Illustration of why combining both person-generic and person-specific trends is important when learning personalized prediction models. The illustrated example is for daily mood prediction. (a) Most people are happier on weekends when they do not have to work. (b) Specific individuals, in our case P1 and P3, may have weekly events impacting their mood, e.g., socializing with friends can be positive, while a stressful meeting can be negative. (c) It is important to further know the baseline mood level of each person, as shown for P1, P2, and P3.



(a) Linear Mixed Effects (LME) [114]

(b) Nonlinear Mixed Effects (NLME) [29]

(c) Neural Networks with Linear Mixed Effects (NN-LME) [212]

(d) Neural Mixed Effects (NME), our approach

Figure 4.2: Visual comparison of our approach, Neural mixed Effects (NME), and previous approaches. NME enables person-specific parameters at any layer to represent nonlinear person-specific trends. Person-generic ($\bar{\theta}$) and person-specific (θ^i) parameters are combined by summing, i.e., $\bar{\theta} + \theta^i$.

challenging for machine learning models, even when trained on data from these people, as they might average out across people: as exemplified in Figure 4.1 when the more positive mood from a person's socializing coincides with the more negative mood of another person's stressful meeting.

Mixed effect models[2] are popular in statistics to study person-generic and person-specific trends by combining person-generic and person-specific parameters [99]. Linear mixed effect (LME) models have recently been gaining popularity in machine learning for personalizing models [85, 104, 105, 122, 132, 169, 170, 177, 183, 212]. Integrating LME with neural networks is currently limited to linear person-specific trends: person-specific parameters can only be in the last linear layer of a neural network as illustrated in Figure 4.2c. This rules out person-specific parameters in the remaining layers, i.e., nonlinear person-specific parameters. Separately from work with neural networks, nonlinear mixed effect approaches were proposed, but their optimization does not scale to large neural networks with many layers and parameters [29].

In this paper, we propose Neural Mixed Effect (NME) models to learn nonlinear person-specific parameters in a scalable manner. Our NME models combine the efficient optimization of neural networks with the person-specific parameters of nonlinear mixed effect models. NME learns nonlinear person-specific parameters by enabling them anywhere in a nonlinear neural network, as shown in Figure 4.2d. We demonstrate integrating our NME approach into two model architectures. We evaluate performance primarily on Multi-Layer Perceptrons (MLPs) for better comparison with previous MLP-LME work. To demonstrate NME for more complex models that yet have some interpretable parameters, we integrate NME with neural Conditional Random Fields (CRFs) to classify states in a temporal sequence [45]. CRFs explicitly model a sequence's temporal dynamics and allow us to interpret the person-specific temporal transitions between states.

We evaluate NME on six unimodal and multimodal datasets, including a smartphone dataset

[2]In statistics, the person-generic trends are often referred to as *fixed effects* and the person-specific trends as *random effects*. The name mixed effects comes from mixing both fixed and random effects.

	Linear Mixed Effects (LME)	Nonlinear Mixed Effects (NLME)	Neural Networks with Linear Mixed Effects (NN-LME)	Neural Mixed Effects (NME)
Nonlinear Model	✗	✓	✓*	✓
Dataset Scalability	✗	✓	✗	✓
Model Scalability	✓	✗	✓	✓

Table 4.1: Comparison of NME with previous approaches. LME models do not scale well with too many observations per person. The sampling-based optimization of NLME does not scale well with too many parameters. NN-LME has nonlinear person-generic parameters, but it re-use the optimization of LME, which (*) limits NN-LME to linear person-specific parameters and it does not scale as well for large datasets. Our proposed NME combines the efficient optimization of neural networks with the nonlinear persons-specific parameters of mixed effect models.

to predict daily mood and a mother-adolescent dataset to predict affective state sequences where half the mothers experience symptoms of depression. We analyze the interpretable person-specific transition parameters in the CRF and hypothesize that they differ between families where mothers experience symptoms of depression.

4.1 Technical and Related Background

Mixed effect models were proposed in statistics for data that is not independent and identically distributed, e.g., longitudinal data from multiple people [99]. In statistics, the goal of mixed effect models is often to study research questions about person-generic trends, referred to as fixed effects, and person-specific trends, referred to as random effects. Mixed effect models include a penalty term to regularize the person-specific parameters (denoted as $\boldsymbol{\theta}^i$) so that they learn only what the person-generic parameters (denoted as $\bar{\boldsymbol{\theta}}$) cannot learn. The technical challenge when optimizing mixed effect models is to separate fixed and random effects since they affect each other, e.g., a random bias term can affect the fixed slope of linear mixed effect models [171].

We briefly highlight the optimization of linear and nonlinear mixed effect models, review related work that explored combinations of neural networks and mixed effect models, and then contrast mixed models with multitask learning.

Linear Mixed Effects (LME): For an observation from the i-th person represented by a feature vector $\boldsymbol{X}$, a linear mixed effects model infers the prediction as $\hat{y} = (\bar{\boldsymbol{\theta}} + \boldsymbol{\theta}^i)^T \boldsymbol{X}$, see Figure 4.2a. For efficient optimization, it is often assumed that the random effects $\boldsymbol{\theta}^i$ follow a multivariate normal distribution with zero mean and covariance $\boldsymbol{\Sigma}$. A popular method to optimize LME models is an Expectation-Maximization (EM) algorithm that minimizes the mean squared error [114]. The challenging part of this EM algorithm is that a matrix needs to be inverted for each person i, where the matrix size is the number of observations for person i. This makes it challenging to optimize LME models when a person has many observations, i.e., LME models do not easily scale to large datasets.

Nonlinear Mixed Effects (NLME): Nonlinear mixed effect models are used to model nonlinear person-specific trends, for example, in pharmacometrics [137]. As shown in Figure 4.2b, random effects can be anywhere in a nonlinear model $\hat{y} = f(\boldsymbol{X}; \bar{\boldsymbol{\theta}} + \boldsymbol{\theta}^i)$ making their optimization more challenging. While multiple optimization approaches exist for nonlinear mixed effects [10, 29, 115, 149], most modern nonlinear mixed effect approaches find an approximate solution using random walk Metropolis sampling [29, 80]. One downside of this sampling approach is that it converges slowly for large models with many parameters [80]. One upside, compared to LME, is that this sampling approach scales well with many observations as it does not require matrix inversions that depend on the number of people or observations.

Neural Networks with Linear Mixed Effects (NN-LME): LME models have been combined with neural networks to improve performance for tasks involving longitudinal data from multiple people, such as for mood and mental health-related tasks [85, 122, 169, 170, 177, 183, 212]. All of these combinations follow the same mathematical formulation of $\hat{y} = (\bar{\boldsymbol{\theta}} + \boldsymbol{\theta}^i)^T f(\boldsymbol{X}; \boldsymbol{\theta}_{\text{neural}})$, see Figure 4.2c, where $\boldsymbol{\theta}_{\text{neural}}$ are the person-generic parameters of the neural

network. These combinations can be seen as simply placing an LME model on top of a neural network. Most NN-LME approaches use the same EM algorithm as LME models [114]. The only difference is that the neural network parameters $\boldsymbol{\theta}_{\text{neural}}$ become part of the fixed effects, meaning the neural network needs to be trained until convergence within every E-step, which can be slow for large neural networks. By re-using the same EM algorithm from LME models, its limitations apply: the random effects will minimize the mean squared error and NN-LME will not easily scale to large datasets. While two approaches extend beyond the means squared error by finding an approximate solution for binary classification[169, 170], their work does not generalize to multiclass classification.

Our proposed Neural Mixed Effects (NME) approach is a significant generalization of previous work by allowing person-specific parameters, i.e., random effects, anywhere in neural networks where even the last layer can be nonlinear. Our proposed NME model is also scalable to large datasets and large models by efficiently optimizing the NLME objective with stochastic gradient descent. We summarize this comparison in Figure 4.2 and Table 4.1

Multitask Models: Assuming not all model parameters have a person-specific component, mixed models are similar to multitask models where each task corresponds to a person [22, 174]. The two main differences are 1) mixed models have a person-generic ("shared") component even for parameters that have a person-specific component and 2) while multitask models can have an additional explicit regularization between the task-specific parameters [49, 179], mixed models do not require a hyper-parameter to determine the strength of this regularization as $\boldsymbol{\Sigma}$ is learned.

4.2 Problem Statement

Our main goal is personalized prediction: predicting a person's future observations by training on their past observations. The problem of personalized prediction using mixed effects can be formalized as follows. Given a training dataset with n people and n_i observations for the i-th person $\{(\boldsymbol{X}_j^i, y_j^i) \mid i \in [1, n],\ j \in [1, n_i]\}$ and a test dataset with unseen observations from the

same people, the goal is to learn a function $f(\boldsymbol{X}_j^i; \boldsymbol{\theta})$ predicting y_j^i where the parameters $\boldsymbol{\theta}$ are expressed as the sum of a person-generic $\bar{\boldsymbol{\theta}}$ and a person-specific component $\boldsymbol{\theta}^i$.

4.3 Neural Mixed Effect Models

Mixed effect models are gaining popularity in machine learning for personalized predictions as they combine person-generic and person-specific parameters. In this section, we present our generalization named Neural Mixed Effects (NME) model to better integrate mixed effect models in neural networks through a more scalable optimization and by allowing person-specific parameters anywhere. The advantage of our proposed NME approach is that it enables any neural network architecture to have person-specific parameters $\boldsymbol{\theta}^i$ as long as its original parameters (which we will refer to as person-generic parameters $\bar{\boldsymbol{\theta}}$) can be optimized with gradient descent. The only difference is that the person-specific components $\boldsymbol{\theta}^i$ also need to be stored and optimized. When making predictions for person i, the neural network parameters become the sum of these two components $\bar{\boldsymbol{\theta}} + \boldsymbol{\theta}^i$. Similar to multitask learning, not all parameters need a person-specific component. If parameters have no person-specific components, the parameters are equal to the person-generic components $\bar{\boldsymbol{\theta}}$.

We first focus on the optimization process in Subsection 4.3.1, then show that NME is a nonlinear mixed effects model in Subsection 4.3.2, and finally, we describe in Subsection 4.3.3 how to predict sequences using a neural Conditional Random Field (CRF) and how we combine it with NME.

4.3.1 Optimization

The goal is to learn person-specific parameters $\boldsymbol{\theta}^i$ representing person-specific trends, i.e., that cannot be learned by the person-generic parameters $\bar{\boldsymbol{\theta}}$. In addition to minimizing a downstream loss function l, mixed effect models separate person-generic and person-specific trends by regu-

larizing the person-specific parameters. This regularizing encourages the person-specific parameters $\boldsymbol{\theta}^i$ to only focus on what cannot be learned by the unregularized person-generic parameters $\bar{\boldsymbol{\theta}}$. Following previous NN-LME work, we regularized the person-specific parameters by assuming that they follow a multivariate normal distribution with zero mean and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{\dim(\boldsymbol{\theta}^i)\times\dim(\boldsymbol{\theta}^i)}$, where $\dim(\boldsymbol{\theta}^i)$ is the number of person-specific parameters. $\boldsymbol{\Sigma}$ is the same for all people. To make the regularization invariant to the scale of different downstream loss functions, mixed effect models have, next to $\boldsymbol{\Sigma}$, a second weighting factor σ^2 that represents the average downstream loss. The resulting loss function of NME is

$$\sum_{i=1}^{n}\left[\frac{1}{\sigma^2}\sum_{j=1}^{n_i} l(y_j^i, f(\boldsymbol{X}_j^i; \bar{\boldsymbol{\theta}} + \boldsymbol{\theta}^i))\right] + \boldsymbol{\theta}^{iT}\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}^i \ . \tag{4.1}$$

The left term of Equation 4.1 optimizes $\bar{\boldsymbol{\theta}} + \boldsymbol{\theta}^i$ for best downstream performance while the right term regularizes the person-specific parameters $\boldsymbol{\theta}^i$. As we have separate persons-specific parameters $\boldsymbol{\theta}^i$ for each person i but apply the same regularization, we are likely to learn larger person-specific parameters when a person has many observations: as the left term, the sum over the number of observations for a person is more likely to outweigh the regularization term on the right when a person has many observations. Intuitively, this improves performance the most when we have many observations for a person and helps prevent overfitting for a person with only a few observations.

Optimization of Equation 4.1 is performed with stochastic gradient descent in batches, where the regularization term on the right is scaled by how many observations a person has in the current batch B. The right part of Equation 4.1 becomes

$$\frac{\sum_{\boldsymbol{X}_j^k \in B} \mathbb{1}(k=i)}{n_i}\boldsymbol{\theta}^{iT}\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}^i \tag{4.2}$$

where the indicator function $\mathbb{1}(k=i)$ is 1 when the observation $\boldsymbol{X}_j^k$ is from the i-th person, i.e., $k=i$.

Table 4.2: Dataset characteristics. With the *calendar* modality we refer to metadata including the year and the weekday.

Dataset	Tasks	Group	#Groups	#Observations	Modalities
Imdb [203]	Movie rating (regression)	Genre	383	83 143	text
News [128]	Number of shares on Facebook (regression)	Outlet	598	60 080	calendar, text
Spotify [126]	Danceability rating (regression)	Genre	58	26 844	acoustic, calendar, text
IEMOCAP [18]	Arousal and valence ratings (regression)	Person	10	4784	acoustic, text, vision
MAPS [4]	Daily self-assessed mood ratings (regression)	Person	38	2122	calendar, GPS, text, typing
TPOT [208]	Four affective states (multiclass classification)	Person	195	15 228	acoustic, text, vision

After each epoch of minimizing Equation 4.1, we update σ^2 to the new average downstream loss l of the training set and $\mathbf{\Sigma}$ to the sample covariance matrix of the person-specific parameters $\boldsymbol{\theta}^i$.

Fortunately, it is common in mixed effect modeling to assume that the person-specific parameters are independent of each other [29, 202], which reduces $\mathbf{\Sigma}$ to an easy-to-invert diagonal matrix. This allows us to efficiently optimize Equation 4.1 even for large models with many person-specific parameters. NMEs with this assumption are as fast as multitask models when having the same person/task-specific parameters. As seen from Equation 4.1, the NME objective scales linearly with the number of people and their observations enabling NME to scale to even large datasets.

To summarize, 1) NME allows person-specific parameters anywhere in a neural network, 2) NME uses stochastic gradient descent to optimize even large models with many person-specific parameters efficiently, and 3) NME scales linearly with the dataset size.

4.3.2 NME as a Nonlinear Mixed Effects Model

NME learns a nonlinear mixed effects model because its optimization procedure follows that of the nonlinear mixed effects solver saemix [29]. saemix is designed to optimize nonlinear mixed

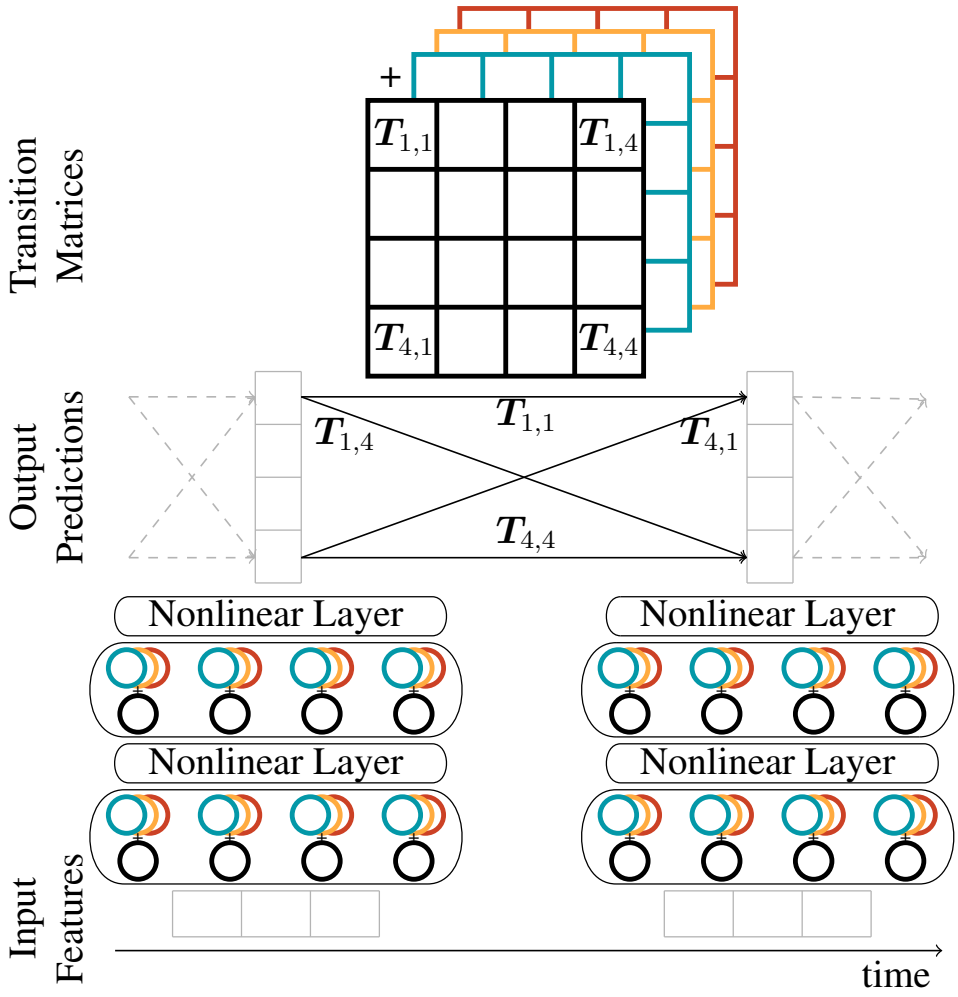


Figure 4.3: Illustration of the NME-CRF with person-specific parameters everywhere. An MLP predicts the initial output predictions which are refined by the CRF using the transition matrix $\boldsymbol{T}$.

effect models in statistics using random walk Metropolis sampling. However, sampling many parameters for neural networks is typically computationally challenging, converges slowly, and might lead to sub-optimal solutions [31, 80, 143]. NME replaces sampling with gradient descent to scale to large neural networks with many person-specific parameters.

saemix is an approximation EM algorithm [38], which means the expectation step (E-step) is not required to have converged before continuing with the maximization step (M-step). When assuming that the person-specific parameters $\boldsymbol{\theta}^i$ follow a multivariate normal distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}$, saemix incrementally minimizes Equation 4.1 during the E-step. During the M-step, saemix updates σ^2 and $\boldsymbol{\Sigma}$. Under general assumptions[3], saemix will converge to a mixed effects model. NME reduces Equation 4.1 during each epoch, corresponding to the E-step. Updating σ^2 and $\boldsymbol{\Sigma}$ between epochs corresponds to the M-steps. As NME follows the optimization procedure of saemix, NME will also converge to a nonlinear mixed effects model.

4.3.3 NME Conditional Random Fields

When predicting states that have a temporal order, such as the sequence of affective states on the mother-adolescent dataset, it can be beneficial to account for temporal dynamics, e.g., how likely it is to transition from one state to the next. Accounting for temporal dynamics may not only improve performance, but it may also be possible to interpret which transition the model infers as more or less likely. If we can further learn person-specific transitions, we can interpret whether they differ, for example, between families where mothers experience symptoms of depression.

Conditional Random Fields (CRFs) are graphical models that can learn state transitions in an interpretable manner [98]. When the transitions are assumed to be time-invariant, i.e., they are constant across time, we can represent all possible transitions from one to the next state through

[3]Assuming $l(y_j^i, f(\boldsymbol{X}_j^i; \bar{\boldsymbol{\theta}} + \boldsymbol{\theta}^i)$ are conditionally independent given the person i and follow a distribution in the exponential family.

one matrix $\boldsymbol{T} \in \mathbb{R}^{|\text{states}| \times |\text{states}|}$ where $|\text{states}|$ is the number of states. CRFs learn such a transition matrix $\boldsymbol{T}$. While CRFs have been combined with neural networks [45], they have not been explored with person-specific parameters, as done in the NME approach. With our NME-CRF, we can learn person-specific transition matrices $\boldsymbol{T} = \bar{\boldsymbol{T}} + \boldsymbol{T}^i$, which allows us to analyze them.

Besides a transition matrix $\boldsymbol{T}$, a CRF needs to know how likely each state is at time t, which we infer using an MLP. Figure 4.3 provides an illustration of NME-CRF. The CRF model can be optimized using gradient descent by minimizing the following loss function

$$-\frac{\exp\left(\sum_t^L f(\boldsymbol{X}_t^i; \bar{\boldsymbol{\theta}} + \boldsymbol{\theta}^i) + (\bar{\boldsymbol{T}} + \boldsymbol{T}^i)_{y_{t-1}, y_t}\right)}{Z([\boldsymbol{X}_1^i, \ldots, \boldsymbol{X}_L^i])} \tag{4.3}$$

where Z is a normalization function. We use the forward-backward algorithm to efficiently calculate Equation 4.3 [13]. To combine the CRF with NME, Equation 4.3 becomes the downstream loss l in Equation 4.1. At inference time, we use the viterbi algorithm to efficiently determine the most likely state sequence [13].

4.4 Experimental Setup

We evaluate our NME approach on six unimodal and multimodal datasets, including both regression and multiclass classification tasks. For better comparison with previous approaches, we primarily integrate NME with MLPs. The mother-adolescent dataset has temporal state sequences allowing us to evaluate the NME-CRF. We perform a more detailed analysis of the learned parameters of the NME-CRF since it learns interpretable state transitions.

4.4.1 Datasets

We conduct experiments on six datasets, summarized in Table 4.2.

Imdb [203], News [128], Spotify [126]: These are three public datasets used by previous NN-LME work [170]. We follow their experimental protocol and use the same features and

Table 4.3: Performance on six datasets with person-specific parameters in the last and all layers of the MLP. Best overall performance is underlined while best performance for the last/all layers is in bold. When a baseline is significantly worse than NME-MLP with person-specific parameters in the last or all layers, L or A are in superscript.

		Imdb	News	Spotify	IEMOCAP-A	IEMOCAP-V	MAPS	TPOT
		NRMSE ↓	NRMSE ↓	NRMSE ↓	CCC ↑	CCC ↑	Pearon's r ↑	Krippendorff α ↑
	Generic-MLP	0.927^{LA}	0.841^{LA}	0.711^{L}	0.510^{A}	0.518^{A}	0.119	0.355
Last	MLP-LME [212]	0.881^{L}	0.630	0.685	0.455^{L}	0.466^{L}	0.143	—
Last	Specific-MLP	0.891^{L}	0.646^{L}	0.794^{L}	0.431^{L}	0.354^{L}	0.074	0.347
Last	NME-MLP (ours)	**0.846**	**0.627**	**0.679**	**0.510**	**0.555**	**0.209**	**0.367**
All	Specific-MLP	0.886^{A}	0.654^{A}	0.770^{A}	0.452^{A}	0.443^{A}	0.124	0.288^{A}
All	NME-MLP (ours)	**0.856**	**0.629**	**0.690**	**0.558**	**0.559**	**0.138**	**0.367**

labels. Instead of people being the grouping variable on these datasets, we have genres on Imdb and Spotify and outlets on the News datasets as a grouping variable, i.e., we learn genre-specific and outlet-specific parameters. Following previous work, we report the root mean squared error (RMSE) for these three datasets. For easier comparison across the three datasets, we normalize the RMSE by the standard deviation of the ground truth labels on the test set (NRMSE).

IEMOCAP [18]: The IEMOCAP dataset [18] consists of dyadic interactions of five pairs of people, a total of ten people. Each pair is asked to improvise a set of emotionally charged interactions spontaneously. We separately predict arousal and valence ratings for each person on short utterances using features extracted by previous work [209], which includes statistics aggregated at the utterance-level of OpenFace 2.0 [7], openSMILE's eGeMaPs [53], and RoBERTa [117]. As is common for IEMOCAP, we use the concordance correlation coefficient (CCC) [101] as the evaluation metrics.

MAPS [4]: Mobile Assessment for the Prediction of Suicide (MAPS) is a longitudinal dataset of smartphone data of adolescents with daily mood self-assessments [4]. We predict the daily

mood self-assessments using their phone activity from the past 24h. Inspired by previous phone-based mood prediction work [3, 76, 107, 151], we extracted the following features: LIWC dimensions [147] and sentiment from Vader [74] of the typed text, the number of words, total time typing, the mean and variance of the typing speed, the weekday, the number of visited places based on GPS data as well as distance traveled and the average walking speed. The evaluation metric is Pearson's correlation coefficient r, which is well suited for evaluating how much of the mood variation we can predict.

TPOT [131]: The Transitions in Parenting of Teens (TPOT) dataset contains video recordings of dyadic interactions between mothers and their adolescents [131]. By design, mothers of half the dyads exhibit at least moderate depression symptoms at recruitment time and further had a treatment history for depression (referred to as the depressed group). The other half of mothers exhibits at most low symptoms, do not have a treatment history of depression, and had further no mental health treatment a month before recruitment (referred to as the non-depressed group). The interactions are typically 15 minutes long and focus on resolving areas of disagreement, such as participation in household chores. These interactions are annotated for each person for a sequence of four affective states (*other*, *aggressive*, *dysphoric*, and *positive*). These affective states are closely related to Living in Familial Environments codes [73, 167]. The affective state annotations are onset annotations, i.e., a state is annotated when enough evidence is available to determine the affective state and last until enough evidence is available for the next onset. This annotation approach means that two consecutive segments will not have the same label, e.g., *positive* will not follow *positive*. For additional details of TPOT, please see Chapter 5[4]. When using the NME-MLP, we predict these segments independently of each other. As the NME-CRF allows us to model temporal dynamics, we jointly predict each person's sequence of segments. In both cases, we use the same features from previous work [208], which are similar to the features on IEMOCAP but uses LIWC [147] instead of RoBERTa. Following previous work, we report

[4]When we use TPOT in Chapter 5 we use data from 268 people. As explained in the implementation details of this chapter, we removed people with fewer than ten annotations, resulting in a total of 195 people.

Krippendorff's α between the ground truth and the predicted labels.

4.4.2 NME Models and Baselines

Similar to previous work, we evaluate NME primarily in the context of MLPs (referred to as **NME-MLP**). Additionally, we evaluate NME using neural CRFs for the sequence prediction task on TPOT (referred to as **NME-CRF**). Since our NME approach allows person-specific parameters anywhere in the model, we explore three approaches: 1) having person-specific parameters in only the last layer (denoted as **last**), 2) for the CRF to additionally have person-specific parameters in its transition matrix $\boldsymbol{T}$ (denoted as **last+$\boldsymbol{T}$**), and 3) having them everywhere in the model (denoted as **all**). Figure 4.3 depicts the NME-CRF with person-specific parameters everywhere, including the transition matrix $\boldsymbol{T}$.

We compare NME-MLP and NME-CRF to three baselines.

Generic-MLP: Generic-MLP is either an MLP or a CRF (**Generic-CRF**) with only person-generic parameters, i.e., $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}$. Generic-MLP corresponds to a conventional MLP that is directly optimized with the downstream loss function l.

Specific-MLP: Specific-MLP is either an MLP or a CRF (**Specific-CRF**) with only person-specific parameters, i.e., $\boldsymbol{\theta} = \boldsymbol{\theta}^i$. The person-specific parameters are optimized with the downstream loss function l, i.e., they do not follow the NME approach. When evaluating person-specific parameters in only the last layer, we use person-generic parameters in all the previous layers of the MLP, i.e., $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}$ (the same as multitask learning with a task-specific last layer).

MLP-LME [212]: Almost all previous MLP-LME work [122, 177, 183, 212] is based on the same EM algorithm [114]. We implement MLP-LME as described in previous work [212], which makes MLP-LME a baseline for regression tasks with person-generic and person-specific parameters in the last layer, i.e., $\boldsymbol{\theta} = \bar{\boldsymbol{\theta}} + \boldsymbol{\theta}^i$. MLP-LME has so far not been extended to multiclass classification, so we cannot evaluate MLP-LME on TPOT.

4.4.3 Experimental Details

For all datasets we have a within-person split of 60% training, 20% validation, and 20% testing. For IEMOCAP, MAPS, and TPOT, the first 60% of the observations per person are used for training, the following observations for validation, and the last observations for testing. This is done to avoid temporally correlated observations that would invalidate the validation or test set.

All models are implemented in PyTorch [144] and optimized with Adam [88]. Their hyper-parameter are determined using a gridsearch which includes the learning rate, the number of layers in the MLP and their width, and L2 weight decay. Model validation is based on the validation set performance. All models are trained on consumer-level graphic cards, such as, the NVidia RTX 3080 Ti.

All input features are z-normalized on the training set. For regression tasks, the ground truth is also z-normalized based on the training set. The mean squared error is the loss function l for all regression tasks. For the MLP on TPOT, we minimize the cross entropy loss, while the forward-backward algorithm is used for the CRF on TPOT to minimize Equation 4.3. Features from different modalities are combined through early fusion.

When reporting performance metrics, we first calculate them within each person and then report the average. This allows us to focus on the within-person performance and avoids Simpson's paradox [171]. Significance tests are conducted with paired person-clustered bootstrapping [155] using $p = 0.05$ and 10,000 resamplings at the person-level[5]. To determine the performance metrics reliably, we need a large enough test set per person: we remove people from all experiments if we have less than ten observations from them.

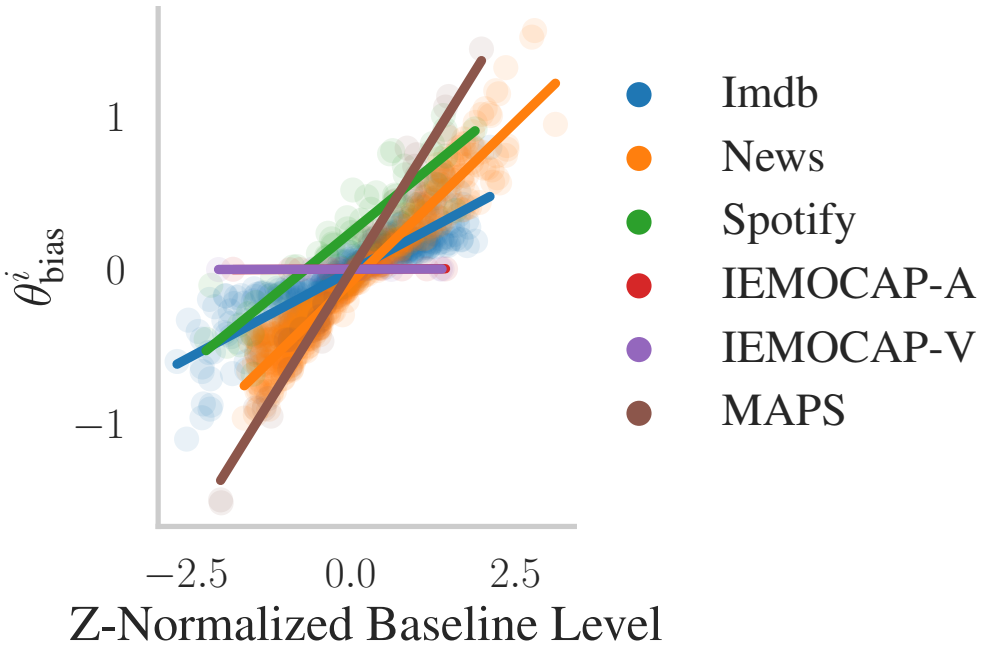


Figure 4.4: Correlation between the baseline level (ground truth on the training set) and the last bias term θ^i_{bias} of NME-MLP.

4.5 Results and Discussion

We first present the NME-MLP experiments across all six datasets and then focus on analyzing the NME-CRF multiclass classification experiments on the TPOT dataset.

4.5.1 NME-MLP Experiments

Last layer with person-specific parameters: We first evaluate NME-MLP with person-specific parameters in only the last layer for a direct comparison with MLP-LME [212]. NME-MLP performs numerically equal or better than all three baselines (Generic-MLP, Specific-MLP, and MLP-LME) on the six datasets, see the top half of Table 4.3. While Specific-MLP incurs a performance drop for the two smaller datasets, i.e., IEMOCAP and MAPS, NME-MLP maintains or improves performance indicating that it is important to have both person-generic and person-

[5]For each person, calculate the performance metric and take their difference between two models. Then bootstrap the differences by resampling 10,000 times with replacement to derive 95% confidence intervals using percentiles.

Table 4.4: Performance of the CRF on TPOT. Best overall performance is underlined while best performance for the last/all layers is in bold.

				Krippendorff $\alpha \uparrow$
			Generic-CRF	0.467
Last	+	$\boldsymbol{T}$	Specific-CRF	0.485
			NME-CRF (ours)	**<u>0.494</u>**
All			Specific-CRF	0.317^{A}
			NME-CRF (ours)	**0.470**

specific parameters. Unlike current MLP-LME implementations, NME-MLP can also be applied to multiclass classification on the TPOT dataset. NME-MLP again performs numerically better than its baselines. As indicated by the superscripts in Table 4.3, NME performs in many cases statistically significantly better compared to its baselines.

All layers with person-specific parameters: As illustrated in Figure 4.2d, NME enables person-specific parameters anywhere in a neural network. The bottom half of Table 4.3 summarizes the performance with person-specific parameters everywhere. NME-MLP numerically outperforms Specific-MLP and Generic-MLP. Having person-specific parameters everywhere also leads to the best performance across all IEMOCAP experiments suggesting that people in IEMOCAP may have nonlinear person-specific trends.

Interpretation of baseline levels: NME-MLPs for regression infer their prediction as $\hat{y} = (\bar{\boldsymbol{\theta}} + \boldsymbol{\theta}^i)^T \boldsymbol{Z}_j^i + \bar{\theta}_{\text{bias}} + \theta_{\text{bias}}^i$ where $\boldsymbol{Z}_j^i$ is the representation learned by previous layers. It is possible that $\bar{\theta}_{\text{bias}} + \theta_{\text{bias}}^i$ will correspond to a person's baseline level on the training set. As can be observed in Figure 4.4, θ_{bias}^i is highly correlated with the baseline level on all datasets, including IEMOCAP ($r = 0.669$ for arousal and $r = 0.543$ for valence). A potential explanation for why the magnitude of θ_{bias}^i is very small on IEMOCAP could be that the improvised dyads might be

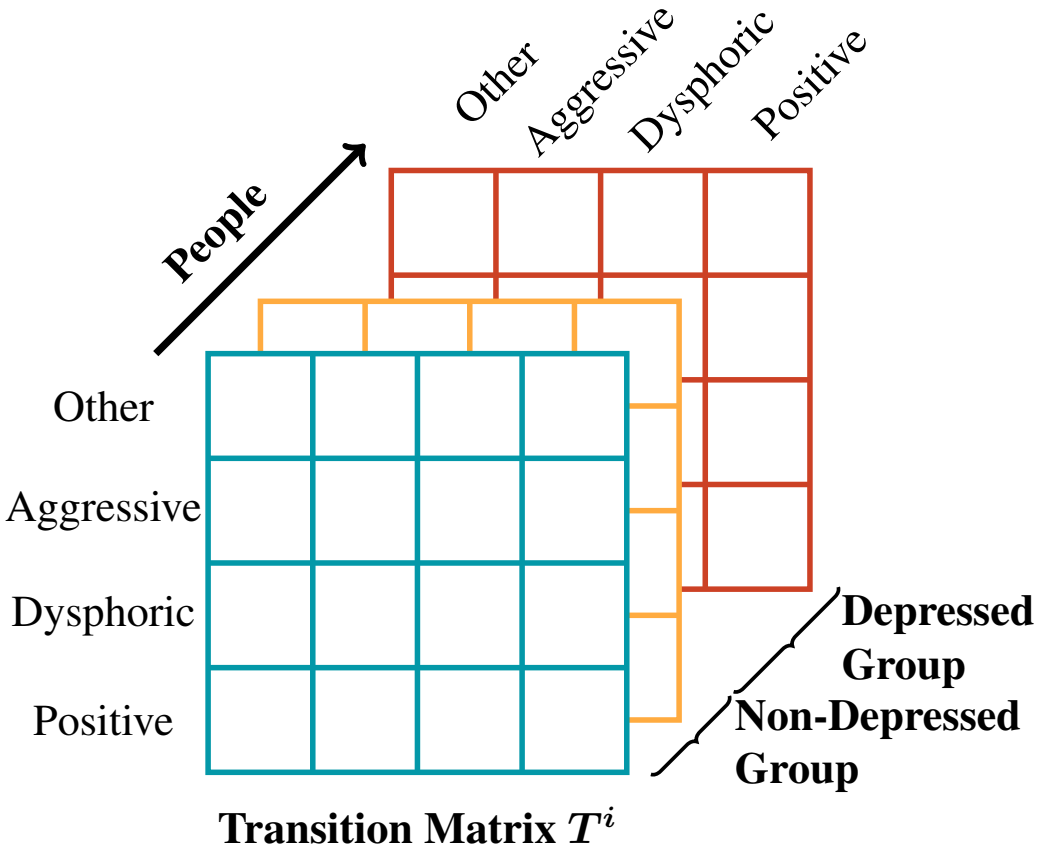


Figure 4.5: Visualization of the person-specific transition matrices. Half of the matrices belong to families where the mother is in the depressed group.

easier to predict, making it unnecessary for the model to encode the baseline levels.

4.5.2 NME-CRF Experiments

NME-CRF improves performance: We study the temporal structure of affective states on TPOT with the NME-CRF. While previous MLP-LME [212] work does not generalize to temporal structures, such as modeled by a CRF, our NME easily extends CRFs. Table 4.4 shows that NME-CRF numerically improves over its baselines, demonstrating that even more complex models benefit from having person-specific parameters and that the transition patterns on TPOT depend on the person.

Interpretation of temporal transitions: The NME-CRF model allows analyzing the learned

person-specific transition parameters. We focus on whether they differ between families (both adolescents and mothers) in the depressed and non-depressed group. We focus on this balanced group for two reasons 1) transition patterns have previously been linked to depression [167], and 2) already the ground truth base rate of the four affective states is different between them as indicated by the Chi-squared test $\chi^2(3, 8946) = 61.0, p < 0.001$. As visualized in Figure 4.5, we group the person-specific transition matrices and then compare their differences. The multivariate Hilbert-Schmidt Independence Criterion (HSIC) [148][6] indicates that the two groups have significantly different transition matrices HSIC $= 0.71, p = 0.006$.

The 95% confidence intervals of the differences in the transition probabilities between families in the depressed and non-depressed group shown in Table 4.5 indicate six significant differences between them. While families in the non-depressed group are more likely to transition from *positive* to the majority class *other*, families in the depressed group are more likely to transition to *aggressive* and *dysphoric*. Similar trends are observed for transitions from *other*: families in the non-depressed group are more likely to transition to *positive* while families in the depressed group are more likely to transition into *aggressive*. These observations seem plausible as more aggressive and less positive behaviors have been associated with depression [91, 166, 167]. As illustrated with the above analyses, it is possible to interpret the learned person-specific parameters learned by NME.

Regularization term needed for many person-specific parameters and small datasets: To test in which situations the regularization term of NME, i.e., the right part of Equation 4.1, is needed for good performance, we train an unregularized NME (uNME) that does not have the regularization term. We evaluate (u)NME with 1) person-specific parameters in different model parts of the CRF, and 2) with less and less training data per person. Figure 4.6 indicates that the regularization term is needed for many person-specific parameters and on smaller datasets. Even with little data, NME-CRF always performs better than the Generic-CRF despite having more

[6]We use the implementation from the R package dHSIC.

Table 4.5: 95% confidence intervals of the learned transition probability differences between families in the depressed and non-depressed group. Positive values indicate a higher transition probability for families in the depressed group. Intervals in bold are significantly different.

Model-implied Transitions		Into Other	Into Aggressive	Into Dysphoric	Into Positive
From	Other	[0.0, 1.8]	[**0.7** , **4.9**]	[−2.0, 3.2]	[**-7.4** , **−1.3**]
	Aggressive	[−1.2, 2.8]	[−1.7, 0.2]	[−0.4, 3.4]	[**-2.1** , **−0.4**]
	Dysphoric	[−5.5, 1.1]	[−0.1, 4.4]	[−0.9, 0.5]	[−1.4, 2.0]
	Positive	[**-8.3** , **−1.6**]	[**0.3** , **2.2**]	[**0.1** , **5.5**]	[0.0, 1.8]

parameters. As described in Subsection 4.3.1, mixed effect models tend to learn smaller person-specific parameters for a person with little data which helps avoid overfitting. In the extreme case of having very little data per person, the NME-CRF should converge to the Generic-CRF as the person-specific parameters will barely be used [150]. This trend can be observed in Figure 4.6 as the performance gap between NME-CRF and Generic-CRF narrows with fewer observations per person.

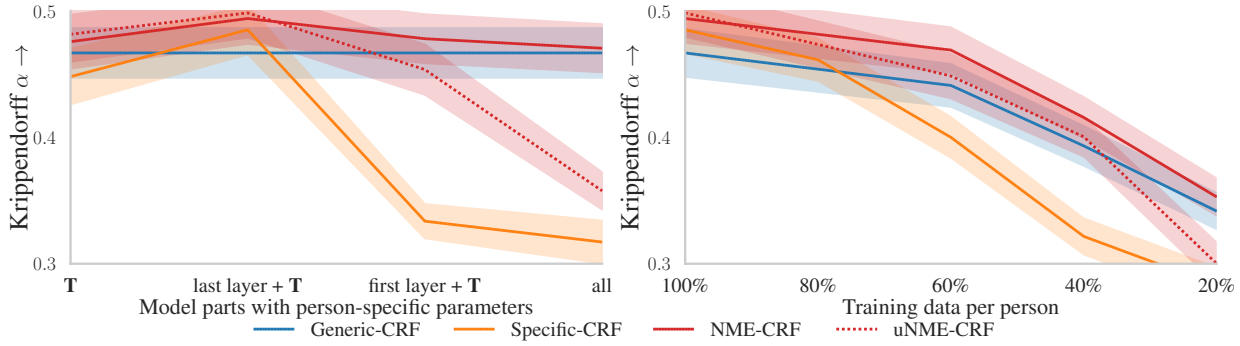


Figure 4.6: Performance on TPOT: (left) with person-specific parameters in different model parts and (right) when trained on smaller subset of data per person.

4.6 Conclusion

We demonstrated that personalized models benefit by combining two types of trends: (a) person-generic trends shared across people and (b) unique person-specific trends. Linear mixed effect models are gaining popularity in machine learning for personalization as they combine these two trends. We proposed Neural Mixed Effect (NME) models to generalize previous work integrating linear mixed effect models in neural networks. NME allows person-specific parameters anywhere in a neural network to learn nonlinear person-specific trends. NME's optimization is further scalable to large datasets and large neural networks. NME achieved this by combining the efficient neural network optimization with the person-specific parameters of nonlinear mixed effect models. We evaluated NME on six unimodal and multimodal datasets covering regression and classification tasks and observed numerical improvements on all six datasets. Further, we showed that NME can be combined with neural conditional random fields to learn interpretable person-specific temporal transitions. Finally, we demonstrated that person-specific parameters can be interpreted, for example, we observed that the person-specific transition matrices of the NME-CRF are different for families in the depressed group.

Chapter 5

Modality Importance Transparency

This is the first chapter of the multimodal transparency thrust, concentrating on multimodal model mechanics. This chapter quantifies how important modalities are for a model but also guides the model's modality importance by how informative modalities are for humans. Similar to the previous chapters, we explore transparency for healthcare-related affective states. In this chapter, we focus on four affective states that have shown a relation to a future onset of depression. The work in this chapter was published at the International Conference on Multimodal Interaction [208][1].

Depression is a prevalent mood disorder affecting more than 264 million people [77]. Detecting depression early is crucial, as depression can affect the development of adolescents [20]. Therefore, we are interested in depression-related affective states during mother-adolescent interactions. To this end, we focus in this chapter on three affective states, i.e., positive, aggressive, and dysphoric, that have shown a relation to a future onset of depression [167] and study how these states are expressed through different modalities.

Combining information from multiple modalities to predict affective states is challenging and does not always improve the predictive performance of machine learning models [196].

[1]The published paper is titled "Human-Guided Modality Informativeness for Affective States" and is available at `https://doi.org/10.1145/3462244.3481004`.

However, humans express themselves through multiple modalities, making it essential to study how humans integrate information from multiple modalities when recognizing affective states. We are also interested in leveraging this knowledge to effectively combine information from multiple modalities in machine learning models. We focus our study on how humans use three important modalities for face-to-face conversations [125], i.e., vision, language, and acoustic modalities. While all three modalities may always be available, we hypothesize that a subset of modalities will be sufficient to predict expressions of affective states. In particular, we expect that these subsets are not the same for each instance of affect expression.

In this chapter, we study the hypothesis of using a subset of modalities to predict affective states from two angels: (1) a human study to understand better which modalities people are paying attention to when recognizing affective states; and (2) the impact of integrating these human ratings to guide machine learning models to attend to a subset of modalities. An interesting aspect of this chapter is that we are holistically studying the relation between modalities and affective states by showing annotators all available modalities simultaneously and asking them to judge the informativeness of each modality. For these judgments, we discretize modality informativeness in three levels: (a) sufficient, when a modality is, by itself, enough to recognize the expressed affective state; (b) relevant, when a modality includes useful information about the expressed affective state but is not sufficient to recognize the affective state; and (c) none, when the modality does not seem to be used to express the affective state. We study whether human annotators can reliably accomplish this task and analyze the distribution of these modality informativeness annotations. Finally, we explore the impact of integrating these annotations in predictive models. Our study and experiments are performed on a recent dataset of mother-adolescence interactions recorded in the context of studying affective states related to the onset of depression [131].

5.1 Related Work

We group the related work into four topics. First, we cover computational approaches for predicting multimodal affective states. Then, we focus on how multimodal machine learning models estimate modality informativeness. Third, we mention multimodal perception experiments highlighting that affective states are differently perceived across modalities. Finally, we highlight some unimodal attempts at integrating human guidance to improve predictions of machine learning models, such as using eye gaze to attend to salient words in NLP tasks.

Multimodal affective recognition: A modality-centric view of affective states is to divide them into how they are expressed, i.e., non-verbally and verbally. Non-verbal affective states include, for example, the basic six emotions, while verbal affective states include more language-driven aspects of affect such as sentiment, complaining or (dis-)agreeing. Challenges such as AVEC [157], ComParE [165], and FERA [190] have focused extensively on predicting non-verbal expressions of affect. Similarly, language-driven aspects of affective states have been predicted as part of sentiment analysis [173, 221] and to some degree as part of dialogue acts [26]. In this chapter, we focus on three multimodal affective states, i.e., they are expressed non-verbally and verbally, that have shown to be statistically related to a future onset of depression [167].

Modality informativeness: Many models that focus on multimodal fusion implicitly or explicitly estimate the informativeness of modalities [135, 186, 187]. Two motivations for modeling modality informativeness are often a better predictive performance [135, 186] and making the model more interpretable as the impact of each modality is estimated [187]. One way to model modality informativeness is to use modality attention with decision-fusion models [135]. As attention is not guaranteed to reflect how important a modality is [153], we guide the modality attention to be similar to human perceived modality informativeness and also evaluate how similar the predicted modality attention is to the perceived modality informativeness.

Modality perception: Multimodal perception studies have been conducted to rate how affective states are perceived in different modality combinations [19, 118, 127, 152]. For example,

some researchers studied whether emotions are perceived differently across modalities [19]. Focusing on individual modalities has the advantage that other modalities cannot hinder the perception of the current modality, but being exposed to only a subset of modalities, i.e., not having all the available information, can lead to different judgments about affective states, as demonstrated by these studies. To avoid this limitation and to focus instead on modality informativeness, we ask human annotators to judge modality informativeness while being exposed to all available modalities.

Human-guidance: Human attention, operationalized as eye gaze fixations, has helped in unimodal tasks to learn more robust attention mechanisms in NLP as a way to attend to words [8]. In computer vision, eye gaze information was also used to attend to salient objects [182, 211]. While eye gaze is an effective way to derive visual attention, it is not well-suited to infer the informativeness of other modalities, such as the acoustic modality. As an alternative, we ask human annotators to rate how informative each modality is.

5.2 Dataset

Our study takes advantage of the recent Transitions in Parenting of Teens (TPOT) [131] dataset, which consists of 134 audio- and video-recorded mother-adolescent interactions (a total of 268 participants). These natural interactions are 20 minutes long and focus on problem-solving tasks. Conversations typically focus on discussing the amount of screen time, the participation in household chores, and the behavior toward other family members. All participating families are considered to have low social economic status in the US. The adolescents are between 11 and 14 years old, and half of the mothers have a history of an unipolar disorder. We are going to use this dataset again in Chapter 4, Chapter 6, and Chapter 7.

Each interaction is annotated for four multimodal affective states: positive, aggressive, dysphoric, and other (mostly neutral). These affective states are closely related to the Living in Familial Environments (LIFE) codes [73] and can directly be derived from them [167]. The

Krippendorff α of the annotated states is 0.66. The four affective states are expressed non-verbally and verbally. For example, being sad is coded as dysphoric, but self-focused complaints are also coded as dysphoric.

The affective state coding focuses on the onsets of events, i.e., when enough evidence is available to determine an affective state. We assume that an annotated state is valid until the next onset. Through preliminary machine learning experiments[2], we determined that the annotations are most likely delayed by one second. We, therefore, shift all annotations by one second. The dataset has a total of 4,117 positive segments, 1,683 aggressive segments, 5,313 dysphoric segments, and 6,221 other segments. The average segment duration is 6.1 seconds.

5.3 Human Judgment of Modality Informativeness

We are interested in how much information each modality contributes when recognizing affective states. Additionally, we want to explore whether interactions between modalities are crucial when predicting affective states or whether modalities can be used independently.

For our study, we recruited and trained two annotators from our local institution[3]. As the TPOT dataset contains sensitive data, all annotators were part of our IRB protocol. The annotation software ELAN [201] is used to display side-by-side videos of the mother and the adolescent. For each family member, we randomly select a balanced subset of twelve segments. We exclude segments of the "other" state from this annotation, as they are primarily characterized by neutral or no expressions. Each family member's video is randomly assigned to one of the two annotators. 10% of the videos (26 videos) are annotated by both annotators to calculate the inter-rater agreement (see section 5.3.1 for details about the Krippendorff α).

[2]By evaluating the performance when predicting the events shifted in 0.25-second increments.

[3]The two female annotators were already familiar with the annotation software. We followed the established approach for annotator training, where annotators are trained on a separate subset of the data (not used in our main study) until reaching a high enough agreement. In our case, we used the threshold of 0.7 Krippendorff α on the training subset.

Modality informativeness is the amount of information a modality contains to recognize an affective state. For each modality (vision, language, and acoustics), the annotators are asked "How much information does the modality contribute to the affective state?" and given the following response options: "sufficient information", "relevant information", "no information", and "not clear / I do not know". A modality is sufficient when the annotators can recognize the affective state using only this modality. In contrast, a modality is relevant if it is not sufficient to recognize the affective state by itself but provides information about the affective state. An example of relevant information is speaking loudly: it can signal a high arousal state, but typically we cannot differentiate between positive and aggressive states with just this cue. We should note that multiple modalities can be sufficient for the same segment. Furthermore, none of the modalities may be sufficient by itself, meaning that the interaction between modalities is crucial.

As a sanity check, we ask the annotators "Do you agree with the affective state?". This allows us to flag segments where the affective state might be too ambiguous. The annotators have the following response options: "agree", "somewhat agree (it could be interpreted as <affective state>)", "disagree", and again "not clear / I do not know". Our two annotators "agree" in 86% of the cases with the originally coded affective states. Our study excludes annotations where the annotators do not "agree" with the affective state.

Annotation Interface: Figure 5.1 shows a screenshot of the annotation interface. The side-by-side videos are located above the tiers (not shown in the screenshot). Knowing that another onset happens immediately before or after an onset that is going to be annotated was pointed out to be important by our two annotators during pilot studies. To contextualize the sampled onsets, nearby onsets are included in the ELAN files if they occur within five seconds.

5.3.1 Annotation Analysis

Annotator agreement: We report the agreement of our modality informativeness annotations using Krippendorff α: 0.50 for the visual modality, 0.66 for the language modality, and 0.65

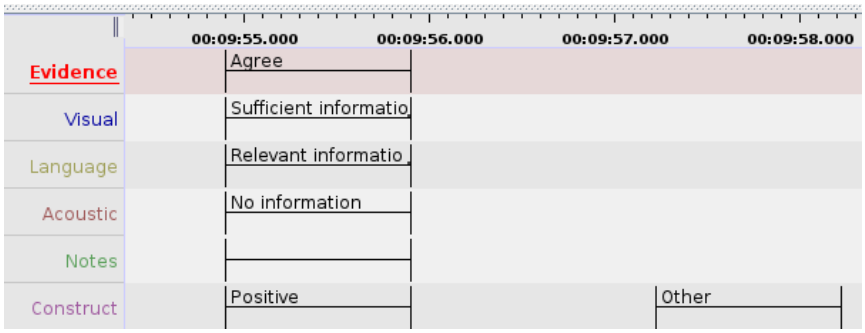


Figure 5.1: The annotation interface. On the left is an annotated onset and on the right is a nearby onset for context. The length of the segments has no meaning. The onset is the start of the segment.

for the acoustic modality. These Krippendorff α are computed using the ordinal weighting scheme [95] since our annotation label scheme is ordinal. If one or both annotators choose ”not clear / I do not know” for a modality, we treat the annotation as missing. Only 6% of the modality annotations are flagged as missing, leaving 2,724 segments for vision, 2,694 for language, and 2,703 for acoustics (15.6% of all TPOT segments). While we sampled the affective states in a balanced manner, not all videos have three aggressive segments, leading to an imbalance between the affective states. Out of the segments with at least one modality annotated (a rating different from ”not clear / I do not know”), 35.55% are positive, 24.02% are aggressive, and 40.38% are dysphoric.

Modality Informativeness: We analyze the informativeness of each modality. As seen in Table 5.1, the vision modality provides the most frequently sufficient information, followed by the language modality. Interestingly, the acoustic modality does not seem to provide as much information for this dataset. A potential explanation might be that it is cognitively difficult to focus on acoustic characteristics when listening to speech [111]. It is further surprising to observe that the annotators did not often choose ”relevant information”. This suggests that, in most cases, individual modalities could be sufficient to predict an affective state.

While we did not annotate which exact behaviors are causing relevant/sufficient information,

Table 5.1: Distribution of the modality informativeness.

Modality	Information		
	No	Relevant	Sufficient
Vision	16%	11%	67%
Language	49%	3%	41%
Acoustic	78%	3%	13%

Table 5.2: Common behaviors related to the three affective states as reported by the annotators.

State	Behaviors
Positive	head node, yes / agree statements, smile, eyebrows raised, laughter
Aggressive	head shake, no / disagreement statements, scowl / glare, eyebrows raised, sigh
Dysphoric	gaze aversion, head facing downwards / away from partner, self-touches (face and head), fiddling, shoulder shrugs, lip suck/bite, sigh

we asked our annotators for the most common behaviors for each of the three affective states and tabulated them in Table 5.2. Behaviors shared among affective states seem to be related to arousal (raised eyebrows) and valence (sigh). This is somewhat expected since both positive and aggressive states tend to be high arousal states, while aggressive and dysphoric states both tend to be low valence.

Informativeness and Missingness: The language and acoustic modalities are not always available since a person does not speak all the time. To validate if this greatly impacts informativeness annotations, we look at how often words are spoken during segments that are annotated as containing ”no information”. If words are spoken during an uninformative (”no information”)

Table 5.3: Percentage of available information for each affective state. 100% means all segments of the affective state.

Modality	Positive			Aggressive			Dysphoric		
	No	Rel	Suf	No	Rel	Suf	No	Rel	Suf
Vision	19%	6%	74%	24%	16%	55%	9%	12%	69%
Language	47%	2%	49%	25%	6%	63%	64%	3%	22%
Acoustic	70%	2%	26%	79%	9%	6%	83%	2%	4%

segment, we know that language and acoustics are available[4] and are not caused because speech is missing. During 51.15% of the uninformative language segments, words were spoken. Similarly for acoustics, 66.18% of the uninformative acoustic segments contain spoken words.

Modalities per affective state: Table 5.3 shows the distribution of informativeness for each affective state. Similar to Table 5.1, vision provides a lot of information across all affective states, but language provides more often information than vision for aggressive. In addition, language is more often informative for positive and aggressive than for dysphoric. A potential reason for this observation is that agreement and disagreement are coded as positive and aggressive, respectively. Another observation is that when the acoustic modality is informative, it tends to be informative for the positive state.

Cross-modal interactions: It is also interesting to study which modalities co-occur. Table 5.4 shows that more than half of the time, when language is informative, vision also provides information. When the acoustic modality is informative, it is often accompanied by visual information. While a single modality is frequently sufficient, affective states are often still expressed in multiple modalities. A predictive model could benefit from this extra information in terms of robustness by integrating uni-modal predictions dynamically based on a predicted modality

[4]This is a simplification for acoustics as people can also express themselves non-verbally, e.g., laughing, crying, and sighing.

Table 5.4: Co-occurrence of available information (relevant or sufficient). Co-occurrence probabilities are relative to how often the row modality is informative, e.g., in 38% of the cases when vision is informative, language is also informative.

Modality	Co-occurs with		
(base rate)	Vision	Language	Acoustic
Vision (83%)	100%	38%	18%
Language (48%)	67%	100%	9%
Acoustic (17%)	89%	26%	100%

informativeness.

5.4 Modality Attention

To guide how much attention a model pays to each modality, we explore two decision-fusion architectures that differ only in how modalities are aggregated. The first architecture averages unnormalized predictions (logits), while the second architecture averages normalized predictions (probabilities). While unnormalized logits contain more information than the normalized probabilities, the weighting of the unimodal predictions (attention) might be misleading as the unimodal models can learn to encode modality informativeness through the magnitude of their unnormalized logits instead of relying on the attention mechanism [153].

We use superscript in the following equations to denote a modality $m \in M$ with $M = \{v, l, a\}$. The prediction $\hat{y}_i$ of the unnormalized model for segment i is expressed as

$$p_i = \text{softmax}\left(\left[\sum_{m \in M} l^m_{i,\text{Pos}} a^m_i, \ldots, \sum_{m \in M} l^m_{i,\text{Oth}} a^m_i\right]\right) \tag{5.1}$$

$$\hat{y}_i = \text{arg max}_{s \in \{\text{Pos, Agg, Neg, Oth}\}} p_{i,s} \tag{5.2}$$

where $[\cdot]$ is the concatenation operator. The unnormalized logits $l^m_i \in \mathbb{R}^4$ for each modality m

are defined as

$$l_i^m = W^m f^m(X_i^m) + b^m \tag{5.3}$$

and the attention vector $a_i \in \mathbb{R}^{|M|}$ is

$$a_i = \text{softmax}\left(g([\, f^v(X_i^v), f^l(X_i^l), f^a(X_i^a)\,])\right) \ . \tag{5.4}$$

W is the projection matrix to the four affective states and b is the bias term. f and g are operationalized using Multi-Layer Perceptrons (MLP). This first model is part of the family of cooperative gating models [75] and is a special case of the multimodal gating unit [135] when used on the predicted output.

The second model averages normalized probabilities. The changes to the first model are

$$l_i^m = \text{softmax}(W^m f^m(X_i^m) + b^m) \tag{5.5}$$

$$p_i = \left[\sum_{m\in M} l_{i,\text{Pos}}^m a_i^m, \ldots, \sum_{m\in M} l_{i,\text{Oth}}^m a_i^m\right] \ . \tag{5.6}$$

5.4.1 Human-Guided Attention

Our goal is to study how models can be guided to prioritize modalities similarly to how humans judge the modality informativeness. Maximizing this similarity has the potential advantage of better interpretability and could also help the model during training to focus on the subset of informative modalities, as it might prevent the model from learning some spurious correlations.

We propose a new auxiliary loss to improve the similarity between model attention and human judgments. To formalize this loss, we define two matrices $A, H \in \mathbb{R}^{n\times|M|}$ where n is the number of segments. These matrices correspond to the predicted attentions (A) and the human informativeness judgments (H). Row i in these matrices corresponds to the importance of the three modalities for segment i. To define a similarity between the human judgments and the algorithmic attentions, we convert the ordinal human judgments to numeric values: no information (0.0), relevant information (0.5), and sufficient information (1.0).

$A_m, H_m \in \mathbb{R}^n$ are the columns of A and H, respectively. They correspond to attention values of modality m across all n segments. We minimize the following auxiliary loss

$$-\lambda \frac{1}{|M|} \sum_{m \in M} \text{pearson}\left(A_m, H_m\right) \ . \tag{5.7}$$

This loss maximizes the modality-averaged correlation between A and H. $\lambda \in \{0.1, 0.5, 1.0\}$ is a hyper-parameter to find a good scale for the auxiliary loss.

5.5 Experimental Methodology

To evaluate our human-guided prediction approach[5], i.e., the unnormalized model with the auxiliary loss from Equation 5.7 (referred to as guided), we define two baseline models: the normalized and unnormalized model, each without the auxiliary loss (referred to as normalized and unnormalized respectively).

We define an interaction-independent five-fold split for testing with a nested holdout split for validation (60% for training, 20% for validation, and 20% for testing). Reported metrics are averaged over the five test sets. The following hyper-parameters are validated for all models: learning rate for Adam [87], number of layers of the individual MLPs, strength of the L2-norm for the learnable parameters, and λ to balance the auxiliary loss. The primary loss function is the categorical cross-entropy, rectified linear units are used as non-linear activation functions, and early stopping is used. All parameters are jointly optimized. Features are z-normalized using the respective training sets and feature selection is performed with a linear support vector classifier [54] on the training sets. The best model is determined by the weighted accuracy averaged over the validation sets.

We report the affective state prediction performance using accuracy (Acc) and Krippendorff α[6]. Krippendorff α is chosen since we can easily compare the model's performance with the

[5]The code is available at `github.com/twoertwein/HumanGuidedAttention`.

[6]Krippendorff α is typically computed between the ratings of annotators. Here, we treat the model and the ground truth as two raters.

inter-rater agreement.

Further, we use two metrics to evaluate how interpretable the learned attention is. For each modality m, we report

$$\rho_m = \text{spearman}(A_m, H_m) \ . \tag{5.8}$$

Compared to section 5.4.1, we replace the differentiable Pearson's correlation with the non-differentiable Spearman's correlation since the human informativeness scale is ordinal. Additionally, the human informativeness and the predicted attention should, for each segment, have a similar/the same ordering. We report the segment-averaged Spearman's rank correlation coefficient

$$\bar{\rho}_{\text{local}} = \frac{1}{n} \sum_{i=1}^{n} spearman(A_i, H_i) \tag{5.9}$$

to evaluate whether segments have, on average, a similar attention ordering as the human informativeness. $A_i, H_i \in \mathbb{R}^3$ are the rows of A and H.

Significance tests are conducted with paired person-clustered bootstrapping [155] using $p = 0.05$ and 10,000 resamplings at the person-level.

5.5.1 Extracted Features

Vision: We use OpenFace [7] and AFAR [48] to extract facial action unit intensities and occurrences, head rotation, and eye gaze angles. When aggregating frame-level features to the labeled segments, we ignore features of video frames that were not correctly tracked according to OpenFace/AFAR. All features are aggregated to the labeled segments using the mean and standard deviation. Additionally, the maximum is used when aggregating facial action unit intensities. As an additional proxy for gaze aversion, we calculate the angular distance from looking straight into the camera [57] as the camera is located approximately on face-level behind the conversation partner. We combine the features from OpenFace and AFAR by concatenating their features, meaning we have action unit statistics from both OpenFace and AFAR.

Table 5.5: Performance on the entire test set and the gating metrics on the annotated test subset.

Model	α	Acc	$\bar{\rho}_{\text{local}}$	ρ_v	ρ_l	ρ_a
Chance	0.000	0.307				
Normalized	0.336	0.528	0.284	0.273	0.356	-0.148
Unnormalized	0.350	0.537	0.372	0.140	0.288	-0.090
Guided	**0.351**	**0.541**	**0.636**	**0.288**	**0.423**	**0.283**

Language: All interactions are manually transcribed. Words are automatically aligned to the audio using the Montreal Forced Aligner [124]. We use the dimensions from LIWC 2015 [178] to represent all words that occur during a labeled segment.

Acoustic: The audio files are first processed with StereoTool's declipper[7] in an attempt to recover clipped amplitudes caused by a too high microphone gain and then volume-normalized with FFmpeg according to the EBU R128 standard. Next to features from COVAREP [36], we extract the feature sets corresponding to the following openSMILE [50] configurations: eGeMAPS v01a [52], prosodyAcf (pitch and voicing probability), and vad_opensource [51] (speech activation detection). Most acoustic features are meaningful only while a person is speaking. When aggregating the audio features to the labeled segments, we consider audio features that happen only while speaking according to the aligned transcripts and when COVAREP/openSMILE detect speech. All low-level features are aggregated to the labeled segments using the mean and standard deviation. The high-level features from eGeMAPS are aggregated using only their mean.

Table 5.6: Performance on the annotated test subset. Oracle refers to using the annotated modality informativeness instead of the learned attention.

Model	Krippendorff α	Accuracy
Unnormalized	0.318	0.518
- Oracle	0.324	0.535
Guided	0.328	0.525
- Oracle	**0.379**	**0.561**

5.6 Results and Discussion

Human-guided attention: Results are summarized in Table 5.5. Our human-guided model shows small improvements over the baseline models, but most importantly, the learned attention weights are much closer to human judgment. The correlation between the attention and the human judgment significantly increased from $\bar{\rho}_{\text{local}} = 0.372$ to $\bar{\rho}_{\text{local}} = 0.636$, meaning that our guided model prioritizes modalities similar to how humans prioritize them. The modality-specific correlations ($\rho_v, \rho_l,$ and ρ_a) increased as well, making it easier to interpret the attention across segments.

Oracle experiment: Table 5.6 shows the hypothetical case when our guided model predicts perfectly the human informativeness. Its performance would significantly improve from $\alpha = 0.328$ to $\alpha = 0.379$. The other models do not improve significantly when using the human informativeness.

Attention per modality and affective state: Finally, Table 5.7 shows the averaged attention of our guided model for the three annotated affective states on the test sets. It is very intriguing to compare Table 5.7 and Table 5.3. This comparison shows similar trends between the human judgment and the model's attention: vision is essential and language is more important for pos-

[7] www.stereotool.com

Table 5.7: Average of the predicted attention for the three annotated affective states on the entire test sets.

Modality	Positive	Aggressive	Dysphoric
Vision	0.607	0.542	0.689
Language	0.329	0.422	0.276
Acoustic	0.064	0.036	0.035

itive and aggressive than for dysphoric. The only obvious difference is that the model amplifies the existing bias [229] of acoustics not being too predictive.

5.7 Conclusion

This chapter studied the hypothesis that a subset of modalities is sufficient to recognize affective states from two perspectives. First, we demonstrated that humans can reliably judge the informativeness of modalities and observe that, in most cases, a single modality is sufficient to recognize affective states while, at the same time, the affective states are still expressed through multiple modalities. Second, we proposed a human-guided auxiliary loss to improve the learned attention to be significantly more similar to human informativeness judgments while not degrading the predictive performance. Finally, the predictions can further be improved by directly using the human informativeness judgments during test time, demonstrating empirically that the human ratings are reliable. This paves the way for more intuitive and easier-to-interpret multimodal models.

Achieving a significant improvement when overwriting the learned attention with the human judgment indicates that our model can be corrected by a trained human, which makes our model more controllable and potentially also more acceptable by users [159]. This significant improvement also highlights the need for more research on how to learn better and more robust attention

mechanisms.

This chapter focused on the first of three multimodal phenomena we will investigate. The next multimodal phenomena are interactions between modalities which we discuss in Chapter 6. While this chapter processed modalities almost independently and combined them in a weighted sum, the next chapter processes modalities simultaneously to learn modality interactions that can only be present when modalities contextualize each other.

Chapter 6

Multimodal Interaction Transparency

After concentrating on the importance of modalities in Chapter 5, we now focus on the second multimodal phenomenon, namely multimodal interactions. Many of today's machine learning models learn multimodal interactions that empirically help their performance, but it is often unclear what interactions are learned and how they influence the model's output. This chapter tries to answer these questions. The work in this chapter was published at Findings of the Association for Computational Linguistics EMNLP [209][1].

Multimodal fusion integrates information from what we say, how we speak, and how we visually express ourselves. While multimodal models have led to performance improvements [186, 219, 222], they often have the downside of being difficult to interpret: it is unclear whether interactions between two modalities (bimodal) or three modalities (trimodal) are learned or whether these models focus on only one modality [204]. Quantifying multimodal interactions is an essential building block for future research: in model debugging as a step to better understand models and improve their performance [44] as well as in AI applications as a step to be more interpretable [61].

Seminal work [69] observed that many multimodal models function like the sum of uni-

[1]The published paper is titled "Beyond Additive Fusion: Learning Non-Additive Multimodal Interactions" and is available at `https://doi.org/10.18653/v1/2022.findings-emnlp.344`.

modal models, so-called additive models. In other words, these models might not be learning as many non-additive (bimodal and trimodal) interactions as expected. The non-additive interaction example in Figure 6.1 exemplifies how humans perceive the whole multimodal example as more than the sum of the two modalities. While the current approach of separating additive and non-additive interactions highlighted the problem of models primarily learning additive contributions, it did not provide solutions to learn non-additive interactions explicitly [69]. However, many multimodal language tasks require explicitly learning unimodal, bimodal, and trimodal interactions, such as, emotion recognition during spoken language [70].

In this paper, we introduce Multimodal Residual Optimization (MRO) to explicitly learn and decompose predictions into the sum of unimodal, bimodal, and trimodal interactions. Inspired by Occam's razor, to prefer simpler solutions, the main intuition of MRO is that (simpler) unimodal contributions should be learned before learning (more complex) bimodal and trimodal interactions. For example, the bimodal predictions should learn to correct the mistakes (residuals) of the unimodal predictions, thereby letting the bimodal predictions focus on the remaining bimodal

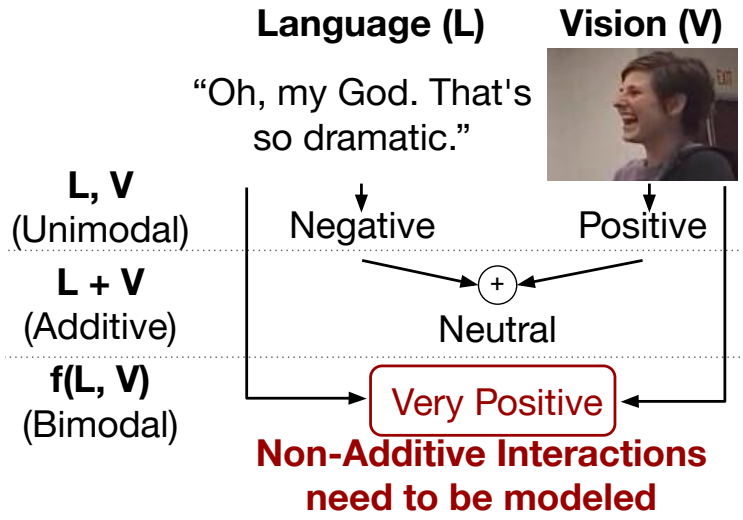


Figure 6.1: The joint assessment of language and vision (denoted as $f(L, V)$) is different from the sum of unimodal assessments (additive). This is an example for valence from the IEMOCAP dataset [18].

interactions. Similarly, trimodal predictions should learn what is not modeled by unimodal and bimodal predictions.

We evaluate MRO on six multimodal language datasets, including tasks for intent, sentiment, and emotion recognition. MRO aims to separate multimodal interactions (unimodal, bimodal, and trimodal) without degrading predictive performance. As part of evaluating MRO, we propose a new evaluation metric that extends prior work to three modalities [69]. We complement our empirical results with a human perceptions study to evaluate whether MRO learns non-additive interactions that align with human judgment.

6.1 Related Work

We review previous research on four aspects related to multimodal interactions: the prevalence of additive interactions, model-specific and model-agnostic quantification of modality interactions, and taxonomies of multimodal interactions.

Prevalence of Additive: Growing empirical evidence [69] and annotation studies [96, 152, 208] highlight that additive interactions are prevalent especially on datasets that are not carefully balanced, e.g., not having the same image contextualized with different captions [69]. An empirical approach highlights that multimodal models can be factorized into additive models without significant loss in performance [69], indicating that the examined models primarily relied on additive interactions. Similarly, multimodal perception studies indicate the importance of additive interactions: unimodal ratings of emotions are predictive of multimodal ratings [152]. Further, annotations of the semiotic mode, how the multimodal meaning emerges from individual modalities [9], of text-image pairs found that modalities provide mostly the same meaning [96]. Moreover, modality importance annotations for affective states found that a single modality often contains sufficient information to confirm an affective state [208]. While additive interactions are sufficient in many cases, non-additive interactions are still needed, especially when datasets contain the same unimodal representation in different multimodal contexts [69, 152].

Model-specific quantification: Models can indicate how much they rely on potentially non-additive interactions [187, 221]. Multimodal routing [187] was recently proposed to interpret the relative importance of multimodal interactions. It uses the routing-by-agreement algorithm [161] to focus more on modalities whose embedding is similar to other modalities' embeddings. The performance gains of the routing model hint at modalities containing partially redundant information [32] for emotion and sentiment prediction. While most model-specific approaches cannot rule out that a multimodal model potentially uses only one modality [204], MRO encourages that a bimodal model focuses on bimodal interactions.

Model-agnostic quantification: Multimodal interactions can be quantified after a model has been trained [69, 120, 188, 198]. EMAP [69] is based on the idea of factorizing any trained model into additive and non-additive interactions. Unfortunately, this marginalizing is very costly: with m modalities and a dataset of N samples, it requires N^m forward passes. Compared to EMAP, MRO learns a model that directly separates multimodal interactions.

Taxonomy of Multimodal Interactions: Many categorizations have been proposed to quantify the relationship between modalities [89, 198, 225]. A recent study [96] uses Koepfer's parallel, amplifying, and divergent. Parallel signals that only one modality is needed for prediction as they all provide the same meaning. Amplifying is sometimes also referred to as "additive" in a non-mathematical sense: modalities provide similar information but their combined meaning is either amplified or diminished. Finally, divergent indicates that modalities provide opposing information. Figure 6.1 is an example of opposing information.

6.2 Quantifying Multimodal Interactions

To learn a multimodal model that separates unimodal, bimodal, and trimodal interactions, we begin by defining how to quantify these three types of multimodal interactions. The work presented in this section is a generalization of prior work [69], which defined metrics to quantify multimodal interactions in the bimodal case.

Consider three modalities T (text), V (vision), and A (acoustic) with corresponding features x_T, x_V, x_A. A bimodal function f is *additive* when it can be factorized into the sum of two unimodal functions, $\forall x_T, x_V \ : \ f(x_T, x_V) \ = \ g(x_T) + h(x_V)$. Further, f contains unimodal contributions when parts of the prediction depend on only one modality: $\exists x_T \ : \mathbb{E}_v \, f(x_T, v) \neq 0$ [120]. This equation is illustrated for the language modality but has the same formulation for the vision modality. Prior work [69] proposed EMAP to quantify unimodal contributions (UC) in the context of two modalities. In this paper, we generalize UC to three modalities.

Claim 1. *A trimodal function f contains unimodal contributions when $UC(f, x_T, x_V, x_A) \neq 0$ with*

$$\begin{aligned} UC(f, x_T, x_V, x_A) = & \\ & \mathop{\mathbb{E}}_{v,a} f(x_T, v, a) + \mathop{\mathbb{E}}_{t,a} f(t, x_V, a) \\ & + \mathop{\mathbb{E}}_{t,v} f(t, v, x_A) - 2 \mathop{\mathbb{E}}_{t,v,a} f(t, v, a) \ . \end{aligned} \tag{6.1}$$

The idea of UC is to evaluate the model with all possible combinations of unimodal features (even feature combinations that are not in a dataset) so that the model cannot use non-additive interactions between modalities. Similarly, we can formulate a function BI to quantify bimodal interactions.

Claim 2. *A trimodal function f contains bimodal interactions (BI) when $BI(f, x_T, x_V, x_A) \neq 0$ with*

$$\begin{aligned} BI(f, x_T, x_V, x_A) = & \\ & \mathop{\mathbb{E}}_{t}[f(t, x_V, x_A) - UC(f, t, x_V, x_A)] \\ & + \mathop{\mathbb{E}}_{v}[f(x_T, v, x_A) - UC(f, x_T, v, x_A)] \\ & + \mathop{\mathbb{E}}_{a}[f(x_T, x_V, a) - UC(f, x_T, x_V, x_a)] \ . \end{aligned} \tag{6.2}$$

The remaining trimodal interactions (TI) are then simply what is not covered by the unimodal

contributions and bimodal interactions:

$$TI(f, x_T, x_V, x_A) = f(x_T, x_V, x_A) \\ - UC(f, x_T, x_V, x_A) - BI(f, x_T, x_V, x_A) \, . \tag{6.3}$$

When computational feasible[2], UC, BI and TI are valuable tools to evaluate whether a model contains unimodal, bimodal, and trimodal interactions.

6.3 Multimodal Residual Optimization

The main contribution of this paper is Multimodal Residual Optimization (MRO) which has the goal of learning and decomposing predictions into unimodal, bimodal and trimodal interactions to quantify them. Inspired by Occam's razor, the intuition of MRO is that (simpler) unimodal interactions should be prioritized before learning (more complex) bimodal and trimodal interactions. MRO has two components to separate modality interactions: an architecture and loss-function component.

6.3.1 MRO Architecture

Instead of using a single trimodal function to make a prediction $\hat{y} = f(x_T, x_V, x_A)$, the goal of MRO is to make predictions as $\hat{y} = UC(f, x_T, x_V, x_A) + BI(f, x_T, x_V, x_A) + TI(f, x_T, x_V, x_A)$ without having to compute UC, BI and TI. Therefore, MRO makes predictions $\hat{y}$ based on three components:

$$\hat{y} = \hat{y}_{\text{uni}} + \hat{y}_{\text{bi}} + \hat{y}_{\text{tri}} \tag{6.4}$$

[2] UC, BI, and TI can computationally be demanding given the expectation terms. While this is not as much of an issue when used as evaluation metrics, the computational cost prohibits us from using them as part of an iterative optimization process, e.g., in the loss function of neural networks.

where $\hat{y}_{\text{uni}}, \hat{y}_{\text{bi}}$ and $\hat{y}_{\text{tri}}$ model the unimodal, bimodal, and trimodal interactions respectively. It is important to note that $\hat{y}_{\text{bi}}$ and $\hat{y}_{\text{tri}}$ are intended to model only non-additive interactions, while $\hat{y}_{\text{uni}}$ is designed to model only additive interactions. $\hat{y}_{\text{uni}}$ is defined as

$$\hat{y}_{\text{uni}} = f_{\theta_T}(x_T) + f_{\theta_V}(x_V) + f_{\theta_A}(x_A) \tag{6.5}$$

where $f_{\theta_T}, f_{\theta_V}$ and f_{θ_A} are models, e.g., neural networks that use only one modality as an input. Each model has its own set of parameters (θ_T, θ_V, and θ_A). We parameterize the bimodal and trimodal models in a similar manner:

$$\begin{aligned}\hat{y}_{\text{bi}} =& f_{\theta_{TV}}(x_T, x_V) + f_{\theta_{TA}}(x_T, x_A) \\ &+ f_{\theta_{AV}}(x_A, x_V)\end{aligned} \tag{6.6}$$

$$\hat{y}_{\text{bi}} = f_{\theta_{TVA}}(x_T, x_V, x_A) \tag{6.7}$$

where $f_{\theta_{TV}}, f_{\theta_{TA}}$ and $f_{\theta_{AV}}$ are the bimodal models that take only two modalities as input, and $f_{\theta_{TVA}}$ takes all three modalities as input. The whole MRO model is parameterized with $\Theta = (\theta_T, \theta_V, \theta_A, \theta_{TV}, \theta_{TA}, \theta_{AV}, \theta_{TVA})$.

This architecture already enforces that $\hat{y}_{\text{uni}}$ can only contain unimodal contributions. While dedicating unimodal, bimodal, and trimodal models was explored in prior work [187, 217, 220], they did not explicitly encourage $\hat{y}_{\text{bi}}$ and $\hat{y}_{\text{tri}}$ not to contain unimodal contributions and similarly $\hat{y}_{\text{tri}}$ not to contain bimodal interactions. The MRO loss function described in the next section addresses this issue.

6.3.2 MRO Loss Function

We first explain MRO for two modalities (language and vision) before presenting the more general formulation for three and more modalities.

Bimodal case: To encourage $\hat{y}_{\text{bi}}$ to not contain unimodal contributions, MRO prioritizes $\hat{y}_{\text{uni}}$. MRO defines the loss function as

$$L(y, \hat{y}) = L(y, \hat{y}_{\text{uni}}) + L(y, sg\left(\hat{y}_{\text{uni}}\right) + \hat{y}_{\text{bi}}) \tag{6.8}$$

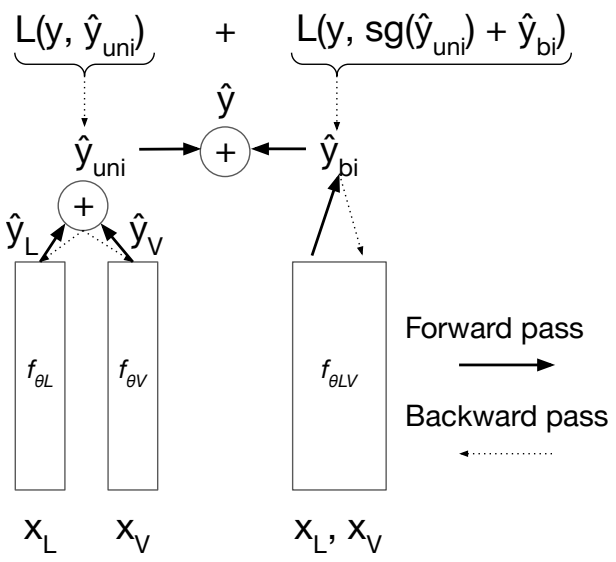


Figure 6.2: Overview of MRO: bimodal model learns what cannot be predicted by the unimodal contributions.

where sg refers to stop-gradient [154], which prevents back-propagation through sg's arguments. The first part of Equation 6.8 updates θ_T and θ_V to predict y using only unimodal contributions $\hat{y}_{\text{uni}} = f_{\theta_T}(x_T) + f_{\theta_V}(x_V)$. The second part of Equation 6.8 updates θ_{TV} so that $L(y, \hat{y}_{\text{uni}} + \hat{y}_{\text{bi}})$ is smaller; i.e., $\hat{y}_{\text{bi}}$ corrects mistakes that $\hat{y}_{\text{uni}}$ makes. We do not backpropagate again to θ_T and θ_V so that $\hat{y}_{\text{bi}}$ does not influence $\hat{y}_{\text{uni}}$; i.e., $\hat{y}_{\text{uni}}$ is optimized independently of $\hat{y}_{\text{bi}}$.

Figure 6.2 summarizes MRO in the bimodal case.

m-modal case: In the case of m modalities, we have m types of interactions: unimodal, bimodal, trimodal, ..., m-modal. Instead of separating just additive from all non-additive interactions, we want to separate these m types of interactions. MRO defines the loss function as

$$L(y, \hat{y}) = \sum_{i=1}^{m} L\left(y, sg\left(\sum_{j=1}^{i-1} \hat{y}_j\right) + \hat{y}_i\right) \tag{6.9}$$

where $\hat{y}_i$ refers to the i-modal predictions, i.e., $\hat{y}_1 = \hat{y}_{\text{uni}}, \hat{y}_2 = \hat{y}_{\text{bi}}, \hat{y}_3 = \hat{y}_{\text{tri}}$. For the trimodal case, $\hat{y}_{\text{uni}}$, $\hat{y}_{\text{bi}}$, and $\hat{y}_{\text{tri}}$ were defined in Subsection 6.3.1. When m is large than three, the models can be defined following the same approach. Similar to the bimodal case, $\hat{y}_{\text{bi}}$ is optimized independently

of $\hat{y}_{\text{tri}}$ as the gradient of $\hat{y}_{\text{bi}}$ is stopped by sg when optimizing $\hat{y}_{\text{tri}}$.

6.3.3 Sequential MRO

An alternative to MRO's approach of simultaneously optimizing all the main components ($\hat{y}_{\text{uni}}$, $\hat{y}_{\text{bi}}$, $\hat{y}_{\text{tri}}$), the sequential MRO (sMRO) proposes to optimize them sequentially.

First, sMRO optimizes the parameters of $\hat{y}_{\text{uni}}$ using the loss $L(y, \hat{y}_{\text{uni}})$ until convergence and then freezes its parameters θ_L, θ_V, and θ_A before optimizing $\hat{y}_{\text{bi}}$ and $\hat{y}_{\text{tri}}$. Next, sMRO optimizes the parameters of $\hat{y}_{\text{bi}}$ using the loss $L(y, \hat{y}_{\text{uni}} + \hat{y}_{\text{bi}})$ until convergence and then freeze the bimodal parameters θ_{LV}, θ_{LA} and θ_{VA}. The trimodal $\hat{y}_{\text{tri}}$ can then be optimized using the loss $L(y, \hat{y}_{\text{uni}} + \hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})$. For cases with more than three modalities, sMRO can optimize the parameters of $\hat{y}_m$ for $L(y, \sum_{i=1}^{m} \hat{y}_i)$ until convergence and then freeze the parameters of $\hat{y}_m$.

6.4 Experimental Methodology

We evaluate whether we can train a model that separates unimodal, bimodal, and trimodal interactions while not degrading predictive performance.

Datasets: We focus on five sentiment- and emotion-annotated datasets for which prior work used multimodal models, see Table 6.1. We also include the Instagram dataset [96] as it has modality interaction annotations (semiotic modes), which we can use to evaluate MRO.

We use the same features across all sentiment and emotion datasets: RoBERTa [117] as a representation of transcribed utterances; OpenFace 2.0 [7] to summarize face-related features, and openSMILE's eGeMAPS [52] to summarize acoustic features. For the Instagram dataset, we use the author-provided ResNet features [67] to summarize the image content and use RoBERTa to represent captions.

Evaluation: We want that the prediction components $\hat{y}_{\text{uni}}, \hat{y}_{\text{bi}}$ and $\hat{y}_{\text{tri}}$ correspond to $UC(\hat{y}), BI(\hat{y})$, and $TI(\hat{y})$ so that the prediction components represent only unimodal, only bimodal, and only

Original Paper	Tasks	Abbreviation	Samples	Modalities
[217]	Sentiment (regression)	MOSI	2.2k	3
[221]	Sentiment, Polarity, Happiness (regression)	MOSEI	22.9k	3
[18]	Arousal and Valence (regression)	IEMOCAP	4.8k	3
[189]	Arousal and Valence (regression)	SEWA	1.9k	3
[131]	Affect categories (4-way classification)	TPOT	17.3k	3
[96]	Intent of Instagram posts (7-way classification)	Instagram	1.3k	2

Table 6.1: Dataset overview.

trimodal interactions. To test this, we use $|UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})|$ to evaluate whether the bimodal and trimodal predictions contain unimodal contributions and $|BI(\hat{y}_{\text{tri}})|$ whether the trimodal prediction contains bimodal contributions. Given the MRO-architecture, $\hat{y}_{\text{uni}}$ cannot include bimodal and trimodal interactions and $\hat{y}_{\text{bi}}$ cannot include trimodal interactions. This means, if $|UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})| + |BI(\hat{y}_{\text{tri}})|$ is 0, the model perfectly separates unimodal, bimodal, and trimodal interactions, i.e., $\hat{y}_{\text{uni}} = UC(\hat{y}), \hat{y}_{\text{bi}} = BI(\hat{y})$, and $\hat{y}_{\text{tri}} = TI(\hat{y})$. We use 5-fold test setup for all datasets.

Models: We compare the MRO-architecture when optimized in different manners: with $L(y, \hat{y}_{\text{uni}} + \hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})$ (referred to as **Joint**), **sMRO**, and **MRO**. For performance comparison, we include the routing model [187] (referred to as **Routing**), a recently proposed model with the goal of modality interpretability. Lastly, we compare the performance against a single trimodal model $\hat{y} = f_{\theta_{TVA}}(x_T, x_V, x_A)$ (referred to as **Tri**) to evaluate whether the larger MRO-architecture has two many parameters for smaller datasets.

Implementation Details: The functions f of Equation 6.4 are instantiated as multi-layer

perceptrons. For each multimodal model, e.g., $f_{\theta_{TV}}$, we implement two popular types of fusion: early fusion (concatenating the modalities) and tensor fusion [219] (outer product between modalities after learning unimodal embeddings). The type of fusion is a hyper-parameter together with the number of layers, their width, learning rate, learning rate decay, L2 weight decay, dropout, and with/without prior feature selection. As a loss function, we use the mean absolute error for regression tasks and the cross-entropy loss for classification tasks.

6.5 Multimodal Perception Study

We conduct a multimodal perception study to evaluate whether MRO learns non-additive interactions, when humans also require non-additive interactions. We choose arousal and valence on the IEMOCAP dataset for this study as arousal and valence are two fundamental dimensions to describe emotional states [130].

Study Design: Crowd workers[3] are asked to rate arousal and valence of video segments when being exposed to only a subset of modalities. The four subsets are: 1) the transcript of what the person says (T); 2) the muted video (V); 3) the low-pass filtered audio (A), and 4) the transcripts, the video, and the original audio (TVA_O). IEMOCAP has ten speakers. We randomly select ten segments for each speaker, i.e. 100 segments.

Audio Processing: It is challenging to disentangle speech content and how we speak [12]. Similar to previous work, we low-pass filter the audio signal [215]. Instead of using 850 Hz as a cut-off [215], we use a lower cut-off frequency, as we could understand spoken words at 850 Hz. We choose 660 Hz[4] as it is the mean of the maximum pitch in an empirical study [106] and it also closely coincides with the maximum pitch of contralto singers (E_5 at 659.25 Hz). We

[3]We recruited 40 US-based crowed works from the platform prolific `https://www.prolific.co/` whose first language is English.

[4]We use ffmpeg for low-pass filtering with the following filter configuration: firequalizer=gain='if(lt(f,660), 0, -INF)':min_phase=1

	Arousal	Valence
Min. age	19	21
Mean age	36	37
Max. age	79	62
Female	20	19
Male	20	21

Table 6.2: Basic demographic information about the annotators.

choose this pitch-focused definition as we believe that prosodic information will predict arousal and valence.

Avoiding learning effects: Raters might be able to infer the missing multi-modal context after having rated some of the unimodal subsets for a specific segment. We therefore use two mechanisms to address learning effects across the modalities. First, each of the raters annotates only 20 randomly selected segments for each modality subset (we have eight raters per segment and modality subset). Second, we structurally randomize the order of the modality subsets by first presenting all unimodal subsets in a random order and in the end the trimodal segments.

Ratings and reliability: Following the annotation setup from IEMOCAP, we use the ordinal arousal and valence manikins scale consisting of five levels [16] to rate the two emotional dimensions. The effective reliability [158] over k raters as measured by the Intra-class Correlation Coefficient ICC(2, k-1) is excellent (above 0.9) [94] for all modality subsets. Further, our new trimodal ratings (TVA_O) correlate highly with the existing annotations on IEMOCAP $r(98) = 0.88, p < 0.001$ for arousal and $r(98) = 0.92, p < 0.001$ for valence, indicating that we can use our new annotations to inspect models trained on the original annotations.

Evaluation: To evaluate when humans require non-additive interactions, we train a linear regression model (an additive model) that predicts TVA_O given T, V, and A. We refer to this

model as $\hat{y}_{\text{uni}}^{\text{human}}$. The model fit of $\hat{y}_{\text{uni}}^{\text{human}}$ shows how important the missing non-additive interactions are [152]. Further, the absolute error $|\text{TVA}_O - \hat{y}_{\text{uni}}^{\text{human}}|$ measures how important the missing non-additive interactions are to humans for each segment. We use $|\text{TVA}_O - \hat{y}_{\text{uni}}^{\text{human}}|$ to answer the question: does MRO learn more non-additive interactions when $|\text{TVA}_O - \hat{y}_{\text{uni}}^{\text{human}}|$ is larger, i.e., when humans require non-additive interactions?

6.5.1 Additional Study Details

In addition to the three unimodal and the trimodal combinations we explored bimodal combinations: 1) the muted video with the transcript (TV); 2) the muted video with the low-pass filtered audio (VA); 3) the transcript with the low-pass filtered audio (TA); 4) for comparison the original audio with the transcript (TA_O).

Reliability: We report two types reliabilities: the averaged pairwise reliability between two random raters (ICC(2,1)) and the effective reliability of the mean over k=8 raters (ICC(2, k-1)). Pairwise and effective reliability address different purposes: pairwise is needed to determine how many raters are needed to achieve a targeted effective reliability [158]. Averaging over raters is important as emotional dimensions are subjective and difficult to annotate (especially when modalities are missing). The effective reliability describes how reliable the mean over the raters is, i.e., if we were to draw a new set of ratings and were to average them, how similar is this new mean to our current mean.

Except for transcripts-only (T) on arousal and acoustic-only (A) on valence, all pairwise reliabilities are moderate (between 0.5 and 0.75) [94], see Table 6.3. The effective reliability [158] of the mean over k raters as measured by ICC(2, k-1) is excellent (above 0.9) for all combinations. Instead of directly taking the mean over the raters, we apply, as common in affective computing, a z-normalization for each rater [18, 189] and take a weighted mean [62] over the raters.

Compensation: All raters are paid the same fixed amount, leading to an average hourly rate

Combination	Avg. ICC(2, 1)		ICC(2, k-1)	
	Arousal	Valence	Arousal	Valence
T	0.36	0.55	0.96	0.98
V	0.52	0.64	0.98	0.99
A	0.57	0.38	0.98	0.96
TV	0.48	0.62	0.97	0.98
VA	0.56	0.61	0.98	0.98
TA	0.60	0.54	0.98	0.98
TA_O	0.55	0.62	0.98	0.98
TVA_O	0.56	0.64	0.98	0.99

Table 6.3: Pairwise and effective reliability across the eight combinations. ICC is calculated with the R package psych.

of 11.14 USD/h.

6.6 Results and Discussion

Sanity Check: Before evaluating MRO on more complex datasets, we conduct a sanity check on two simpler datasets: $x_T + x_V + x_A$ which requires only unimodal contributions (we refer to it as **Sanity Check Unimodal**) and $x_T x_V + x_T x_A + x_V x_A$ which requires only bimodal interactions (we refer to it as **Sanity Check Bimodal**). Figure 6.3 shows that the joint and the routing model do not separate unimodal, bimodal, and trimodal interactions well as $|UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})| + |BI(\hat{y}_{\text{tri}})|$ is high. As expected, sMRO and MRO separate the interacts almost perfectly as $|UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})| + |BI(\hat{y}_{\text{tri}})|$ is very close to 0.

To test how many epochs are needed to minimize $|UC(\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}})| + |BI(\hat{y}_{\text{tri}})|$, we evaluate

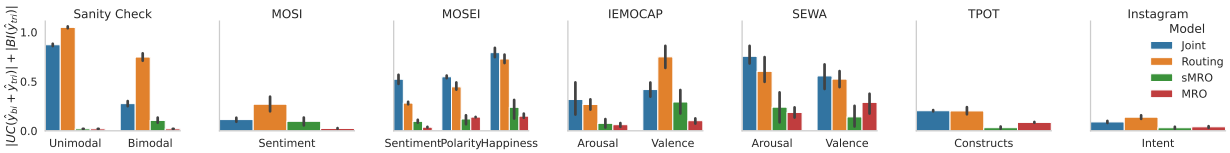
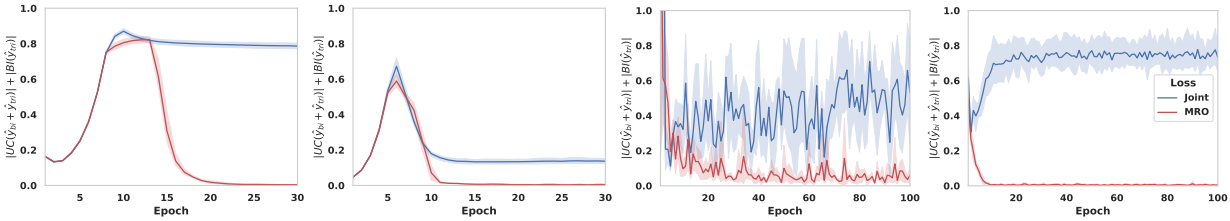


Figure 6.3: Average $|UC(\hat{y}_{\text{bi}}+\hat{y}_{\text{tri}})|+|BI(\hat{y}_{\text{tri}})|$ for all models and datasets. Lower values indicate a better separation of unimodal, bimodal, and trimodal contributions.



(a) Sanity Check Unimodal (b) Sanity Check Bimodal (c) Arousal on IEMOCAP (d) Valence on IEMOCAP

Figure 6.4: $|UC(\hat{y}_{\text{bi}}+\hat{y}_{\text{tri}})|+|BI(\hat{y}_{\text{tri}})|$ for the same model optimized with either $L(y, \hat{y}_{\text{uni}}+\hat{y}_{\text{bi}}+\hat{y}_{\text{tri}})$ (Joint, in blue) or with MRO (in red). Lower values indicate a better separation of unimodal, bimodal, and trimodal interactions.

it after each epoch. The results in Figure 6.4a show that the separation during the first epochs becomes worse as $\hat{y}_{\text{uni}}$ has not yet learned much, meaning $\hat{y}_{\text{bi}}$ and $\hat{y}_{\text{tri}}$ try to predict unimodal contributions which increases $|UC(\hat{y}_{\text{bi}}+\hat{y}_{\text{tri}})|$. However, after a few epochs the separation becomes better and $|UC(\hat{y}_{\text{bi}}+\hat{y}_{\text{tri}})|+|BI(\hat{y}_{\text{tri}})|$ reaches 0. The same can be observed for the bimodal sanity check in Figure 6.4b.

MRO significantly reduces $|UC(\hat{y}_{\text{bi}}+\hat{y}_{\text{tri}})|+|BI(\hat{y}_{\text{tri}})|$. Similar to the sanity check on simpler dataset, we want that $|UC(\hat{y}_{\text{bi}}+\hat{y}_{\text{tri}})|+|BI(\hat{y}_{\text{tri}})|$ is as small as possible. For easier comparison across datasets, we normalize $|UC(\hat{y}_{\text{bi}}+\hat{y}_{\text{tri}})|+|BI(\hat{y}_{\text{tri}})|$ by the standard deviation of the ground truth from the training set. Figure 6.3 shows that sMRO and MRO significantly reduce $|UC(\hat{y}_{\text{bi}}+\hat{y}_{\text{tri}})|+|BI(\hat{y}_{\text{tri}})|$ compared to models optimized with $L(y, \hat{y}_{\text{uni}}+\hat{y}_{\text{bi}}+\hat{y}_{\text{tri}})$ (Joint) and the routing model.

As it is computationally very expensive to evaluate $|UC(\hat{y}_{\text{bi}}+\hat{y}_{\text{tri}})|+|BI(\hat{y}_{\text{tri}})|$ after each

epoch, we plot it only for arousal and valence on IEMOCAP in Figure 6.4c and Figure 6.4d. We focus on IEMOCAP as we also conduct the perception study on it, see Section 6.5. While the plot for arousal in Figure 6.4c is a bit noisy, MRO quickly reduces $|UC(\hat{y}_{\text{bi}} + tri)| + |BI(\hat{y}_{\text{tri}})|$. The same can be observed for valence in Figure 6.4d.

MRO does not degrade performance. The secondary goal of MRO is not degrading performance. Table 6.4 lists the models' performance. Models optimized with MRO are in no case significantly worse than any other model. However, they are statistically significantly better than the joint model for valence on SEWA and happiness on MOSEI.

MRO might generalize slightly better because, similar to structural risk minimization [192], it prioritizes simpler models and relies on more complex multimodal models only when needed. Another reason is that MRO has similar effects as having auxiliary unimodal loss functions which seems beneficial for multimodal models [196, 223].

Ablating $\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}}$ decreases performance. We quantify the average performance impact of post-hoc removing $\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}}$ across datasets, i.e., $\hat{y} = \hat{y}_{\text{uni}}$. When comparing Table 6.5 with Table 6.4, we observe that removing $\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}}$ (the non-additive predictions), hurts performance. While additive contributions are very important, non-additive interactions are needed for best performance.

MRO learns more non-additive interactions when two modalities are informative. The TPOT dataset has human judgments for how important modalities are to confirm the current affective state [208]. Three importance levels were annotated: 1) a modality is sufficient to confirm the affective state (while ignoring other modalities), 2) a modality contains relevant information for the affective state (information from a second modality is needed), and 3) a modality contains no information for the current affective state.

We hypothesize that MRO uses more non-additive interactions ($\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}}$) for samples with at least two informative modalities (relevant or sufficient) compared to samples with only one informative modality. To measure whether $\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}}$ are used more, we calculate how much

	Tri	Routing	Joint	sMRO	MRO
MOSI (Pearson's r)					
Sentiment	0.662	0.658	0.657	0.656	0.661
MOSEI (Pearson's r)					
Sentiment	0.723	0.727	0.727	0.726	0.727
Polarity	0.599	0.597	0.606	0.593	0.605
Happiness	0.637	0.642	0.637	0.630	0.641
IEMOCAP (Concordance Correlation Coefficient)					
Arousal	0.588	0.613	0.622	0.624	0.611
Valence	0.647	0.655	0.624	0.603	0.634
SEWA (Concordance Correlation Coefficient)					
Arousal	0.317	0.263	0.293	0.292	0.304
Valence	0.268	0.335	0.268	0.310	0.337
TPOT (Accuracy)					
Constructs	0.565	0.554	0.566	0.566	0.574
Instagram (macro ROC AUC)					
Intent	0.876	0.731	0.891	0.888	0.891
Mean	0.588	0.595	0.589	0.589	**0.599**

Table 6.4: Average performance over the test folds. Higher is better.

	sMRO	MRO
Mean	0.577	0.587

Table 6.5: Average performance when post-hoc removing $\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}}$, i.e., $\hat{y} = \hat{y}_{\text{uni}}$.

the softmax probabilities (TPOT is a classification task) change when removing $\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}}$, i.e., $\sum_{k=1}^{4} |\text{softmax}(\hat{y})^{(k)} - \text{softmax}(\hat{y}_{\text{uni}})^{(k)}|$ where k indexes the probability vector for the four classes. The means of samples with two informative modalities (0.299) and only one informative modality (0.264) are significantly different according to an independent t-test, $t(2671) = 5.059, p < 0.001$. This suggests that MRO not only mathematically separates unimodal, bimodal, and trimodal interactions but that its separation also correlates with human assessments. Further, this observation provides evidence that models are more likely to learn non-additive interactions when several modalities are themselves informative.

MRO learns more non-additive interactions when modalities amplify each other. We included the Instagram dataset [96] because it has modality interaction annotations (semiotic modes) that are inspired by Kloepfer [89]. To test whether $\hat{y}_{\text{bi}}$ (this dataset has only two modalities) contributes more depending on the semiotic mode (parallel, amplifying, and divergent), we conduct a one-way ANOVA on the probability changes when removing $\hat{y}_{\text{bi}}$. The means between the semiotic modes are significantly different, $F(2, 1296) = 5.059, p = 0.006$, with the highest average change for amplifying (0.317), followed by parallel (0.272), and then divergent (0.256). The means between amplifying and parallel are significantly different $t(1297) = 2.432, p = 0.015$ as well as between amplifying and divergent $t(1297) = 2.874, p = 0.004$. Similar to the results on TPOT, it is confirming that MRO learned significantly larger non-additive contributions ($\hat{y}_{\text{bi}}$) for amplifying than for parallel. A possible explanation why diverging seems to require the least non-additive interactions is that the definition of diverging requires that only the meaning of the modalities is opposing but it does not specify how the combined meaning is formed. Even if the combined meaning of Figure 6.1 was neutral (additive), the semiotic mode is still divergent.

MRO learns non-additive interaction when humans need non-additive interactions. The additive model $\hat{y}_{\text{uni}}^{\text{human}}$ of predicting the mutlimodal ratings TVA_O given the uni-modal ratings, fits very well ($r^2 = 0.85$ for arousal and $r^2 = 0.85$ for valence) which is inline with similar prior

work [152]. Even though our multimodal model is not on par with $\hat{y}_{\text{uni}}^{\text{human}}$ ($r^2 = 0.68$ for arousal and $r^2 = 0.66$ for valence), we observe a significant correlation of $r(98) = 0.202, p = 0.043$ for valence between $|\text{TVA}_O - \hat{y}_{\text{uni}}^{\text{human}}|$ (the missing non-additive interactions) and $|\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}}|$ (non-additive contributions). This indicates that $\hat{y}_{\text{bi}} + \hat{y}_{\text{tri}}$ learned non-additive interactions that cannot be explained by $\hat{y}_{\text{uni}}^{\text{human}}$. For arousal, we do not observe a significant correlation, potentially because the optimization seems far nosier for arousal then for valence, see Figure 6.4c.

6.7 Conclusion

We proposed MRO to explicitly learn and separate unimodal, bimodal, and trimodal interactions in a multimodal model. This separation is essential for quantifying how much a model uses multimodal interactions and is a step towards more interpretable models. Based on prior work [69] we proposed a new evaluation metrics to quantify whether a trimodal models uses unimodal, bimodal, and trimodal interactions. Empirically, we observed that MRO successfully separated unimodal, bimodal, and trimodal interactions while not degrading predictive performance. Beyond the empirical evaluation, MRO learns non-additive interactions in accordance with human judgments on three datasets.

This and the previous chapter focused on two core multimodal phenomena. The next chapter, Chapter 7, will extend our efforts on those two phenomena by studying which contributions of a modality are unique to it and which are redundantly expressed through multiple modalities.

Chapter 7

Modality Contribution Transparency

The previous two chapters focused on two core multimodal model mechanics. In this chapter, we focus on the last core multimodal model mechanics for this thesis: learning a model that factorizes its output into 1) unique modality contributions that can be derived from only one modality and 2) pairwise redundant contributions that can be derived from at least two modalities. This allows us to understand better how uniquely useful modalities are and how much two modalities share for a task.

Humans express their affective states simultaneously through multiple modalities, often in a redundant manner, giving people and computational models multiple ways to perceive affective states, e.g., someone might be perceived as being in a positive state because they are visually smiling or because they are also simultaneously audibly laughing. This redundancy makes it challenging in multimodal machine learning to determine what a modality uniquely contributes and what can redundantly be contributed by multiple modalities when predicting affective states. Computational models that learn what is uniquely predicted by one modality and what can redundantly be predicted by multiple modalities have the potential to a) help interpretability as machine learning practitioners can inspect these two types of contributions and b) improve robustness to missing modalities as they have to learn redundant contributions from multiple modalities, for example, by relying on both visual smiles and audible laughter.

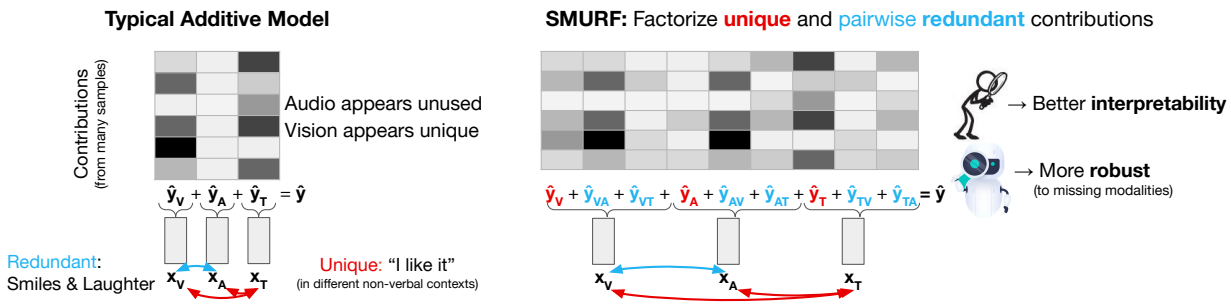


Figure 7.1: Typical additive models might ignore a redundant modality, which can mislead machine learning practitioners and make the model less robust. SMURF factorizes unique and pairwise redundant contributions, which includes extracting pairwise redundant contributions from both modalities. SMURF's factorization has the potential to improve interpretability and also improves robustness to missing modalities.

Learning unique and redundant contributions is difficult for even additive models [65], such as the models used in Chapter 5, that express their prediction as the sum of separately processed modality contributions, as the contributions of two modalities can be correlated. To overcome this difficulty, two challenges need to be addressed. First, we need to ensure that a model derives the same information from two modalities simultaneously to represent pairwise redundant contributions, as typical models might ignore a highly correlated modality, such as vision being ignored in multimodal machine translation [1, 220, 224] as illustrated in Figure 4.1 on the left. And second, we need a factorization to separate unique and redundant contributions that does not degrade the model's predictive performance while also directly describing how those contributions relate to the model's prediction to make them easier to inspect for machine learning practitioners.

In this paper, we propose SMURF (Statistical Modality Uniqueness and Redundancy Factorization), an additive model that learns to factorize its predictions into unique contributions and pairwise redundant contributions by expressing them as the sum of a) unique contributions that are uncorrelated with all other modalities and b) redundant contributions that are maximally

correlated between pairs of modalities. SMURF learns these two types of contributions by maximizing the covariance between pairwise redundant contributions and by minimizing the absolute value of the covariance between a modality's unique contributions and its pairwise redundant contributions. One crucial implication of maximizing the covariance between pairs of redundant contributions is that SMURF extracts the same information from both modalities, e.g., SMURF relies on both the audible laughter and the visual smiles. Our evaluation of SMURF is structured by two research questions:

RQ1 Can SMURF learn its factorization while not degrading predictive performance?

RQ2 Does SMURF's maximization of pairwise redundancies improve its robustness to missing modalities?

We evaluate that SMURF does not degrade predictive performance on eight affective datasets and one synthetic dataset. We further use the synthetic dataset, on which we have a ground truth of the unique and pairwise redundant contributions, to verify SMURF's learned factorization in both a bimodal and trimodal setting (RQ1). We hypothesize that SMURF might be more robust to missing modalities at test time as it maximizes pairwise redundant contributions (RQ2): if visual smiles would make audible laughter completely redundant, SMURF will still rely on the audible laughter, which becomes crucial when the vision modality is missing at test time. Finally, we explore whether SMURF has the potential to improve interpretability by testing whether its unique and pairwise redundant contributions have a significant relationship with human judgment studies on three datasets.

7.1 Related Work

We cover three related topics: approaches that quantify the amount of unique and redundant information, coordinated representations to learn redundancy, and multimodal collinearity.

Unique and Redundant Information: Mutual information [28, 230] and Partial Information

Decompositions (PID) [109] have been used to train neural networks that estimate statistics of those theoretic measures, for example, the redundant amount of information in bits between two modalities for a task [28]. While these models predict a task and estimate statistics of information theoretical measures, they do not couple these two goals as much as SMURF does, meaning it is unclear, for example, if the estimated amount of redundancy is also used for the task prediction. SMURF overcomes this issue by directly factorizing its prediction as the sum of unique and pairwise redundant contributions. We want to note that estimating the amount of redundancy expressed as bits and estimating the pairwise redundant contributions expressed as additive values making up the model's predictions are related but different tasks. Further, some of the PID redundancy measures have unexpected properties as two variables can have redundant information for a task, even when the two variables are independent (and are therefore also uncorrelated) [93].

Coordinated Representation for Redundancy: Coordinated representation learning tries to learn a representation by minimizing a similarity measure between two modalities [6], meaning this representation focuses on redundant information simultaneously present in both modalities. Many different similarity measures have been proposed to learn such a coordinated representation [2, 121, 195] and previous work also learned representations that represent what the other modality does not contain [66, 214, 227]. The main difference between these representation approaches and SMURF is that SMURF is applied to the output of a model to directly express the prediction as the sum of unique and pairwise redundant contributions. SMURF's approach has two advantages: 1) it ensures that the unique and pairwise redundant contributions impact the prediction (embedding spaces undergo further layers which might learn not to use, for example, the correlated information), and 2) it is easier to inspect the low-dimensional contributions for a machine learning practitioners than to inspect the high-dimensional embedding spaces.

Multimodal Collinearity: Multimodal models can learn to ignore a modality even though the modality contains predictive information when other modalities provide the same and more

information. This has been observed in multimodal machine translation [205] and multimodal sentiment recognition [220, 224]. Ignoring an informative modality is related to collinearity in statistics, where at least one feature is a linear combination of the other features [1]. In such a situation, a model might use the redundant feature to some degree or ignore it. SMURF avoids this ambiguity by maximally relying on pairwise redundant contributions, meaning a redundant modality will be used by SMURF. This can potentially improve interpretability and robustness to missing modalities, for example, when only a completely redundant modality is available at test time.

7.2 Problem: Unique and Pairwise Redundant Contributions

Our goal is training a model that factorizes its prediction into unique and pairwise redundant modality contributions where the prediction is directly derived from those contributions to make them easier to inspect for machine learning practitioners. Inspired by the interpretable but linear factor analysis in statistics, which predicts a variable as the sum of uncorrelated factors [216], we 1) express the prediction $\hat{y}$ as the sum of the unique and pairwise redundant contributions and 2) define the unique contributions as being uncorrelated with other modalities and the pairwise redundant contributions as being maximally correlated between modalities.

For N samples expressed as a tuple of vectors $(\mathbf{x}_A, \mathbf{x}_B)$ from two modalities A and B, our goal is to factorize the predictions $\hat{\mathbf{y}}$ as

$$\hat{\mathbf{y}} = \hat{\mathbf{y}}_A + \hat{\mathbf{y}}_B + \hat{\mathbf{y}}_{AB} + \hat{\mathbf{y}}_{BA} \tag{7.1}$$

where we use a single-letter subscript to denote the unique contributions ($\hat{\mathbf{y}}_A$ and $\hat{\mathbf{y}}_B$) and a two-letter subscript to denote the pairwise redundant contributions derived from the first mentioned modality with the other modality ($\hat{\mathbf{y}}_{AB}$ and $\hat{\mathbf{y}}_{BA}$), meaning $\hat{\mathbf{y}}_{AB}$ is derived from modality A and represents what the two modalities A and B redundantly explain about $\mathbf{y}$. The unique contributions should be uncorrelated with each other, meaning their Pearson's correlation coefficient r

should be 0.0

$$\min |r(\hat{\mathbf{y}}_A, \hat{\mathbf{y}}_B)|. \tag{7.2}$$

Further, the unique contributions should also be uncorrelated with the pairwise redundant contributions, which are represented as the sum $\hat{\mathbf{y}}_{AB} + \hat{\mathbf{y}}_{BA}$

$$\min |r(\hat{\mathbf{y}}_A, \hat{\mathbf{y}}_{AB} + \hat{\mathbf{y}}_{BA})| \tag{7.3}$$

$$\min |r(\hat{\mathbf{y}}_B, \hat{\mathbf{y}}_{AB} + \hat{\mathbf{y}}_{BA})| \, . \tag{7.4}$$

For the sum of $\hat{\mathbf{y}}_{AB}$ and $\hat{\mathbf{y}}_{BA}$ to represent the pairwise redundant contributions, they should be maximally correlated to represent the same information

$$\max r(\hat{\mathbf{y}}_{AB}, \hat{\mathbf{y}}_{BA}) \, . \tag{7.5}$$

To ensure that $\hat{\mathbf{y}}_A$ and $\hat{\mathbf{y}}_{AB}$ are derived from only A, they are predicted using a unimodal model that uses only $\mathbf{x}_A$ (and similar for the contributions from B)

$$[\hat{\mathbf{y}}_A, \ \hat{\mathbf{y}}_{AB}] = f_{\theta_A}(\mathbf{x}_A) \tag{7.6}$$

$$[\hat{\mathbf{y}}_B, \ \hat{\mathbf{y}}_{BA}] = f_{\theta_B}(\mathbf{x}_B) \tag{7.7}$$

where f_θ is model with learnable parameters θ. The next section explains how SMURF learns these models.

7.3 SMURF

We first explain SMURF (Statistical Modality Uniqueness and Redundancy Factorization) with two modalities for regression where the prediction $\hat{\mathbf{y}}$ is a vector of $N \times 1$. We then outline SMURF more generally with m modalities.

7.3.1 Bimodal SMURF

SMURF learns its factorization for an additive model as in Equation 7.1 through two auxiliary loss terms (L_{uncor} and L_{cor}) to achieve the two desired properties: the uncorrelated unique contributions (Equation 7.2 to Equation 7.4) and the correlated pairwise redundant contributions (Equation 7.5). The loss function of SMURF is

$$L(\mathbf{y}, \hat{\mathbf{y}}) + \lambda(L_{\text{uncor}} + L_{\text{cor}}) \tag{7.8}$$

where $L(\mathbf{y}, \hat{\mathbf{y}})$ is a downstream loss, $\mathbf{y}$ is the ground truth, and λ is a hyper-parameter determining the trade-off between the different loss terms.

As Pearson's r is differentiable, we could operationalize the two auxiliary loss terms through Equation 7.2 to Equation 7.5. Pearson's r is, however, scale-invariant, meaning its value can widely fluctuate when one of the two contributions is close to zero, which makes the optimization unstable. Instead, we use the sample covariance

$$\text{cov}(\hat{\mathbf{y}}_A, \hat{\mathbf{y}}_{AB}) = \frac{\sum_{i=1}^{N}(\hat{\mathbf{y}}_A^i - \bar{\hat{\mathbf{y}}}_A)(\hat{\mathbf{y}}_{AB}^i - \bar{\hat{\mathbf{y}}}_{AB})}{N-1}, \tag{7.9}$$

where $\bar{\hat{\mathbf{y}}}$ is the average over the N samples in $\hat{\mathbf{y}}$, to uncorrelate the unique contribution with the modality's pairwise redundant contributions by minimizing

$$L_{\text{uncor}} = \frac{1}{2}\big(|\text{cov}(\hat{\mathbf{y}}_A,\ \hat{\mathbf{y}}_{AB})| + |\text{cov}(\hat{\mathbf{y}}_B,\ \hat{\mathbf{y}}_{BA})|\big)\ . \tag{7.10}$$

Figure 7.2 visually illustrates that L_{uncor} uncorrelates the unique contribution from the modality's pairwise redundant contributions. To learn the pairwise redundant contributions ($\hat{\mathbf{y}}_{AB}$ and $\hat{\mathbf{y}}_{AB}$) we minimize

$$L_{\text{cor}} = -\text{cov}(\hat{\mathbf{y}}_{AB}, \hat{\mathbf{y}}_{BA}) + \frac{1}{2}\text{var}(\hat{\mathbf{y}}_{AB})\text{var}(\hat{\mathbf{y}}_{BA}) \tag{7.11}$$

where var is the sample variance

$$\text{var}(\hat{\mathbf{y}}) = \frac{\sum_{i=1}^{N}(\hat{\mathbf{y}}^i - \bar{\hat{\mathbf{y}}})}{N-1}. \tag{7.12}$$

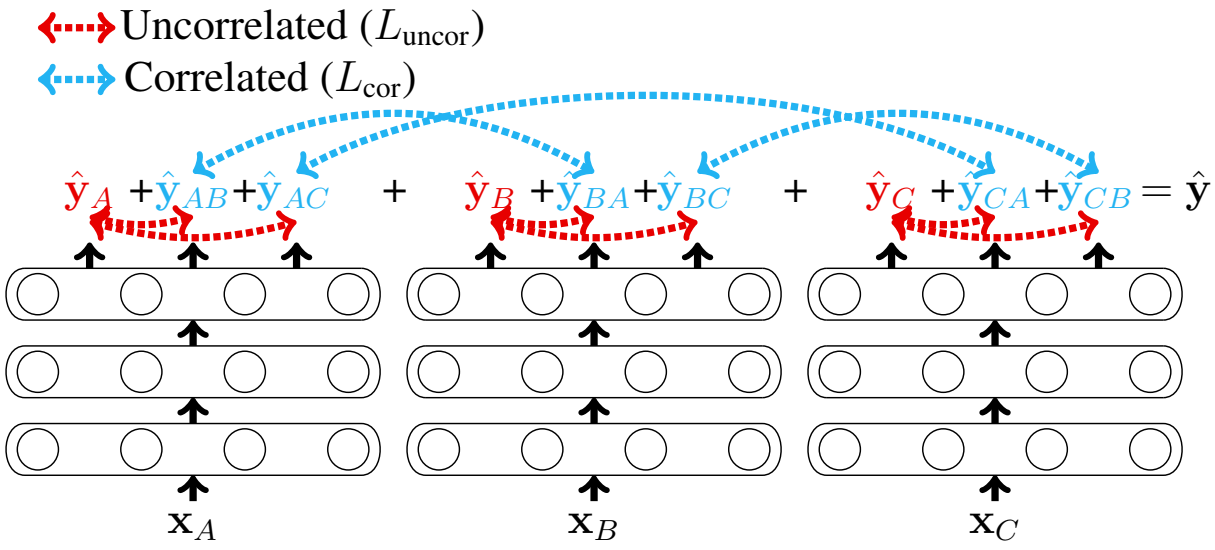


Figure 7.2: Illustration of SMURF for three modalities.

The first term of L_{cor} maximizes the covariance between the pairwise redundant contributions. The second term limits the individual variances as even weakly correlated contributions with increasingly larger variances would increase the covariance without increasing their correlation. Equation 7.11 is known as the scalar-version of an Hirschfeld-Gebelein-Ŕenyi (HGR) correlation [71] approximation proposed to learn maximally correlated representation in neural networks and was demonstrated to perform better than maximizing Pearson's r, the sample covariance, and canonical correlation analysis [121].

As L_{cor} maximizes the covariance between pairwise redundant contributions, the model is incentivized to use the pairwise redundant contributions as much as it can. Together with L_{uncor}, L_{cor} penalizes the model if it would learn correlated unique contributions.

7.3.2 m-modal SMURF

In the case of m modalities (A, B, ..., M), we learn again the unique contributions and but also all pairwise redundant contributions as illustrated in Figure 7.2 for $m = 3$. This means that the

model architecture becomes

$$[\hat{\mathbf{y}}_A,\ \hat{\mathbf{y}}_{AB},\ \ldots,\ \hat{\mathbf{y}}_{AM}] = f_{\theta_A}(\mathbf{x}_A) \tag{7.13}$$

$$\ldots$$

$$[\hat{\mathbf{y}}_M,\ \hat{\mathbf{y}}_{MA},\ \ldots,\ \hat{\mathbf{y}}_{MN}] = f_{\theta_M}(\mathbf{x}_M)\ . \tag{7.14}$$

L_{uncor} uncorrelates the unique contribution of modality I from all the pairwise redundant contributions with modality J

$$L_{\text{uncor}} = \alpha \sum_{(I,J),I\neq J} |\text{cov}(\hat{\mathbf{y}}_I, \hat{\mathbf{y}}_{IJ})| \tag{7.15}$$

where $\alpha = \frac{1}{m^2-m}$ is a normalization term to average over all the covariance terms so that the same λ can be used across bimodal and m-modal experiments. Similarly, L_{cor} maximizes the covariance between redundant contributions of all modality pairs (I, J)

$$L_{\text{cor}} = \beta \sum_{(I,J),I<J} -\text{cov}(\hat{\mathbf{y}}_{IJ},\ \hat{\mathbf{y}}_{JI}) + \frac{1}{2}\text{var}(\hat{\mathbf{y}}_{IJ})\text{var}(\hat{\mathbf{y}}_{JI})\ . \tag{7.16}$$

where $\beta = \frac{2}{m^2-m}$ is again a normalization term to average over the covariance terms.

7.4 Non-additive SMURF (MRO-SMURF)

So far we focused on additive models, but conceptually we can also extend SMURF to non-additive models. The focus of this chapter is SMURF for additive models but we want to highlight how SMURF can be combined with our work in Chapter 6. A non-additive model with modalities A, B, and C can learn non-additive interactions between pairs of modalities and between the triplet of modalities. Multimodal Residual Optimization (MRO) [209] was proposed in Chapter 6 to separate additive, pairwise non-additive (bimodal) and triplet non-additive (trimodal) interactions from each other. We can extend MRO by applying SMURF within the additive and within the bimodal interactions. This allows us to explore a) whether there are

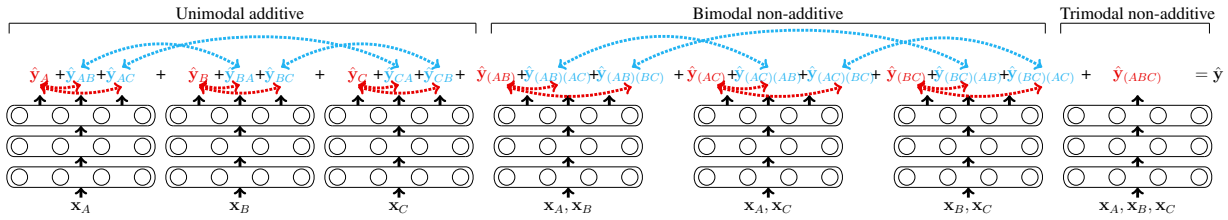


Figure 7.3: Illustration of combining SMURF and MRO for three modalities. MRO factorizes additive, bimodal non-additive and trimodal non-additive interactions and SMURF further factorizes the additive and bimodal non-additive interactions into unique and redundant contributions.

non-additive interactions between A and B that derive the same information as non-additive interactions between other modality pairs (redundant bimodal interactions) and b) what the unique non-additive interactions between modalities contribute to the prediction.

We have the following seven models when combining SMURF and MRO to factorize non-additive interactions for three modalities A, B, and C as illustrated in Figure 7.3: the three unimodal models

$$[\hat{\mathbf{y}}_A,\ \hat{\mathbf{y}}_{AB},\ \hat{\mathbf{y}}_{AC}] = f_{\theta_A}(\mathbf{x}_A) \tag{7.17}$$

$$[\hat{\mathbf{y}}_B,\ \hat{\mathbf{y}}_{BA},\ \hat{\mathbf{y}}_{BC}] = f_{\theta_B}(\mathbf{x}_B) \tag{7.18}$$

$$[\hat{\mathbf{y}}_C,\ \hat{\mathbf{y}}_{CA},\ \hat{\mathbf{y}}_{CB}] = f_{\theta_C}(\mathbf{x}_C)\ ; \tag{7.19}$$

the three bimodal models

$$[\hat{\mathbf{y}}_{(AB)},\ \hat{\mathbf{y}}_{(AB)(AC)},\ \hat{\mathbf{y}}_{(AB)(BC)}] = f_{\theta_{AB}}(\mathbf{x}_A, \mathbf{x}_B) \tag{7.20}$$

$$[\hat{\mathbf{y}}_{(AC)},\ \hat{\mathbf{y}}_{(AC)(AB)},\ \hat{\mathbf{y}}_{(AC)(BC)}] = f_{\theta_{AC}}(\mathbf{x}_A, \mathbf{x}_C) \tag{7.21}$$

$$[\hat{\mathbf{y}}_{(BC)},\ \hat{\mathbf{y}}_{(BC)(AB)},\ \hat{\mathbf{y}}_{(BC)(AC)}] = f_{\theta_{BC}}(\mathbf{x}_B, \mathbf{x}_C)\ , \tag{7.22}$$

where $\hat{\mathbf{y}}_{(AB)}$ are the unique non-additive contributions from the modality pair AB, $\hat{\mathbf{y}}_{(AB)(AC)}$ are the redundant non-additive contributions between the pairs AB and AC; and the trimodal model

$$\hat{\mathbf{y}}_{(ABC)} = f_{\theta_{ABC}}(\mathbf{x}_A, \mathbf{x}_B, \mathbf{x}_C)\ . \tag{7.23}$$

To define the loss function of MRO, we define $\hat{\mathbf{y}}_{\text{uni}}$ as the sum of all unimodal contributions (the sum over all outputs from the three unimodal models)

$$\hat{\mathbf{y}}_{\text{uni}} = \hat{\mathbf{y}}_A + \hat{\mathbf{y}}_{AB} + \hat{\mathbf{y}}_{AC} + \hat{\mathbf{y}}_B + \hat{\mathbf{y}}_{BA} + \hat{\mathbf{y}}_{BC} + \hat{\mathbf{y}}_C + \hat{\mathbf{y}}_{CA} + \hat{\mathbf{y}}_{CB} \tag{7.24}$$

and $\hat{\mathbf{y}}_{\text{bi}}$ as the sum of all bimodal contributions (the sum over all outputs from the three bimodal models)

$$\begin{aligned}\hat{\mathbf{y}}_{\text{bi}} = &\hat{\mathbf{y}}_{(AB)} + \hat{\mathbf{y}}_{(AB)(AC)} + \hat{\mathbf{y}}_{(AB)(BC)} + \hat{\mathbf{y}}_{(AC)} + \hat{\mathbf{y}}_{(AC)(AB)} + \hat{\mathbf{y}}_{(AC)(BC)} \\ &+ \hat{\mathbf{y}}_{(BC)} + \hat{\mathbf{y}}_{(BC)(AB)} + \hat{\mathbf{y}}_{(BC)(AC)} \ .\end{aligned} \tag{7.25}$$

To ensure that the bimodal models learn only the non-additive bimodal interactions, we use the MRO loss formulation to define the loss L as

$$L(\mathbf{y}, \hat{\mathbf{y}}) = L(\mathbf{y}, \hat{\mathbf{y}}_{\text{uni}}) + L(\mathbf{y}, sg(\hat{\mathbf{y}}_{\text{uni}}) + \hat{\mathbf{y}}_{\text{bi}}) + L(\mathbf{y}, sg(\hat{\mathbf{y}}_{\text{uni}} + \hat{\mathbf{y}}_{\text{bi}}) + \hat{\mathbf{y}}_{(ABC)}) \tag{7.26}$$

where sg means stop gradient [154] which prevents back-propagation through sg's arguments.

To achieve the factorization constraints from SMURF within the unimodal and within the bimodal contributions, we define the two auxiliary loss terms L_{uncor} and L_{cor} as following

$$
\begin{aligned}
L_{\text{uncor}} = \frac{1}{12}\big(&\text{cov}(\hat{\mathbf{y}}_A, \hat{\mathbf{y}}_{AB}) + \text{cov}(\hat{\mathbf{y}}_A, \hat{\mathbf{y}}_{AB}) \\
&+ \text{cov}(\hat{\mathbf{y}}_B, \hat{\mathbf{y}}_{BA}) + \text{cov}(\hat{\mathbf{y}}_B, \hat{\mathbf{y}}_{BC}) \\
&+ \text{cov}(\hat{\mathbf{y}}_C, \hat{\mathbf{y}}_{CA}) + \text{cov}(\hat{\mathbf{y}}_C, \hat{\mathbf{y}}_{CB}) \\
&+ \text{cov}(\hat{\mathbf{y}}_{(AB)}, \hat{\mathbf{y}}_{(AB)(AC)}) \\
&+ \text{cov}(\hat{\mathbf{y}}_{(AB)}, \hat{\mathbf{y}}_{(AB)(BC)}) \\
&+ \text{cov}(\hat{\mathbf{y}}_{(AC)}, \hat{\mathbf{y}}_{(AC)(AB)}) \\
&+ \text{cov}(\hat{\mathbf{y}}_{(AC)}, \hat{\mathbf{y}}_{(AC)(BC)}) \\
&+ \text{cov}(\hat{\mathbf{y}}_{(BC)}, \hat{\mathbf{y}}_{(BC)(AB)}) \\
&+ \text{cov}(\hat{\mathbf{y}}_{(BC)}, \hat{\mathbf{y}}_{(BC)(AC)})\big)
\end{aligned} \tag{7.27}
$$

$$
\begin{aligned}
L_{\text{cor}} = \frac{1}{6}\big(&\text{cov2}(\hat{\mathbf{y}}_{AB}, \hat{\mathbf{y}}_{BA}) \\
&+ \text{cov2}(\hat{\mathbf{y}}_{AC}, \hat{\mathbf{y}}_{CA}) \\
&+ \text{cov2}(\hat{\mathbf{y}}_{BC}, \hat{\mathbf{y}}_{CB}) \\
&+ \text{cov2}(\hat{\mathbf{y}}_{(AB)(AC)}, \hat{\mathbf{y}}_{(AC)(AB)}) \\
&+ \text{cov2}(\hat{\mathbf{y}}_{(AB)(BC)}, \hat{\mathbf{y}}_{(BC)(AB)}) \\
&+ \text{cov2}(\hat{\mathbf{y}}_{(AC)(BC)}, \hat{\mathbf{y}}_{(BC)(AC)})\big)
\end{aligned} \tag{7.28}
$$

where for brevity we use $\text{cov2}(\mathbf{a}, \mathbf{b})$ to refer to $-\text{cov}(\mathbf{a}, \mathbf{b}) + \frac{1}{2}\text{var}(\mathbf{a})\text{var}(\mathbf{b})$. Intuitively, we apply the same constraints as for for the additive trimodal SMURF but in addition also apply them for the bimodal non-additive interactions.

Table 7.1: Dataset characteristics.

Dataset	Tasks	Samples	Modalities (abbreviations)
MOSEI [221]	Sentiment and happiness (regression)	23.3k	audio (A), text (T), video (V)
MOSI [218]	Sentiment (regression)	2.2k	audio (A), text (T), video (V)
IEMOCAP [18]	Arousal and valence (regression)	4.8k	audio (A), text (T), video (V)
RECOLA [156]	Arousal and valence (regression)	1.0k	audio (A), ECG (E), video (V)
SEWA [189]	Arousal and valence (regression)	2.2k	audio (A), text (T), video (V)
UMEME [152]	Arousal and valence (regression)	1.6k	audio (A), text (T), video (V)
TPOT [131]	Four affective states (multiclass classification)	15.2k	audio (A), text (T), video (V)
VREED [176]	Arousal-valence quadrants (multiclass classification)	312	ECG (E), GSR (G), gaze (V)

7.5 Experimental Setup

7.5.1 Datasets

As affective states are often expressed through multiple modalities, we focus on eight affective datasets that include sentiment and emotion annotations. See Table 7.1 for a summary. To evaluate SMURF's factorization, we also create a synthetic dataset with a ground truth of the unique and pairwise redundant contributions.

MOSI [218] and MOSEI [221]: These two datasets consist of single-person YouTube videos where the person expresses an opinion, e.g., about a movie. In both cases, we predict the continuous sentiment ratings (MOSI-S and MOSEI-S) and the happiness intensity ratings on MOSEI (MOSEI-H).

IEMOCAP [18]: We use the improvised dyadic interactions of IEMOCAP and predict their continuous arousal (IEMOCAP-A) and valence (IEMOCAP-V) ratings separately for each person and utterance.

RECOLA [156]: This dataset consists of French-speaking dyadic interactions. Similar to IEMOCAP, we predict arousal (RECOLA-A) and valence (RECOLA-V) ratings for each person

and utterance.

SEWA [189]: This dataset consists of German-speaking dyadic interactions. As previously, we predict arousal (SEWA-A) and valence (SEWA-V) ratings for each person and utterance.

UMEME [152]: The University of Michigan Emotional McGurk Effect (UMEME) dataset contains a set of sentences enacted in different emotional settings. We predict arousal (UMEME-A) and valence (UMEME-V) separately for each enacted sentence. UMEME has further combinations of mismatched audio and video, e.g., the video from a positive enactment but the audio from a negative enactment. As we focus on more natural interactions, we exclude those mismatched combinations.

TPOT [131]: As previously detailed in Chapter 5, the Transitions in Parenting of Teens (TPOT) dataset contains video recordings of dyadic interactions between mothers and their adolescents. These interactions consist of segments annotated for four affective states (other, aggressive, dysphoric, and positive). We classify these segments for each person independently of the previous and following segments.

VREED [176]: VREED is a virtual reality dataset of people watching emotion-eliciting 360-degree videos. We predict the four quadrants of the arousal-valence space (the four combinations of low/high arousal and low/high valence) separately for each person and video.

Synthetic: To test whether SMURF recovers the intended unique and pairwise redundant contributions, we create a synthetic dataset. We define $\mathbf{y}$ as

$$\mathbf{y} = \mathbf{a} + \mathbf{b} + \mathbf{c} + \mathbf{d} + \mathbf{e} + \mathbf{f} \tag{7.29}$$

where $\mathbf{a}, \ldots, \mathbf{f} \sim \mathbb{N}(0, 1)$ are randomly sampled and we define the three modalities A, B, and C as containing three features each

$$A = [\mathbf{a}, \mathbf{d}, \mathbf{e}],\ B = [\mathbf{b}, \mathbf{d}, \mathbf{f}],\ \text{and}\ C = [\mathbf{c}, \mathbf{e}, \mathbf{f}] \tag{7.30}$$

where [] is the concatenation operator. $\mathbf{a}, \mathbf{b}$, and $\mathbf{c}$ are the unique contributions that are in only one modality, and $\mathbf{d}, \mathbf{e}$, and $\mathbf{f}$ are the pairwise redundant contributions that are in multiple modalities,

for example, $\mathbf{d}$ is in A and B. We use this synthetic test case in two settings: in the introduced trimodal setting where the model has access to modalities A, B, and C and in a bimodal setting, where the model has access to only modalities A and B.

Synthetic (Non-Additive): We validate MRO-SMURF on a synthetic dataset that has two non-additive interactions in the form of two multiplications between $\mathbf{k}$ and $\mathbf{l}$ and between $\mathbf{m}$ and $\mathbf{n}$

$$\mathbf{y} = \mathbf{kl} + \mathbf{mn} \tag{7.31}$$

where $\mathbf{k}, \ldots, \mathbf{n} \sim \mathbb{N}(0, 1)$ are randomly sampled. We define the three modalities A, B, and C as

$$A = [\mathbf{k}, \mathbf{m}],\ B = [\mathbf{k}, \mathbf{n}],\ \text{and}\ C = [\mathbf{l}]\ . \tag{7.32}$$

$\mathbf{kl}$ is a pairwise redundant non-additive interaction between modality pairs (A, C) and (B, C), while $\mathbf{mn}$ is a unique non-additive interaction present only in the modality pair (A, B). We evaluate MRO-SMURF only on this dataset as MRO learned few non-additive interactions on affective datasets in previous work [209].

7.5.2 Features

We use the same features for each modality: MiniLM-L12-v2's sentence embedding [197] for text, openSMILE's eGeMaPs [52] features for audio, and OpenFace 2.2 [7] features for video. In all cases, we use statistics, such as mean and standard deviation, to aggregate the extracted features at the labeled utterance level.

RECOLA does not provide transcripts of the spoken text: we use the heart rate related features (ECG) that are provided by the dataset authors as a third modality instead of the text modality. VREED does not share the raw audio-video recordings and has no transcripts. We use the author-provided features for eye-gaze, skin-conductance (GSR), and heart rate (ECG).

7.5.3 Baselines

We compare SMURF to two baselines: **SMURF w/o $L_{\text{cor}} + L_{\text{uncor}}$** and **E-HGR** [121]. For better comparison, we use the same additive architecture for all models [65]: all models express their predictions as in Equation 7.1. The only difference between the models are their auxiliary loss terms, such as, L_{cor} and L_{uncor} in Equation 7.8 for SMURF.

SMURF w/o $L_{\text{cor}}+L_{\text{uncor}}$: To evaluate whether SMURF's two auxiliary loss terms negatively impact the predictive performance, we compare it without having those two additional terms, i.e., $\lambda = 0$ in Equation 7.8.

E-HGR [121]: While we are not aware of other approaches that explicitly factorize the prediction into unique and redundant contributions, we compare SMURF to previous work that also uses the Hirschfeld-Gebelein-Renyi (HGR) correlation approximation [121] to maximize redundancy. Unlike SMURF, which maximizes HGR between the pairwise redundant contributions, E-HGR maximizes the HGR correlation in an embedding space between all modalities simultaneously. Intuitively, this means that E-HGR might learn fewer redundancies than SMURF as the "intersection" of three modalities can not be larger than the intersection of two modalities, e.g., while laughter and smiles frequently co-occur, they might not also be co-occurring together with textual expressions.

7.5.4 Evaluation Methodology

We evaluate SMURF through our two research questions: whether SMURF achieves its factorization while not degrading predictive performance (RQ1) and whether SMURF's maximization of redundant contributions makes it more robust to missing modalities (RQ2). We later also analyze whether SMURF's factorization relates to human judgments.

RQ1: Factorization and Performance: We evaluate SMURF's factorization on the synthetic dataset to verify that it achieves a) the two desired properties: uncorrelated unique contributions (Equation 7.2 to Equation 7.4) and correlated pairwise redundant contributions (Equa-

tion 7.5); and b) whether SMURF recovers the known ground truth of the unique and pairwise redundant contributions, e.g., SMURF's $\hat{\mathbf{y}}_A$ should correspond to the feature $\mathbf{a}$ on the synthetic dataset, meaning $r(\hat{\mathbf{y}}_A, \mathbf{a})$ should be high. On all nine datasets, we further report Pearson's r for regression tasks and accuracy for classification tasks to evaluate whether SMURF impacts predictive performance.

RQ2: Redundancy and Robustness: SMURF's covariance maximization L_{cor} explicitly encourages it to derive the same contributions from multiple modalities (pairwise redundancy), e.g., SMURF will focus on both the audible laughter and the visual smile even if they always cooccur. We hypothesize that this pairwise redundancy maximization might make SMURF more robust to missing modalities compared to other additive models, such as SMURF w/o $L_{\text{cor}}+L_{\text{uncor}}$ and E-HGR, that might learn pairwise redundancies to a lesser degree.

We evaluate this hypothesis by assuming that only one modality is available at test time, for example, only A, which means we have only $[\hat{\mathbf{y}}_A, \hat{\mathbf{y}}_{AB}]$ available. To test how much information is present in these available modality contributions, e.g., $[\hat{\mathbf{y}}_A, \hat{\mathbf{y}}_{AB}]$, we train a linear model to predict the original model output $\hat{\mathbf{y}}$ using only the available modality contributions as an input. If the original model did not extract all the pairwise redundant information from a modality, e.g., if a model ignored audible laughter and relied only on visual smiles, this linear model will perform poorly on the downstream task. This performance allows us to quantify whether the explicitly encouraged pairwise redundancy in SMURF improves robustness to missing modalities.

7.5.5 Implementation Details

The unimodal models (f_{θ_A}, f_{θ_B}, and f_{θ_B}) are instantiated as multi-layer perceptions (MLP) using PyTorch [144]. The bimodal and trimodal models of MRO-SMURF are instantiated using Tensor Fusion Networks [219], where modalities are combined by first learning unimodal representations and then taking the outer product between the unimodal representations [219]. All models are learned with the optimizer Adam [88] and have their hyper-parameters validated on the vali-

dation sets. Hyper-parameters include $\lambda \in [0.1, 1]$ (for both SMURF and E-HGR, SMURF w/o $L_{\text{cor}} + L_{\text{uncor}}$ uses $\lambda = 0$), the number of layers of the MLP and their number of neurons, the learning rate, and the strength of L2 weight decay. For SMURF, we minimize Equation 7.8 for each batch.

Early stopping is performed on the loss values on the validation set, which includes auxiliary loss terms for SMURF and E-HGR. The predictive performance metric on the validation set determines the best model of the hyperparameter search. We use a 5-fold testing for all datasets. These folds are person-independent except for MOSI and MOSEI for which we use the official test set.

SMURF for Classification

When predicting one out of c class labels, $\hat{y}$ can be represented as a vector of c logits for one sample, meaning $\hat{\mathbf{y}}$ is now a matrix of $N \times c$. We enforce the covariance constraints separately for each class label, e.g., the unique contribution for the i-th class should be uncorrelated of the pairwise redundant contributions of i-th class but we do not enforce covariance constraints between different classes. The two loss terms in case of c classes and two modalities are

$$L_{\text{uncor}} = \frac{1}{2c} \sum_{i \in [1,c]} |\text{cov}(\hat{\mathbf{y}}_A^i,\ \hat{\mathbf{y}}_{AB}^i)| + |\text{cov}(\hat{\mathbf{y}}_B^i,\ \hat{\mathbf{y}}_{BA}^i])| \tag{7.33}$$

$$L_{\text{cor}} = \frac{1}{c} \sum_{i \in [1,c]} -\text{cov}(\hat{\mathbf{y}}_{AB}^i,\ \hat{\mathbf{y}}_{BA}^i) + \frac{1}{2}\text{var}(\hat{\mathbf{y}}_{AB}^i)\text{var}(\hat{\mathbf{y}}_{BA}^i) \tag{7.34}$$

where we select the i-th column in $\hat{\mathbf{y}}$ with $\hat{\mathbf{y}}^i$.

7.6 Results and Discussion

7.6.1 RQ1: Factorization and Performance

The goal of the first research question is to evaluate whether SMURF achieves the factorization outlined in Section 7.2 without degrading predictive performance.

Pearson's r	Uncorrelated Unique Contributions $\downarrow$	Correlated Pairwise Redundant Contributions $\uparrow$	Ground Truth Contributions $\uparrow$
Averaged over	Equation 7.2 to Equation 7.4	Equation 7.5	$r(\hat{\mathbf{y}}_A, \mathbf{a}), r(\hat{\mathbf{y}}_B, \mathbf{b}), \ldots$
Bimodal Synthetic Dataset	0.001	0.925	0.945
Trimodal Synthetic Dataset	0.001	0.800	0.817

Table 7.2: Averaged Pearson's r on the bimodal and trimodal synthetic dataset: (left) within the contributions from SMURF; (right) between the contributions from SMURF and the known ground truth contributions.

Achieves factorization: The goal of SMURF is to learn a model that factorizes its predictions $\hat{\mathbf{y}}$ as in Equation 7.1 subject to the constraints that the unique contributions are uncorrelated with all other contributions (Equation 7.2 to Equation 7.4) and that the pairwise redundant contributions are highly correlated (Equation 7.5). We observe that SMURF achieves these two constraints on the bimodal and trimodal synthetic dataset, see Table 7.2 on the left.

On the synthetic dataset, we know which contributions SMURF should learn, i.e., we have a ground truth of the unique and pairwise redundant contributions. As reported in Table 7.2 on the right, we observe that SMURF's contributions correlate highly with the ground truth contributions.

Maintains predictive performance: SMURF statistically significantly improves performance in some cases, often achieves numerically the best performance, and never significantly decreases the predictive performance, see Table 7.3.

Factorizes non-additive interactions: We test whether MRO-SMURF is able to reconstruct the one unique ($\mathbf{mn}$) and the one pairwise redundant ($\mathbf{kl}$) non-additive interaction of the non-additive synthetic dataset. MRO-SMURF closely reconstructs the unique non-additive interaction, i.e., $r(\hat{\mathbf{y}}_{(AB)}, \mathbf{mn}) = 0.944$, and also the pairwise redundant non-additive interaction, i.e., $r(\hat{\mathbf{y}}_{(AC)(BC)}, \mathbf{kl}) = 0.999$ and $r(\hat{\mathbf{y}}_{(BC)(AC)}, \mathbf{kl}) = 0.998$, indicating that MRO-SMURF can conceptually learn to factorize non-additive interactions.

	E-HGR [121]	SMURF w/o $L_{\text{cor}} + L_{\text{uncor}}$	**SMURF** (proposed)
Pearson's r (regression)			
Synthetic	**1.000** $\pm$ 0.000	**1.000** $\pm$ 0.000	**1.000** $\pm$ 0.000
MOSEI-S	**0.713** $\pm$ 0.006	**0.713** $\pm$ 0.005	**0.713** $\pm$ 0.005
MOSEI-H	**0.623** $\pm$ 0.004	0.621 $\pm$ 0.006	**0.623** $\pm$ 0.008
MOSI-S	**0.700** $\pm$ 0.019	0.690 $\pm$ 0.025	0.690 $\pm$ 0.024
IEMOCAP-A	0.646 $\pm$ 0.061$^{\downarrow}$	0.663 $\pm$ 0.048	**0.665** $\pm$ 0.048
IEMOCAP-V	0.664 $\pm$ 0.078	0.662 $\pm$ 0.085	**0.667** $\pm$ 0.079
RECOLA-A	0.553 $\pm$ 0.054$^{\downarrow}$	0.586 $\pm$ 0.065	**0.596** $\pm$ 0.047
RECOLA-V	0.450 $\pm$ 0.107$^{\downarrow}$	0.418 $\pm$ 0.058$^{\downarrow}$	**0.474** $\pm$ 0.103
SEWA-A	0.499 $\pm$ 0.055	**0.525** $\pm$ 0.029	0.509 $\pm$ 0.048
SEWA-V	**0.472** $\pm$ 0.044	0.470 $\pm$ 0.036	0.465 $\pm$ 0.024
UMEME-A	0.660 $\pm$ 0.083$^{\downarrow}$	0.671 $\pm$ 0.060$^{\downarrow}$	**0.695** $\pm$ 0.072
UMEME-V	0.739 $\pm$ 0.046$^{\downarrow}$	0.740 $\pm$ 0.049	**0.750** $\pm$ 0.040
Accuracy (classification)			
TPOT	0.514 $\pm$ 0.020$^{\downarrow}$	0.521 $\pm$ 0.015$^{\downarrow}$	**0.528** $\pm$ 0.015
VREED	0.551 $\pm$ 0.073$^{\downarrow}$	0.602 $\pm$ 0.083	**0.608** $\pm$ 0.041

Table 7.3: Performance of the trimodal E-HGR, SMURF w/o $L_{\text{cor}} + L_{\text{uncor}}$, and SMURF. Higher is better in all cases and bold indicates best performance. $^{\downarrow}$ (and $^{\uparrow}$) indicates when a baseline performs significantly worse (or better) than SMURF at $\alpha = 0.05$.

7.6.2 RQ2: Redundancy and Robustness

The goal of the second research question is to evaluate whether SMURF's maximization of pairwise redundant contributions is beneficial for robustness to missing modalities.

More robust to missing modalities: To evaluate how robust SMURF and the baselines are to missing modalities, we evaluate the performance of using the learned contributions from just one modality, i.e., we try to reconstruct $\hat{\mathbf{y}}$ using the contributions from only one modality $[\hat{\mathbf{y}}_A, \hat{\mathbf{y}}_{AB}]$. The resulting downstream performance indicates to which degree these contributions contain redundant information in addition to the unique contributions. The original trimodal models should always derive unique contributions as the model would otherwise perform worse, so the main performance difference between the models should reflect the degree to which the trimodal model extracted redundant contributions from multiple modalities. In Table 7.4, we observe that SMURF's contributions lead to a better performance than for its baselines, indicating that SMURF is more robust to missing modalities.

7.7 Analysis

To explore whether SMURF might make a model more interpretable, we test whether its factorization relates to three existing human judgment studies. As there are no large-scale human judgments on the used datasets about uniqueness and redundancy, we compare SMURF's factorization to human judgments that might have indirectly a relationship to uniqueness and redundancy.

Human unimodal judgments: UMEME and IEMOCAP have human unimodal judgments of arousal and valence, where humans are, for example, given only the muted video to rate valence. All samples of UMEME have these judgments for the muted video ($\mathbf{y}_V^H$) and the audio ($\mathbf{y}_{AT}^H$; including the spoken texts). A subset of 100 samples has these judgments on IEMOCAP [209] for the muted video ($\mathbf{y}_V^H$), the low-pass filtered audio ($\mathbf{y}_A^H$), and the text ($\mathbf{y}_T^H$). While

	Available Modality	Approach		
		E-HGR [121]	SMURF w/o $L_{\text{cor}} + L_{\text{uncor}}$	**SMURF** (proposed)
Pearson's r (regression)				
Synthetic	A	$0.644 \pm 0.033^{\downarrow}$	$0.693 \pm 0.006^{\downarrow}$	**0.702** ± 0.003
	B	$0.578 \pm 0.037^{\downarrow}$	$0.693 \pm 0.008^{\downarrow}$	**0.703** ± 0.004
	C	$0.399 \pm 0.044^{\downarrow}$	$0.646 \pm 0.004^{\downarrow}$	**0.685** ± 0.004
MOSEI-S	A	0.321 ± 0.014	$0.313 \pm 0.012^{\downarrow}$	**0.324** ± 0.012
	T	0.690 ± 0.006	**0.691** ± 0.007	**0.691** ± 0.006
	V	**0.255** ± 0.009	0.243 ± 0.007	0.245 ± 0.007
MOSEI-H	A	$0.313 \pm 0.007^{\downarrow}$	$0.304 \pm 0.004^{\downarrow}$	**0.319** ± 0.011
	T	**0.356** ± 0.009	0.349 ± 0.009	0.355 ± 0.012
	V	**0.551** ± 0.003	0.550 ± 0.001	**0.551** ± 0.003
MOSI-S	A	$-0.067 \pm 0.056^{\downarrow}$	$-0.050 \pm 0.051^{\downarrow}$	**0.034** ± 0.069
	T	**0.707** ± 0.021	0.696 ± 0.028	0.697 ± 0.023
	V	0.073 ± 0.030	0.082 ± 0.030	**0.088** ± 0.039
IEMOCAP-A	A	$0.632 \pm 0.048^{\downarrow}$	0.644 ± 0.052	**0.652** ± 0.059
	T	0.322 ± 0.053	$0.301 \pm 0.029^{\downarrow}$	**0.334** ± 0.020
	V	0.357 ± 0.138	0.348 ± 0.163	**0.358** ± 0.133
IEMOCAP-V	A	0.444 ± 0.099	$0.425 \pm 0.085^{\downarrow}$	**0.448** ± 0.093
	T	**0.554** ± 0.041	$0.541 \pm 0.046^{\downarrow}$	0.552 ± 0.049
	V	**0.416** ± 0.176	0.408 ± 0.182	**0.416** ± 0.168
RECOLA-A	A	$0.530 \pm 0.044^{\downarrow}$	$0.544 \pm 0.047^{\downarrow}$	**0.561** ± 0.057
	E	0.187 ± 0.125	**0.248** ± 0.077	0.213 ± 0.144
	V	0.277 ± 0.106	0.290 ± 0.144	**0.309** ± 0.121
RECOLA-V	A	0.116 ± 0.124	$0.084 \pm 0.132^{\downarrow}$	**0.199** ± 0.082
	E	$0.151 \pm 0.119^{\downarrow}$	$0.120 \pm 0.101^{\downarrow}$	**0.260** ± 0.065
	V	0.413 ± 0.127	0.428 ± 0.115	**0.435** ± 0.130
SEWA-A	A	$0.145 \pm 0.106^{\downarrow}$	$0.172 \pm 0.108^{\downarrow}$	**0.258** ± 0.055
	T	$0.049 \pm 0.067^{\downarrow}$	0.090 ± 0.067	**0.123** ± 0.044
	V	0.499 ± 0.049	**0.514** ± 0.028	0.508 ± 0.031
SEWA-V	A	0.182 ± 0.047	0.192 ± 0.042	**0.194** ± 0.032
	T	$0.081 \pm 0.024^{\downarrow}$	$0.026 \pm 0.029^{\downarrow}$	**0.101** ± 0.037
	V	**0.525** ± 0.029	0.522 ± 0.030	**0.525** ± 0.027
UMEME-A	A	0.460 ± 0.115	0.490 ± 0.076	**0.504** ± 0.104
	T	$0.137 \pm 0.092^{\downarrow}$	0.152 ± 0.047	**0.189** ± 0.067
	V	0.450 ± 0.138	0.472 ± 0.072	**0.495** ± 0.096
UMEME-V	A	$0.105 \pm 0.056^{\downarrow}$	$0.109 \pm 0.079^{\downarrow}$	**0.155** ± 0.060
	T	$0.258 \pm 0.054^{\downarrow}$	$0.254 \pm 0.053^{\downarrow}$	**0.280** ± 0.050
	V	0.661 ± 0.051	0.670 ± 0.044	**0.677** ± 0.043
Accuracy (classification)				
TPOT	A	**0.373** ± 0.033	0.362 ± 0.013	0.360 ± 0.017
	T	0.331 ± 0.040	$0.318 \pm 0.024^{\downarrow}$	**0.334** ± 0.026
	V	0.514 ± 0.012	0.515 ± 0.010	**0.517** ± 0.013
VREED	E	0.282 ± 0.030	0.273 ± 0.054	**0.288** ± 0.018
	G	0.273 ± 0.015	**0.300** ± 0.059	0.297 ± 0.057
	V	$0.532 \pm 0.080^{\downarrow}$	0.578 ± 0.070	**0.604** ± 0.060

Table 7.4: Performance of the trimodal additive model when recovering the performance from only one modality. Higher is better in all cases and bold indicates best performance. $^{\downarrow}$ (and $^{\uparrow}$) indicates when a baseline performs significantly worse (or better) than SMURF at $\alpha = 0.05$.

	UMEME	IEMOCAP		
Samples	1544	100		
Modalities	A+T × V	A × T	A × V	T × V
	$\rho(\lvert\mathbf{y}^H_{AT} - \mathbf{y}^H_V\rvert, \lvert\hat{\mathbf{y}}_{(AT)V} + \hat{\mathbf{y}}_{V(AT)}\rvert)$	$\rho(\lvert\mathbf{y}^H_A - \mathbf{y}^H_T\rvert, \lvert\hat{\mathbf{y}}_{AT} + \hat{\mathbf{y}}_{TA}\rvert)$	$\rho(\lvert\mathbf{y}^H_A - \mathbf{y}^H_V\rvert, \lvert\hat{\mathbf{y}}_{AV} + \hat{\mathbf{y}}_{VA}\rvert)$	$\rho(\lvert\mathbf{y}^H_T - \mathbf{y}^H_V\rvert, \lvert\hat{\mathbf{y}}_{TV} + \hat{\mathbf{y}}_{VT}\rvert)$
Arousal	$0.089, p < 0.001$	$0.175, p = 0.002$	$0.126, p = 0.002$	ns
Valence	$0.087, p < 0.001$	$0.124, p = 0.031$	ns	$0.122, p = 0.033$

Table 7.5: Spearman's ρ between the magnitude of the pairwise redundant contributions from SMURF and the absolute difference of human unimodal judgments.

we could characterize redundancy overall at the dataset level using the correlation between the human unimodal judgments, for example, $r(\mathbf{y}^H_T, \mathbf{y}^H_V)$, or using a recently proposed partial information decomposition-based approach [110], we do not have enough statistical power (only two datasets with each two tasks) to evaluate whether those dataset level measures relate to SMURF's factorization. Instead, we focus on the sample-level with the expectation that we are more likely to observe larger pairwise redundant contributions between two modalities when the human unimodal judgments of these modalities are more similar [110]. We hypothesize that, for example, $|\hat{\mathbf{y}}_{TV} + \hat{\mathbf{y}}_{VT}|$ (the magnitude of the predicted pairwise redundant contributions) has a nonlinear correlation with $|\mathbf{y}^H_T - \mathbf{y}^H_V|$ (the difference between the human unimodal judgments). Table 7.5 tabulates the Spearman's ρ for these nonlinear correlation and shows that these hypotheses are confirmed for arousal and valence on UMEME and that we see similar tendencies on the much smaller subset of IEMOCAP.

Human modality judgments: TPOT has judgments of how informative modalities appear to humans when confirming its four affective states [208] ranging from *no*, *relevant*, and *sufficient* information. While these judgments do not separate between unique and redundant information, we expect more unique contributions for samples where the modality is judged as *relevant* or *sufficient* compared to different samples of the same modality that are judged as being uninformative. Wilcoxon's unpaired ranksums test confirms this for both text ($6.416, p < 0.001$) and

video ($2.639, p = 0.004$) but cannot confirm the hypothesis for audio ($0.725, p = 0.234$). This observation could be because the annotators found audio mostly uninformative on TPOT [208].

7.8 Conclusion

Affective states are often expressed in a redundant manner where multiple modalities convey a similar meaning. This makes it challenging to attribute what is uniquely predicted by a modality and what can redundantly be predicted by multiple modalities. We proposed SMURF to learn a model that factorizes its prediction into the sum of unique contributions and pairwise redundant contributions. Besides its factorization, SMURF often improved performance and, importantly, never significantly decreased performance. Further, SMURF became more robust to missing modalities as its maximization of the covariance between modalities encourages it to extract the same information from multiple modalities. Lastly, we observed that SMURF has the potential to improve interpretability as its factorization correlates with human judgments on three datasets.

Chapter 8

Conclusion and Future Directions

In this thesis, we focused on improving transparency for machine learning practitioners by delving into three dimensions of transparency: data transparency, which aims to provide more information about the data used to train models, including understanding how person-specific differences affect the model; reliability transparency, which aims to provide more information about how confident the model output is to enable better risk management, and model mechanics transparency, which aims to provide more information about the decision-making process of a model.

We first focused on analyzing data patterns consistent across people (population-level data transparency) in Chapter 2 by reviewing existing statistical approaches to analyze how acoustic features relate to symptom severities of psychosis. We further demonstrated in machine learning experiments that the acoustic features significantly predict the symptom severities.

Next, we provided prediction intervals in Chapter 3 to offer asymptotical guarantees for the margin of error in regression tasks to improve reliability transparency. We estimated the reliability of a primary predictive model using a secondary model and converted the predicted reliability to prediction intervals using the framework of inductive conformal prediction [138]. Our approach resulted in smaller prediction intervals than other approaches, enabling better risk management.

We revisited data transparency in Chapter 4, focusing on patterns that differ between people

while accounting for patterns consistent across people. We integrated mixed effect models with neural networks to learn complex nonlinear patterns that can be both different or consistent across people by learning person-generic and person-specific model parameters. We demonstrated that an affective model that has person-specific temporal transition patterns between affective states learns person-specific transition differences related to whether the person experienced symptoms of depression.

In the second half of the thesis, we focused on model mechanics of multimodal models as affect is often expressed through multiple modalities, such as audibly laughing while visually smiling. We initially explored in Chapter 5 how important modalities are for a model (modality importance transparency) to inform machine learning practitioners how much modalities influence the model output. For this, we first studied how informative humans perceive modalities to then guide models using those annotations. Our guided model focused on modalities similar to humans, and our experiments showed that the additional guidance increased predictive performance. Finally, we observed that machine learning practitioners might be able to manually improve the predictive performance in the future since our experiments showed that replacing the learned importance with the annotated human informativeness increased performance.

We then made multimodal models more transparent in Chapter 6 by quantifying how much a model uses unimodal additive, bimodal non-additive, and trimodal non-additive interactions (multimodal interaction transparency). We achieved this by prioritizing simpler interactions, such as prioritizing the unimodal additive over the bimodal non-additive interactions. This approach does not only inform machine learning practitioners about a particular model but also indicates to which degree more complex non-additive interactions are needed for a particular task on a given dataset.

Chapter 7 focused on the challenge that modalities often contain redundant information, such as visually smiling while audibly laughing, which makes it difficult to determine what a modality uniquely contributes (modality contribution transparency). We proposed a covariance-based

factorization to separate unique and pairwise redundant contributions between modalities. Our factorization not only improved robustness to missing modalities but also correlated with human judgments, indicating that our approach can make models more human interpretable.

8.1 Future Directions

8.1.1 Data: Multiple, Missing, and Unobserved Grouping Factors

In Chapter 4, we focused on scenarios where the grouping factor is known; in our case, all data from a person formed one group. In practice, there are other ways of grouping data, including gender, age, and the severity of task-relevant conditions, such as the severity of mood disorder symptoms. In the following paragraphs, we outline three research questions as possible extensions of the Neural Mixed Effects (NME) model to account for multiple, missing, or unobserved grouping factors.

The first research question: how do we analyze data patterns from multiple grouping factors simultaneously, such as for people and age? This is important when studying person-specific behaviors, as some behavioral differences might already be explained by another grouping factor, such as the person being younger. Multilevel models in statistics already provide a framework for this [17, 191]. Can we efficiently integrate this framework with neural networks to further tease apart complex patterns?

The second research question: what do we do when we do not observe the grouping factor at test time? The most prominent limitation of NME is that we need to know the grouping factor at test time, which required us to perform within-person testing. If we observe a new person, can we make a better guess than falling back to the ”average” person? Directly predicting a new person's person-specific parameters might be feasible [212]. Alternatively, learning a task-relevant similarity between a new person and previously known people might make it possible to express a new person as a combination of the known people.

The third research question: can we learn differences for an unobserved grouping factor? Assuming we have an observable fine-grained grouping factor, such as people, and we hypothesize that there is a more important coarse grouping factor, but we do not observe it. For example, the more coarse grouping factor might be based on the Big Five personality dimensions [34]. We could use an approach similar to Latent Dirichlet Allocation (LDA) [14], which can describe a book as a set of multiple unobserved topics based on the words in the book, by describing a person (a book) as a set of unobserved personality clusters (set of topics) based on the person's observations (the book's words). While this might make the training with known people more difficult, it might extend more naturally to unknown people at test time as we, even for known people, need first to infer their personality clusters. Such an approach might also be more scalable to more people, assuming there are fewer personality clusters than people.

8.1.2 Model Mechanics: Many Modalities

With models incorporating more and more modalities [108], the question arises whether our proposed methods can scale to a larger number of modalities. If we have dozens of modalities, it might be overwhelming to present even machine learning practitioners with all these values representing a modality's importance, their interactions, or unique and pairwise redundant contributions. All three chapters on multimodal model mechanics require dedicated model branches to focus on specific modalities and Chapter 6 requires even more for all possible interactions between modalities. How can we scale these approaches to a larger number of modalities? Sharing model parameters between branches might be a starting point, such as re-using unimodal representations learned by unimodal branches in the bimodal branches. In settings where only the inference time after training needs to scale, an approach could be first to learn many unimodal, bimodal, ..., and m-modal teacher branches. To then train one m-modal student model to mirror the output of the many teacher models.

8.1.3 Reliability: Out-of-Domain Data

Throughout this thesis, we assumed that we always work with in-domain data, i.e., that our testing data has the same patterns as our training data. Datasets such as UMEME [152] that purposefully combine audio with a mismatched video might lead to an incorrect estimation of reliability (Chapter 3) or an incorrect prediction of what is unique and redundant between modalities (Chapter 7). The problem of out-of-domain data will likely also occur when a model is deployed in real-world settings. Two potential workarounds might be a) estimating an in-domain probability to indicate when to rely on the provided transparency measures or b) providing ways to test the plausibility of model mechanics, for example, in the case of the pairwise redundant contributions, we can test whether they are highly correlated.

8.1.4 Towards Transparency for Everyone

With more and more people actively using machine learning models, such as large language models (LLMs) in the form of chatbots, it becomes important for people to understand a) what data these models were trained on, b) how reliable these models are, and c) have a better understanding of their model mechanics. This thesis focused on improving transparency for machine learning practitioners from the perspective of a machine learning practitioner. This allowed us to focus on technical concepts, such as model mechanics, that might be challenging to present to non-machine learning practitioners. A starting point for future work to improve transparency for everyone is conducting user studies [175, 194] and reviewing proposed policies regulating artificial intelligence [23, 40] to understand better what information users would like to know about machine learning models. In this thesis, we directly ingrained the transparency of model mechanics in the model architecture and how the model is trained. A potentially more flexible approach for transparency might be to let the user ask the model questions to describe how it derived the output. This is already a common pattern when interacting with LLMs [199, 226]. One interesting research direction when approaching this interactive question-asking approach

is ensuring that the model describes how it actually derived the output instead of describing how one could potentially derive the output [82, 116].

Bibliography

[1] Aylin Alin. Multicollinearity. *Wiley interdisciplinary reviews: computational statistics*, 2 (3):370–374, 2010. 7, 7.1

[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR, 2013. 7.1

[3] Randy P Auerbach, Apoorva Srinivasan, Jaclyn S Kirshenbaum, J John Mann, and Stewart A Shankman. Geolocation features differentiate healthy from remitted depressed adults. *Journal of psychopathology and clinical science*, 131(4):341, 2022. 4.4.1

[4] Randy P Auerbach, Ranqing Lan, Hanga Galfalvy, Kira lqueza, Jeffrey F Cohn, Ryan Crowley, Katherine Durham, Karla Joyce, Lauren E Kahn, Rahil Kamath, Louis-Philippe Morency, Giovanna Porta, Apoorva Srinivasan, Jamie Zelazny, David A Brent, and Nicholas B Allen. Intensive longitudinal assessment of adolescents to predict suicidal thoughts and behaviors. *Journal of the American Academy of Child and Adolescent Psychiatry*, 2023. 4.2, 4.4.1

[5] Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014. 1

[6] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 7.1

[7] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66. IEEE, 2018. 3.4.1, 4.4.1, 5.5.1, 6.4, 7.5.2

[8] Maria Barrett, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard. Sequence classification with human attention. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 302–312, 2018. 5.1

[9] John A Bateman. *Text and image: A critical introduction to the visual/verbal divide*. Routledge, 2014. 6.1

[10] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01. URL `https://www.jstatsoft.org/index.php/jss/article/view/v067i01`. 4.1

[11] Matthias Bauer, Mark van der Wilk, and Carl Edward Rasmussen. Understanding probabilistic sparse gaussian process approximations. In *Advances in neural information processing systems*, pages 1533–1541, 2016. 3.5

[12] Pranesh Bhargava and Deniz Başkent. Effects of low-pass filtering on intelligibility of periodically interrupted speech. *The Journal of the Acoustical Society of America*, 131(2): EL87–EL92, 2012. 6.5

[13] John Binder, Kevin Murphy, and Stuart Russell. Space-efficient inference in dynamic probabilistic networks. *Bclr*, 1:t1, 1997. 4.3.3

[14] David Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. *Advances in neural information processing systems*, 14, 2001. 8.1.1

[15] Henrik Boström, Henrik Linusson, Tuve Löfström, and Ulf Johansson. Accelerating difficulty estimation for conformal regression forests. *Annals of Mathematics and Artificial*

Intelligence, 81(1-2):125–144, 2017. 3

[16] Margaret M Bradley and Peter J Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994. 6.5

[17] Paul-Christian Bürkner. brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, 80:1–28, 2017. 8.1.1

[18] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4): 335–359, 2008. (document), 4.2, 4.4.1, 6.1, **??**, 6.5.1, 7.1, 7.5.1

[19] Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80, 2016. 5.1

[20] Alison L Calear and Helen Christensen. Systematic review of school-based prevention and early intervention programs for depression. *Journal of adolescence*, 33(3):429–438, 2010. 5

[21] Dallas Card, Michael Zhang, and Noah A Smith. Deep weighted averaging classifiers. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 369–378. ACM, 2019. 3.2.1

[22] Rich Caruana. Multitask learning. *Machine learning*, 28:41–75, 1997. 4.1

[23] Corinne Cath. Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133):20180080, 2018. 8.1.4

[24] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh

Ghassemi. Ethical machine learning in healthcare. *Annual review of biomedical data science*, 4:123–144, 2021. 1

[25] Lawrence S Chen, Thomas S Huang, Tsutomu Miyasato, and Ryohei Nakatsu. Multimodal human emotion/expression recognition. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, pages 366–371. IEEE, 1998. 1

[26] Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. Dialogue act recognition via crf-attentive structured network. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 225–234, 2018. 5.1

[27] Jeffrey F Cohn, Tomas Simon Kruez, Iain Matthews, Ying Yang, Minh Hoai Nguyen, Margara Tejera Padilla, Feng Zhou, and Fernando De la Torre. Detecting depression from facial actions and vocal prosody. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–7. IEEE, 2009. 1

[28] Pierre Colombo, Emile Chapuis, Matthieu Labeau, and Chloe Clavel. Improving multimodal fusion via mutual dependency maximisation. *arXiv preprint arXiv:2109.00922*, 2021. 7.1

[29] E Comets, A Lavenu, and M Lavielle. SAEMIX, an R version of the SAEM algorithm. *20^{th} meeting of the Population Approach Group in Europe, Athens, Greece*, 2011. 4.2b, 4, 4.1, 4.3.1, 4.3.2

[30] Michael A Covington, SL Anya Lunden, Sarah L Cristofaro, Claire Ramsay Wan, C Thomas Bailey, Beth Broussard, Robert Fogarty, Stephanie Johnson, Shayi Zhang, and Michael T Compton. Phonetic measures of reduced tongue movement correlate with negative symptom severity in hospitalized patients with first-episode schizophrenia-spectrum disorders. *Schizophrenia research*, 142(1):93–95, 2012. 2.1, 2.3.1, 2.4

[31] Joao FG de Freitas, Mahesan Niranjan, Andrew H. Gee, and Arnaud Doucet. Sequential monte carlo methods to train neural network models. *Neural computation*, 12(4):955–993,

2000. 4.3.2

[32] Beatrice De Gelder and Paul Bertelson. Multisensory integration, perception and ecological validity. *Trends in cognitive sciences*, 7(10):460–467, 2003. 6.1

[33] NivjaH de Jong and Ton Wempe. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2):385–390, 2009. 2.2.2

[34] Boele De Raad. *The big five personality factors: the psycholexical approach to personality.* Hogrefe & Huber Publishers, 2000. 8.1.1

[35] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. COVAREP - a collaborative voice analysis repository for speech technologies. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 960–964, 2014. 2.2.2

[36] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep—a collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 960–964. IEEE, 2014. 5.5.1

[37] Frank Dellaert, Thomas Polzin, and Alex Waibel. Recognizing emotion in speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1970–1973. IEEE, 1996. 1

[38] Bernard Delyon, Marc Lavielle, and Eric Moulines. Convergence of a stochastic approximation version of the em algorithm. *Annals of statistics*, pages 94–128, 1999. 4.3.2

[39] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068, 2014. 1

[40] Natalia Díaz-Rodríguez, Javier Del Ser, Mark Coeckelbergh, Marcos López de Prado, Enrique Herrera-Viedma, and Francisco Herrera. Connecting the dots in trustworthy artificial intelligence: From ai principles, ethics, and key requirements to responsible ai systems and regulation. *Information Fusion*, page 101896, 2023. 8.1.4

[41] Niels J Dingemanse and Ned A Dochtermann. Quantifying individual variation in behaviour: mixed-effect modelling approaches. *Journal of Animal Ecology*, 82(1):39–54, 2013. 1, 1.1.3

[42] T. Drugman and A. Abeer. Joint robust voicing detection and pitch estimation based on residual harmonics. In *Proceedings of Interspeech*, pages 1973–1976, 2011. 2.2.2

[43] Thomas Drugman, Yannis Stylianou, Yusuke Kida, and Masami Akamine. Voice activity detection: Merging source and filter-based information. *IEEE Signal Processing Letters*, 23(2):252–256, 2015. 2.2.1

[44] Mengnan Du, Ninghao Liu, and Xia Hu. Techniques for interpretable machine learning. *Communications of the ACM*, 63(1):68–77, 2019. 6

[45] Greg Durrett and Dan Klein. Neural CRF parsing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 302–312. Association for Computational Linguistics, July 2015. 4, 4.3.3

[46] Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208, 2018. 1

[47] Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978. 1.1.2

[48] Itir Onal Ertugrul, Laszlo A Jeni, Wanqiao Ding, and Jeffrey F Cohn. Afar: A deep learn-

ing based tool for automated facial affect recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pages 1–1. IEEE, 2019. 5.5.1

[49] Theodoros Evgeniou and Massimiliano Pontil. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 109–117, 2004. 4.1

[50] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462, 2010. 5.5.1

[51] Florian Eyben, Felix Weninger, Stefano Squartini, and Björn Schuller. Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 483–487. IEEE, 2013. 5.5.1

[52] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 2015. 5.5.1, 6.4, 7.5.2

[53] Florian Eyben, Klaus Scherer, Björn Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Devillers, Julien Epps, Petri Laukka, Shrikanth Narayanan, and Khiet Phuong Truong. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing*, 7(2):190–202, 4 2016. ISSN 1949-3045. doi: 10.1109/TAFFC.2015.2457417. Open access. 4.4.1

[54] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. Liblinear: A library for large linear classification. *the Journal of machine Learning research*,

9:1871–1874, 2008. 5.5

[55] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016. 3.3.1

[56] Ilaria Gaudiello, Elisabetta Zibetti, Sébastien Lefort, Mohamed Chetouani, and Serena Ivaldi. Trust as indicator of robot functional and social acceptance. an experimental study on user conformation to icub answers. *Computers in Human Behavior*, 61:633–655, 2016. 1

[57] Jeffrey M Girard, Jeffrey F Cohn, Mohammad H Mahoor, S Mohammad Mavadati, Zakia Hammal, and Dean P Rosenwald. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses. *Image and vision computing*, 32(10):641–647, 2014. 5.5.1

[58] Ella Glikson and Anita Williams Woolley. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2):627–660, 2020. 1, 1

[59] William Goldfarb, Nathan Goldfarb, Patricia Braunstein, and Hannah Scholl. Speech and language faults of schizophrenic children. *Journal of autism and childhood schizophrenia*, 2(3):219–233, 1972. 2, 2.1, 2.3.1

[60] Matthew C Gombolay, Reymundo A Gutierrez, Shanelle G Clarke, Giancarlo F Sturla, and Julie A Shah. Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams. *Autonomous Robots*, 39(3):293–312, 2015. 1

[61] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57, 2017. 6

[62] Michael Grimm and Kristian Kroschel. Evaluation of natural emotions using self assessment manikins. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, pages 381–385. IEEE, 2005. 6.5.1

[63] T. Hacki. Klassifizierung von glottisdysfunktionen mit hilfe der elektroglottographie. *Folia Phoniatrica*, 41(1):43–48, 1989. 2.3.1

[64] Kyu J Han, Panayiotis G Georgiou, and Shrikanth S Narayanan. The sail speaker diarization system for analysis of spontaneous meetings. In *Multimedia Signal Processing, 2008 IEEE 10th Workshop on*, pages 966–971. IEEE, 2008. 2.2.1

[65] Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical Science*, 1 (3):297–310, 1986. ISSN 08834237. 7, 7.5.3

[66] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1122–1131, 2020. 7.1

[67] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6.4

[68] Tom Heskes. Practical confidence and prediction intervals. In *Advances in neural information processing systems*, pages 176–182, 1997. 3.3.1

[69] Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020. 6, 6.1, 6.2, 6.7

[70] Verena Heusser, Niklas Freymuth, Stefan Constantin, and Alex Waibel. Bimodal speech emotion recognition using pre-trained language models. *arXiv preprint arXiv:1912.02610*, 2019. 6

[71] Hermann O Hirschfeld. A connection between correlation and contingency. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 31, pages 520–524. Cambridge University Press, 1935. 7.3.1

[72] Kevin Anthony Hoff and Masooda Bashir. Trust in automation: Integrating empirical

evidence on factors that influence trust. *Human factors*, 57(3):407–434, 2015. 1

[73] Hyman Hops, Betsy Davis, and Nancy Longoria. Methodological issues in direct observation: Illustrations with the living in familial environments (life) coding system. *Journal of Clinical Child Psychology*, 24(2):193–203, 1995. 4.4.1, 5.2

[74] Clayton Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225, 2014. 4.4.1

[75] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 5.4

[76] Nicholas C Jacobson and Yeon Joo Chung. Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones. *Sensors*, 20(12):3572, 2020. 4.4.1

[77] Spencer L James, Degu Abate, Kalkidan Hassen Abate, Solomon M Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159): 1789–1858, 2018. 5

[78] Ulf Johansson, Henrik Boström, Tuve Löfström, and Henrik Linusson. Regression conformal prediction with random forests. *Machine Learning*, 97(1-2):155–176, 2014. 3

[79] J. Kane and C. Gobl. Identifying regions of non-modal phonation using features of the wavelet transform. In *Proceedings of Interspeech*, pages 177–180, 2011. 2, 2.3.1

[80] Belhal Karimi, Marc Lavielle, and Eric Moulines. f-saem: A fast stochastic approximation of the em algorithm for nonlinear mixed effects models. *Computational Statistics & Data Analysis*, 141:123–138, 2020. 4.1, 4.3.2

[81] KS Kasiviswanathan and KP Sudheer. Methods used for quantifying the prediction uncertainty of artificial neural network based hydrologic models. *Stochastic environmental research and risk assessment*, 31(7):1659–1670, 2017. 3

[82] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023. 8.1.4

[83] Stanley R Kay, Abraham Flszbein, and Lewis A Opfer. The positive and negative syndrome scale (panss) for schizophrenia. *Schizophrenia bulletin*, 13(2):261, 1987. 2, 2.1

[84] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017. 3, 3.3.2

[85] Pascal Kilian, Sangbeak Ye, and Augustin Kelava. Mixed effects in machine learning – a flexible mixedML framework to add random effects to supervised machine learning regression. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL `https://openreview.net/forum?id=MKZyHtmfwH`. 4, 4.1

[86] Helen Killaspy, Sarah White, Nabeela Lalvani, Rachel Berg, Ajoy Thachil, Sen Kallumpuram, Omar Nasiruddin, Christine Wright, and Gill Mezey. The impact of psychosis on social inclusion and associated factors. *International Journal of Social Psychiatry*, 60(2): 148–154, 2014. 2

[87] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 5.5

[88] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In

Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations*, 2015. 4.4.3, 7.5.5

[89] Rolf Kloepfer. Komplementarität von sprache und bild am beispiel von comic, karikatur und reklame. *Sprache in Technischen Zeitalter Stuttgart*, 1976. 6.1, 6.6

[90] Charles Knapp and Glifford Carter. The generalized correlation method for estimation of time delay. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 24(4): 320–327, 1976. 2.2.1

[91] Michele Knox, Cheryl King, Gregory L Hanna, Deirdre Logan, and Neera Ghaziuddin. Aggressive behavior in clinically depressed adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry*, 39(5):611–618, 2000. 4.5.2

[92] Christian G Kohler, Jeffrey B Walker, Elizabeth A Martin, Kristin M Healey, and Paul J Moberg. Facial emotion perception in schizophrenia: a meta-analytic review. *Schizophrenia bulletin*, 36(5):1009–1019, 2009. 2.1

[93] Artemy Kolchinsky. A novel approach to the partial information decomposition. *Entropy*, 24(3):403, 2022. 7.1

[94] Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163, 2016. 6.5, 6.5.1

[95] Klaus Krippendorff. Computing krippendorff's alpha-reliability. 2011. 5.3.1

[96] Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. Integrating text and image: Determining multimodal document intent in instagram posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019. 6.1, 6.4, **??**, 6.6

[97] Viraj Kulkarni, Milind Kulkarni, and Aniruddha Pant. Survey of personalization tech-

niques for federated learning. In *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, pages 794–797. IEEE, 2020. 1

[98] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001. 4.3.3

[99] Nan M Laird and James H Ware. Random-effects models for longitudinal data. *Biometrics*, pages 963–974, 1982. 4, 4.1

[100] John David Michael Henry Laver. *Individual features in voice quality.* PhD thesis, University of Edinburgh, 1987. (document), 2.1, 2.3.1

[101] I Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989. 4.4.1

[102] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998. 3.4.1

[103] David I Leitman, John J Foxe, Pamela D Butler, Alice Saperstein, Nadine Revheim, and Daniel C Javitt. Sensory contributions to impaired prosodic processing in schizophrenia. *Biological psychiatry*, 58(1):56–61, 2005. 2.1

[104] Joshua J Levy, Carly A Bobak, Mustafa Nasir-Moin, Eren M Veziroglu, Scott M Palisoul, Rachael E Barney, Lucas A Salas, Brock C Christensen, Gregory J Tsongalis, and Louis J Vaickus. Mixed effects machine learning models for colon cancer metastasis prediction using spatially localized immuno-oncology markers. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2022*, pages 175–186. World Scientific, 2021. 4

[105] Robert A Lewis, Asma Ghandeharioun, Szymon Fedor, Paola Pedrelli, Rosalind Picard, and David Mischoulon. Mixed effects random forests for personalised predictions of clinical depression severity. *arXiv preprint arXiv:2301.09815*, 2023. 4

[106] Nicole YK Li and Edwin M-L Yiu. Acoustic and perceptual analysis of modal and falsetto registers in females with dysphonia. *Clinical linguistics & phonetics*, 20(6):463–481,

2006. 6.5

[107] Paul Pu Liang, Terrance Liu, Anna Cai, Michal Muszynski, Ryo Ishii, Nick Allen, Randy Auerbach, David Brent, Ruslan Salakhutdinov, and Louis-Philippe Morency. Learning language and multimodal privacy-preserving markers of mood from mobile data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4170–4187, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.322. URL `https://aclanthology.org/2021.acl-long.322`. 4.4.1

[108] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Jeffrey Tsaw, Yudong Liu, Shentong Mo, Dani Yogatama, Louis-Philippe Morency, and Russ Salakhutdinov. High-modality multimodal transformer: Quantifying modality & interaction heterogeneity for high-modality representation learning. *Transactions on Machine Learning Research*, 2022. 8.1.2

[109] Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Faisal Mahmood, Ruslan Salakhutdinov, and Louis-Philippe Morency. Quantifying & modeling feature interactions: An information decomposition framework. *arXiv preprint arXiv:2302.12247*, 2023. 7.1

[110] Paul Pu Liang, Yun Cheng, Ruslan Salakhutdinov, and Louis-Philippe Morency. Multimodal fusion interactions: A study of human and automatic quantification. *arXiv preprint arXiv:2306.04125*, 2023. 7.7

[111] Alvin M Liberman, Katherine Safford Harris, Howard S Hoffman, and Belver C Griffith. The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental psychology*, 54(5):358, 1957. 5.3.1

[112] Robert LiKamWa, Yunxin Liu, Nicholas D Lane, and Lin Zhong. Moodscope: Building a mood sensor from smartphone usage patterns. In *Proceeding of the 11th annual inter-*

national conference on Mobile systems, applications, and services, pages 389–402, 2013. 1

[113] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1):18, 2020. 1, 1.1.1

[114] Mary J Lindstrom and Douglas M Bates. Newton—raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022, 1988. 4.2a, 4.1, 4.4.2

[115] Mary J Lindstrom and Douglas M Bates. Nonlinear mixed effects models for repeated measures data. *Biometrics*, pages 673–687, 1990. 4.1

[116] Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, et al. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, page 100017, 2023. 8.1.4

[117] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach, 2020. `https://openreview.net/forum?id=SyxS0T4tvS`. 4.4.1, 6.4

[118] Sydney Lolli, Ari D Lewenstein, Julian Basurto, Sean Winnik, and Psyche Loui. Sound frequency affects speech emotion perception: Results from congenital amusia. *Frontiers in Psychology*, 6:1340, 2015. 5.1

[119] Ventura J. Lukoff D., Nuechterlein KH. Manual for expanded brief psychiatric rating scale. *Schizophrenia Bulletin*, 12:594–602, 1986. 2, 2.2

[120] Yiwei Lyu, Paul Pu Liang, Zihao Deng, Ruslan Salakhutdinov, and Louis-Philippe Morency. Dime: Fine-grained interpretations of multimodal models via disentangled local explanations. *arXiv preprint arXiv:2203.02013*, 2022. 6.1, 6.2

[121] Fei Ma, Shao-Lun Huang, and Lin Zhang. An efficient approach for audio-visual emotion recognition with missing labels and missing modalities. In *2021 IEEE international conference on multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 7.1, 7.3.1, 7.5.3, **??**, **??**

[122] Francesca Mandel, Riddhi Pratim Ghosh, and Ian Barnett. Neural networks for clustered and longitudinal data using mixed effects models. *Biometrics*, 2021. 4, 4.1, 4.4.2

[123] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013. 3.4.1

[124] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, volume 2017, pages 498–502, 2017. 5.5.1

[125] Albert Mehrabian. *Silent messages*. Wadsworth Publishing Company, Belmont, California, 1971. 5

[126] Thomas Mock. Tidy tuesday: A weekly data project aimed at the r ecosystem, 2022. URL `https://github.com/rfordatascience/tidytuesday`. 4.2, 4.4.1

[127] Gelareh Mohammadi, Sunghyun Park, Kenji Sagae, Alessandro Vinciarelli, and Louis-Philippe Morency. Who is persuasive? the role of perceived personality and communication modality in social multimedia. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 19–26, 2013. 5.1

[128] Nuno Moniz and Luís Torgo. Multi-source social feedback of online news feeds. *arXiv preprint arXiv:1801.07055*, 2018. 4.2, 4.4.1

[129] Vera A Morgan, Anna Waterreus, Vaughan Carr, David Castle, Martin Cohen, Carol Harvey, Cherrie Galletly, Andrew Mackinnon, Patrick McGorry, John J McGrath, et al. Responding to challenges for people with psychotic illness: Updated evidence from the sur-

vey of high impact psychosis. *Australian & New Zealand Journal of Psychiatry*, 51(2): 124–140, 2017. 2

[130] Myriam Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing*, 5(2):101–111, 2014. 1, 6.5

[131] Benjamin W Nelson, Lisa Sheeber, Jennifer Pfeifer, and Nicholas B Allen. Psychobiological markers of allostatic load in depressed and nondepressed mothers and their adolescent offspring. *Journal of Child Psychology and Psychiatry*, 62(2):199–211, 2021. 4.4.1, 5, 5.2, **??**, 7.1, 7.5.1

[132] Che Ngufor, Holly Van Houten, Brian S Caffo, Nilay D Shah, and Rozalina G McCoy. Mixed effect machine learning: A framework for predicting longitudinal change in hemoglobin a1c. *Journal of biomedical informatics*, 89:56–67, 2019. 4

[133] Ian R Nicholson, Jonathan E Chapman, and Richard WJ Neufeld. Variability in bprs definitions of positive and negative symptoms. *Schizophrenia research*, 17(2):177–185, 1995. 2, 2.2

[134] National Institute of Standards and Technology. Spring 2006 (rt-06s) rich transcription meeting recognition evaluation plan, 2006. http://www.itl.nist.gov/iad/mig/tests/rt/2006-spring/docs/rt06s-meeting-eval-plan-V2.pdf. 2.2.1

[135] John Edison Arevalo Ovalle, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. In *ICLR (Workshop)*, 2017. 5.1, 5.4

[136] John E Overall and Donald R Gorham. The brief psychiatric rating scale. *Psychological reports*, 10(3):799–812, 1962. 1a, 2.2

[137] Joel S Owen and Jill Fiedler-Kelly. *Introduction to population pharmacokinetic/pharmacodynamic analysis with nonlinear mixed effects models*. John Wiley &

Sons, 2014. 4.1

[138] Harris Papadopoulos. Inductive conformal prediction: Theory and application to neural networks. In *Tools in artificial intelligence*. IntechOpen, 2008. 8

[139] Harris Papadopoulos and Haris Haralambous. Neural networks regression inductive conformal predictor and its application to total electron content prediction. In *International Conference on Artificial Neural Networks*, pages 32–41. Springer, 2010. 3

[140] Harris Papadopoulos and Haris Haralambous. Reliable prediction intervals with regression neural networks. *Neural Networks*, 24(8):842–851, 2011. 3

[141] Harris Papadopoulos, Alex Gammerman, and Volodya Vovk. Normalized nonconformity measures for regression conformal prediction. In *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications (AIA 2008)*, pages 64–69, 2008. 3, 3.4.3

[142] Harris Papadopoulos, Vladimir Vovk, and Alexander Gammerman. Regression conformal prediction with nearest neighbours. *Journal of Artificial Intelligence Research*, 40:815–840, 2011. 3.4.3

[143] Theodore Papamarkou, Jacob Hinkle, M Todd Young, and David Womble. Challenges in markov chain monte carlo for bayesian neural networks. *Statistical Science*, 37(3): 425–442, 2022. 4.3.2

[144] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035.

Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-librar pdf. 4.4.3, 7.5.5

[145] Sophie Pautex, François Herrmann, Paulette Le Lous, Malika Fabjan, Jean-Pierre Michel, and Gabriel Gold. Feasibility and reliability of four pain self-assessment scales and correlation with an observational rating scale in hospitalized elderly demented patients. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, 60(4):524–529, 2005. 1.1.3

[146] Paola Pedrelli, Szymon Fedor, Asma Ghandeharioun, Esther Howe, Dawn F Ionescu, Darian Bhathena, Lauren B Fisher, Cristina Cusin, Maren Nyer, Albert Yeung, et al. Monitoring changes in depression severity using wearable and mobile sensors. *Frontiers in psychiatry*, 11:584711, 2020. 4

[147] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. The development and psychometric properties of liwc2015. Technical report, University of Texas at Austin, 2015. 4.4.1

[148] Niklas Pfister, Peter Bühlmann, Bernhard Schölkopf, and Jonas Peters. Kernel-based tests for joint independence. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 80(1):5–31, 2018. 4.5.2

[149] José C Pinheiro and Douglas M Bates. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of computational and Graphical Statistics*, 4 (1):12–35, 1995. 4.1

[150] José C Pinheiro and Douglas M Bates. Linear mixed-effects models: basic concepts and examples. *Mixed-effects models in S and S-Plus*, pages 3–56, 2000. 4.5.2

[151] Abhishek Pratap, David C Atkins, Brenna N Renn, Michael J Tanana, Sean D Mooney, Joaquin A Anguera, and Patricia A Areán. The accuracy of passive phone sensors in

predicting daily mood. *Depression and anxiety*, 36(1):72–81, 2019. 4, 4.4.1

[152] Emily Mower Provost, Yuan Shangguan, and Carlos Busso. Umeme: University of michigan emotional mcgurk effect data set. *IEEE Transactions on Affective Computing*, 6(4): 395–409, 2015. 1, 5.1, 6.1, 6.5, 6.6, 7.1, 7.5.1, 8.1.3

[153] Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. Learning to deceive with attention-based explanations. In *The 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019. 5.1, 5.4

[154] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Advances in neural information processing systems*, pages 14866–14876, 2019. 6.3.2, 7.4

[155] Shiquan Ren, Hong Lai, Wenjing Tong, Mostafa Aminzadeh, Xuezhang Hou, and Shenghan Lai. Nonparametric bootstrapping for hierarchical data. *Journal of Applied Statistics*, 37(9):1487–1498, 2010. 4.4.3, 5.5

[156] Fabien Ringeval, Björn Schuller, Michel Valstar, Roddy Cowie, and Maja Pantic. Summary for avec 2018: Bipolar disorder and cross-cultural affect recognition. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 2111–2112, 2018. 7.1, 7.5.1

[157] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, pages 3–12, 2019. 5.1

[158] Robert Rosenthal. Conducting judgment studies: Some methodological issues. *The new handbook of methods in nonverbal behavior research*, pages 199–234, 2005. 6.5, 6.5.1, 6.5.1

[159] Quentin Roy, Futian Zhang, and Daniel Vogel. Automation accuracy is good, but high controllability may be better. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2019. 5.7

[160] Ognjen Rudovic, Jaeryoung Lee, Miles Dai, Björn Schuller, and Rosalind W Picard. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics*, 3(19):eaao6760, 2018. 1, 1.1.3

[161] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, 2017. 6.1

[162] Klaus R Scherer. What are emotions? and how can they be measured? *Social science information*, 44(4):695–729, 2005. 1

[163] Stefan Scherer, Giota Stratou, Jonathan Gratch, and Louis-Philippe Morency. Investigating voice quality as a speaker-independent indicator of depression and ptsd. In *Proceedings of Interspeech*, pages 847–851, 2013. 2.1, 2.2.2, 2.3.1, 2.4

[164] Stefan Scherer, Gale Lucas, Jonathan Gratch, Albert Rizzo, and Louis-Philippe Morency. Self-reported symptoms of depression and ptsd are associated with reduced vowel space in screening interviews. *IEEE Transactions on Affective Computing*, 7(1):59–73, 2015. 2, 2.1, 2.2.2, 2.3.1, 2.4

[165] Björn Schuller, Stefan Steidl, Anton Batliner, Florian Schiel, and Jarek Krajewski. The interspeech 2011 speaker state challenge. 2011. 5.1

[166] Orli S Schwartz, Paul Dudgeon, Lisa B Sheeber, Marie BH Yap, Julian G Simmons, and Nicholas B Allen. Observed maternal responses to adolescent behaviour predict the onset of major depression. *Behaviour research and therapy*, 49(5):331–338, 2011. 4.5.2

[167] Orli S Schwartz, Michelle L Byrne, Julian G Simmons, Sarah Whittle, Paul Dudgeon, Marie BH Yap, Lisa B Sheeber, and Nicholas B Allen. Parenting during early adolescence

and adolescent-onset major depression: A 6-year prospective longitudinal study. *Clinical Psychological Science*, 2(3):272–286, 2014. 1.2.1, 4.4.1, 4.5.2, 5, 5.1, 5.2

[168] K Shailaja, B Seetharamulu, and MA Jabbar. Machine learning in healthcare: A review. In *2018 Second international conference on electronics, communication and aerospace technology (ICECA)*, pages 910–914. IEEE, 2018. 1

[169] Jun Shi, Chengming Jiang, Aman Gupta, Mingzhou Zhou, Yunbo Ouyang, Qiang Charles Xiao, Qingquan Song, Yi (Alice) Wu, Haichao Wei, and Huiji Gao. Generalized deep mixed models. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, page 3869–3877, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393850. doi: 10.1145/3534678.3539103. URL https://doi.org/10.1145/3534678.3539103. 4, 4.1

[170] Giora Simchoni and Saharon Rosset. Integrating random effects in deep neural networks. *Journal of Machine Learning Research*, 24(156):1–57, 2023. URL http://jmlr.org/papers/v24/22-0501.html. 4, 4.1, 4.4.1

[171] Edward H Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 13(2):238–241, 1951. 4.1, 4.4.3

[172] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012. 2.2.3

[173] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013. 5.1

[174] Siyang Song, Zilong Shao, Shashank Jaiswal, Linlin Shen, Michel Valstar, and Hatice Gunes. Learning person-specific cognition from facial reactions for automatic personality

recognition. *IEEE Transactions on Affective Computing*, 2022. 4.1

[175] Harini Suresh, Steven R Gomez, Kevin K Nam, and Arvind Satyanarayan. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2021. 8.1.4

[176] Luma Tabbaa, Ryan Searle, Saber Mirzaee Bafti, Md Moinul Hossain, Jittrapol Intarasisrisawat, Maxine Glancy, and Chee Siang Ang. Vreed: Virtual reality emotion recognition dataset using eye tracking & physiological measures. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, 5(4):1–20, 2021. 7.1, 7.5.1

[177] Reeti Tandon, Sudeshna Adak, and Jeffrey A Kaye. Neural networks for longitudinal studies in alzheimer's disease. *Artificial intelligence in medicine*, 36(3):245–255, 2006. 4, 4.1, 4.4.2

[178] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29 (1):24–54, 2010. 5.5.1

[179] Sara Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. Personalized multitask learning for predicting tomorrow's mood, stress, and health. *IEEE Transactions on Affective Computing*, 11(2):200–213, 2017. 4.1

[180] Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial Intelligence and Statistics*, pages 567–574, 2009. 3.3.1

[181] F Tolkmitt, Hede Helfrich, Rainer Standke, and Klaus Rainer Scherer. Vocal indicators of psychiatric treatment effects in depressives and schizophrenics. *Journal of communication disorders*, 15(3):209–222, 1982. (document), 2, 2.1, 2.2.2, 2.1, 2.3.1

[182] Takumi Toyama, Thomas Kieninger, Faisal Shafait, and Andreas Dengel. Gaze guided object recognition using a head-mounted eye tracker. In *Proceedings of the Symposium*

on Eye Tracking Research and Applications, pages 91–98, 2012. 5.1

[183] Minh-Ngoc Tran, Nghia Nguyen, David Nott, and Robert Kohn. Random effects models with deep neural network basis functions: Methodology and computation. Technical report, University of Sydney Business School, 2017. 4, 4.1, 4.4.2

[184] Talia Tron, Abraham Peled, Alexander Grinsphoon, and Daphna Weinshall. Automated facial expressions analysis in schizophrenia: A continuous dynamic approach. In *International Symposium on Pervasive Computing Paradigms for Mental Health*, pages 72–81. Springer, 2015. 2

[185] Paula T Trzepacz and Robert W Baker. *The psychiatric mental status examination*. Oxford University Press, 1993. 2

[186] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2019. 5.1, 6

[187] Yao-Hung Hubert Tsai, Martin Q Ma, Muqiao Yang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Multimodal routing: Improving local and global interpretability of multimodal language analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, 2020. 5.1, 6.1, 6.3.1, 6.4

[188] Michael Tsang, Dehua Cheng, Hanpeng Liu, Xue Feng, Eric Zhou, and Yan Liu. Feature interaction interpretability: A case for explaining ad-recommendation systems via neural interaction detection. *arXiv preprint arXiv:2006.10966*, 2020. 6.1

[189] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. Avec 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceed-*

ings of the 6th international workshop on audio/visual emotion challenge, pages 3–10, 2016. **??**, 6.5.1, 7.1, 7.5.1

[190] Michel F Valstar, Enrique Sánchez-Lozano, Jeffrey F Cohn, László A Jeni, Jeffrey M Girard, Zheng Zhang, Lijun Yin, and Maja Pantic. Fera 2017-addressing head pose in the third facial expression recognition and analysis challenge. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 839–847. IEEE, 2017. 5.1

[191] Wim Van den Noortgate, Marie-Christine Opdenakker, and Patrick Onghena. The effects of ignoring a level in multilevel analysis. *School Effectiveness and School Improvement*, 16(3):281–303, 2005. 8.1.1

[192] Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999. 6.6

[193] Supriya Vijay, Tadas Baltrušaitis, Luciana Pennant, Dost Ongür, Justin T Baker, and Louis-Philippe Morency. Computational study of psychosis symptoms and facial expressions. In *Computing and Mental Health, Workshop at SIGCHI*, 2016. 2, 2.2

[194] Eric S Vorm and Andrew D Miller. Assessing the value of transparency in recommender systems: an end-user perspective. 2018. 8.1.4

[195] Tinghua Wang, Xiaolu Dai, and Yuze Liu. Learning with hilbert–schmidt independence criterion: A review and new perspectives. *Knowledge-based systems*, 234:107567, 2021. 7.1

[196] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020. 5, 6.6

[197] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Advances in Neural Information Processing Systems, 33:5776–5788, 2020. 7.5.2

[198] Xingbo Wang, Jianben He, Zhihua Jin, Muqiao Yang, Yong Wang, and Huamin Qu. M2lens: Visualizing and explaining multimodal models for sentiment analysis. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):802–812, 2021. 6.1

[199] Zihao Wang, Shaofei Cai, Anji Liu, Xiaojian Ma, and Yitao Liang. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *arXiv preprint arXiv:2302.01560*, 2023. 8.1.4

[200] James R Williamson, Elizabeth Godoy, Miriam Cha, Adrianne Schwarzentruber, Pooya Khorrami, Youngjune Gwon, Hsiang-Tsung Kung, Charlie Dagli, and Thomas F Quatieri. Detecting depression using vocal, facial and semantic communication cues. In *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, pages 11–18, 2016. 1

[201] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. Elan: a professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1556–1559, 2006. 5.3

[202] Russ Wolfinger. Covariance structure selection in general mixed models. *Communications in statistics-Simulation and computation*, 22(4):1079–1106, 1993. 4.3.1

[203] Wrandrall. Imdb new dataset, 2021. URL `https://www.kaggle.com/datasets/wrandrall/imdb-new-dataset`. 4.2, 4.4.1

[204] Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In *"Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 6153–6166, 2021. 6, 6.1

[205] Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. *arXiv preprint arXiv:2105.14462*, 2021. 7.1

[206] Torsten Wörtwein and Louis-Philippe Morency. Simple and effective approaches for uncertainty prediction in facial action unit intensity regression. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pages 452–456. IEEE, 2020. 2, 3

[207] Torsten Wörtwein, Tadas Baltrusaitis, Eugene Laksana, Luciana Pennant, Elizabeth S Liebson, Dost Öngür, Justin T Baker, and Louis-Philippe Morency. Computational analysis of acoustic descriptors in psychotic patients. In *INTERSPEECH*, pages 3256–3260, 2017. 1, 2

[208] Torsten Wörtwein, Lisa B Sheeber, Nicholas Allen, Jeffrey F Cohn, and Louis-Philippe Morency. Human-guided modality informativeness for affective states. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pages 728–734, 2021. 4, 4.2, 4.4.1, 5, 6.1, 6.6, 7.7

[209] Torsten Wörtwein, Lisa Sheeber, Nicholas Allen, Jeffrey Cohn, and Louis-Philippe Morency. Beyond additive fusion: Learning non-additive multimodal interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4681–4696, 2022. 5, 4.4.1, 6, 7.4, 7.5.1, 7.7

[210] Torsten Wörtwein, Nicholas B Allen, Lisa B Sheeber, Randy P Auerbach, Jeffrey F Cohn, and Louis-Philippe Morency. Neural mixed effects for nonlinear personalized predictions. In *Proceedings of the 2023 International Conference on Multimodal Interaction*, pages 445–454, 2023. 3, 4

[211] Fen Xiao, Liangchan Peng, Lei Fu, and Xieping Gao. Salient object detection based on eye tracking data. *Signal Processing*, 144:392–397, 2018. 5.1

[212] Yunyang Xiong, Hyunwoo J Kim, and Vikas Singh. Mixed effects neural networks (menets) with applications to gaze estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7743–7752, 2019. 4.2c, 4, 4.1, 4.3, 4.4.2, 4.5.1, 4.5.2, 8.1.1

[213] Zhaoyi Xu and Joseph Homer Saleh. Machine learning for reliability engineering and safety applications: Review of current status and future opportunities. *Reliability Engineering & System Safety*, 211:107530, 2021. 1

[214] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1642–1651, 2022. 7.1

[215] Ying Yang, Catherine Fairbairn, and Jeffrey F Cohn. Detecting depression severity from vocal prosody. *IEEE transactions on affective computing*, 4(2):142–150, 2012. 6.5

[216] An Gie Yong, Sean Pearce, et al. A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in quantitative methods for psychology*, 9(2):79–94, 2013. 7.2

[217] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2016. 6.3.1, **??**

[218] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016. 7.1, 7.5.1

[219] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017. 6, 6.4, 7.5.5

[220] Amir Zadeh, Chengfeng Mao, Kelly Shi, Yiwei Zhang, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. Factorized multimodal transformer for multimodal sequential learning. In *Elsevier Information Fusion Journal (IF 11.21)*, 2019. 6.3.1, 7, 7.1

[221] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pages 2236–2246, 2018. 5.1, 6.1, **??**, 7.1, 7.5.1

[222] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. In *Advances in Neural Information Processing Systems 34*, 2021. 6

[223] Ying Zeng, Sijie Mai, and Haifeng Hu. Which is making the contribution: Modulating unimodal and cross-modal dynamics for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1262–1274. Association for Computational Linguistics, 2021. 6.6

[224] Yufei Zeng, Zhixin Li, Zhenjun Tang, Zhenbin Chen, and Huifang Ma. Heterogeneous graph convolution based on in-domain self-supervision for multimodal sentiment analysis. *Expert Systems with Applications*, 213:119240, 2023. 7, 7.1

[225] Mingda Zhang, Rebecca Hwa, and Adriana Kovashka. Equal but not the same: Understanding the implicit relationship between persuasive images and text. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 6.1

[226] Xijia Zhang, Yue Guo, Simon Stepputtis, Katia Sycara, and Joseph Campbell. Explaining agent behavior with large language models. *arXiv preprint arXiv:2309.10346*, 2023. 8.1.4

[227] Yuhao Zhang, Ying Zhang, Wenya Guo, Xiangrui Cai, and Xiaojie Yuan. Learning disentangled representation for multimodal cross-domain sentiment analysis. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 7.1

[228] Zheng Zhang, Jeff M Girard, Yue Wu, Xing Zhang, Peng Liu, Umur Ciftci, Shaun Canavan, Michael Reale, Andy Horowitz, Huiyuan Yang, et al. Multimodal spontaneous emotion corpus for human behavior analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3438–3446, 2016. 3.4.1

[229] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. 5.6

[230] Jiahao Zheng, Sen Zhang, Xiaoping Wang, and Zhigang Zeng. Multimodal representations learning based on mutual information maximization and minimization and identity embedding for multimodal sentiment analysis. *arXiv preprint arXiv:2201.03969*, 2022. 7.1