Master Thesis

July 21, 2022

Emotion Recognition in Couples using Transfer Learning

A Multimodal Approach

Madhav Sachdeva

of Delhi, India (19-764-869)

supervised by

Prof. Dr. Thomas Fritz
George Boateng

University of Zurich UZH

HASEL

Master Thesis

Emotion Recognition in Couples using Transfer Learning

A Multimodal Approach

Madhav Sachdeva

University of Zurich^UZH

HASEL

Master Thesis

Author: Madhav Sachdeva, madhav.sachdeva@uzh.ch

Project period: 21 January 2022 - 21 July 2022

Human Aspects of Software Engineering Lab
Department of Informatics, University of Zurich

Acknowledgements

First and foremost, I would like to extend my sincere thanks to George Boateng for his guidance, support, and dedication to this project. Also, thanks for the several discussions, assistance, and feedback meetings that helped me to think more in-depth over the course of the project.

I would also like to extend my sincere thanks to Professor Thomas Fritz for the opportunity to work on such a interesting topic. I very much appreciate the feedback and suggestions, which gave me new perspectives.

Finally, I would like to thank my family and friends for their strong belief and the constant motivation they instilled in me.

Abstract

Automatic emotion recognition in couples may be useful for understanding mental health risks or evaluating outcomes for chronic disease management. However, emotion recognition in couples is under-researched due to difficulties in obtaining data and overcoming issues such as limited samples, noise, and imbalance. In this thesis, a novel in-the-wild dataset called DyMand is investigated for emotion recognition in couples. Furthermore, transfer learning models are developed using the public datasets VAM and K-EmoCon, and the best neural network layers for fine-tuning are demonstrated. In addition, multimodal fusion approaches (early fusion and late fusion) are investigated to utilize different modalities of physiological, acoustic, and linguistic data. Additionally, multi-modal fusion is compared across these modalities and this thesis demonstrates which modalities can improve couples emotion recognition. Furthermore, the developed transfer learning models could improve performance across all modalities by up to 12%.

Zusammenfassung

Die automatische Erkennung von Emotionen bei Paaren kann nützlich sein, um Risiken für die psychische Gesundheit zu verstehen oder die Ergebnisse der Behandlung chronischer Krankheiten zu bewerten. Die Erkennung von Emotionen bei Paaren ist jedoch noch wenig erforscht, da es schwierig ist, Daten zu erhalten und Probleme wie begrenzte Stichproben, Rauschen und Unausgewogenheit zu überwinden. In dieser Arbeit wird ein neuartiger "in-the-wild"-Datensatz namens DyMand für die Emotionserkennung bei Paaren untersucht. Darüber hinaus werden Transfer-Learning-Modelle unter Verwendung der öffentlichen Datensätze VAM und K-EmoCon entwickelt und die besten neuronalen Netzwerkschichten für die Feinabstimmung aufgezeigt. Darüber hinaus werden multimodale Fusionsansätze (frühe Fusion und späte Fusion) untersucht, um verschiedene Modalitäten physiologischer, akustischer und linguistischer Daten zu nutzen. Darüber hinaus wird die multimodale Fusion zwischen diesen Modalitäten verglichen und es wird gezeigt, welche Modalitäten die Emotionserkennung von Paaren verbessern können. Darüber hinaus konnten die entwickelten Transfer-Learning-Modelle die Leistung über alle Modalitäten hinweg um bis zu 12% verbessern.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Understanding the emotions of couples can provide insights about their mental health, help to identify risk factors in relationships [12, 28], or facilitate the prediction of health outcomes in chronic illness management [13, 16]. The emotions of partners in a romantic relationship are known to reflect the emotional support they receive from each other. For instance, studies have shown that positive emotional support from spouses leads to healthier habits in diabetes patients [3,43,56]. Towards this direction, researchers are interested in recognizing emotions of couples to understand the link between romantic relationships and health. [42].

Assessing couples' emotion recognition is a challenging task. Often, couples are asked to self-report their emotions using questionnaires such as PANAS [71]. They need to keep track of their emotions periodically to help researchers in assessing their emotions, which takes time and effort. Another approach is that couples record their conversations, which is later analyzed and the emotions are assessed [5]. However, automatic emotion recognition can provide a means of more effectively recognizing emotions from couples that has less time, effort and more practicality [5].

However, there are several challenges in automatic emotion recognition for couples in everyday life. First, research in automatically detecting emotions of couples has been mostly limited to works that use data collected in ideal laboratory environments [35, 44], which do not reflect the real world. Second, several work also use data from actors who communicate with complete or partially scripted conversations to elicit emotions [23,35,44]. While limited work has focused on real life couples for eliciting emotions [6]. Real life data on emotion recognition is often limited in samples, noisy, and imbalanced as compared to data collected from controlled experiments in lab environments. There are also not many public datasets that focus particularly on couples emotion recognition. Those that do, lack in many ways such as they do not employ real couples, or they give them scripts to read for inducing particular emotions. Therefore, automatic emotion recognition in couples from the real life is still underexplored in this field of research due to a lack of data.

The DyMand dataset [8] was recently created to study the link between emotions and the dyadic management of diabetes. The dataset employed 13 heterosexual couples, in which one partner had type-2 diabetes [8]. The study collected physiological data (heart rate, accelerometer, and more) and speech from patients and their partners over the course of a week. The data was collected using smartwatches and the couples were provided smartphones with a pre-installed application to self-report their emotion data using a standardized methodology. This dataset could help researchers in the direction of building an emotion recognition system that could automatically recognize the emotions of patients and consequently their social support from spouses, using commodity smartwatches. There have been no prior attempts to evaluate the DyMand dataset for emotion recognition due to the dataset being recently introduced. The DyMand dataset also has limited samples, noisy and imbalanced data. These factors impede efforts in building automatic emotion recognition models. In efforts to build automatic emotion recognition models, several approaches must be explored that will help the machine learning models to address the issues of limited samples, noise and imbalanced data.

In emotion recognition tasks, multimodal fusion has shown to improve the classification accuracy in real life datasets [18]. Multimodal fusion leverages two or more modalities simultaneously to recognize emotions. By using several modalities, the models are able to combine information from several sources to enhance pattern recognition. However, there are several challenges with multimodal fusion. For instance, determining which modalities to combine [18], or which multimodal fusion approach to use. There are two main approaches in multimodal fusion. One approach fuses modalities at feature level, called early fusion. This approach enables the machine learning model to use features from the modalities to leverage different sources together. Another approach fuses modalities at the decision level, called late fusion. This approach trains separate machine learning models for each modality, and their respective predictions for the emotions are combined using approaches such as averaging, or majority voting. Couples emotion recognition, in particular, has not investigated multimodal fusion approaches of early fusion and late fusion for datasets on real life couples.

For addressing the limitations of less training data in emotion recognition tasks, transfer learning is often proposed. Transfer learning is the process of using a pre-trained model on a different, large dataset and applying it to improve the prediction of the target data on another dataset [49]. It relies on the assumption that a model trained on a large dataset can generalize well to another dataset in the same domain. For this reason, transfer learning is applied when there is limited training data for improving accuracy. In several domains, transfer learning has achieved success in improving prediction performance even with less training samples [47,75]. It has also demonstrated success in the domain of emotion recognition [24,34,60].

Therefore, this thesis aims to leverage transfer learning and multi-modal fusion for improving real life couples emotion recognition for the DyMand dataset. Specifically, the following research questions will be focused on:

Research Questions

- **RQ 1:** How does the performance with early multi-modal fusion compare with late multi-modal fusion?

- **RQ 2:** How does transfer learning by fine-tuning affect the performance of the DyMand dataset in the early multi-modal fusion approach as compared with the late multi-modal fusion approaches?

Contribution of work

This thesis contributes to the limited research on couples emotion recognition. Firstly, multi-modal fusion approaches will be examined on the DyMand dataset to investigate the effect of different modalities (speech, heart rate, accelerometer), multi-modality (i.e, combining different modalities) and multi-modal fusion approaches (i.e, how and when to combine the modalities) on classifying emotions in a real-life DyMand dataset. No prior work on the DyMand dataset has been done before that explores automatic emotion recognition models. Broadly, this work will also add to the research on couples emotion recognition using early fusion and late fusion approaches which is underexplored for real-life datasets. Secondly, transfer learning will be leveraged to investigate the knowledge transfer capability of emotion recognition datasets on improving the classification of emotions in the DyMand dataset. The main contribution will be transfer learning models in couples emotion recognition that are specifically trained on real life data. The broader contributions of these models are that they can be applied to other real life dyadic datasets and be used in the on-going project for dyadic management of diabetes.

Overview of the thesis

This remainder of this thesis discusses the related work and background knowledge in chapter 2. Followed by chapter 3 which explains the methodology of this thesis including information on pre-training datasets, DyMand dataset, the methodology of multi-modal fusion, and transfer learning by fine-tuning. Subsequently, the results from the experiments are presented in chapter 4. An analysis of the results are discussed in-depth in chapter 5. The limitations and future work are explained in chapter 6. Lastly, the conclusion of this thesis is in chapter 7.

Chapter 2

Related Work and Background

This section introduces the related work and background necessary for understanding the context of this thesis. First, quantifying emotions will be introduced to understand how emotions are classified in literature. Second, multi-modal fusion approaches are explained in relation to related work and how they can be leveraged to improve emotion recognition. Third, couples emotion recognition is explained with a distinction in those who use data from ideal laboratory environments as compared to real life. Fourth, an overview of transfer learning in emotion recognition is provided to compare with existing work and provide information on how transfer learning can enhance emotion recognition. Fifth, a distinction is made between self-reported emotions and externally evaluated emotions as it changes the context of emotion recognition tasks. The former is used to recognize felt emotions and the latter is used to recognize perceived emotions. Finally, the evaluation metric that is used is explained.

Quantifying emotions

In literature, various efforts have been made by psychologists to classify emotions of humans into categories [20,25,51,53] or dimensions [55,62]. Our work focuses on the definition by Russell et al. [59], which identified emotions as dimensions in its "circumplex model of affect [55]." This model is commonly used in emotion recognition research [2, 61, 77]. Specifically, this work classifies emotions on the dimensions of arousal, and valence. Arousal refers to alertness (or activation) and ranges from low (sleepy) to high (excited). Valence refers to pleasantness and ranges from negative (misery) to positive (pleasant) [55].

Multi-modal fusion

Previous work in couples emotion recognition explored uni-modal [35] and multi-modal approaches [9, 18, 19, 54] along with transfer learning [9]. Many work used acoustic data only for emotion recognition [38, 68]. Other works have used linguistic and multi-modal data [64, 65]. There are two main approaches for multi-modal fusion; early fusion and late fusion. For early fusion the features of two or more modalities are concatenate and used as input to the machine learning model as highlighted by [36, 64]. Early fusion is also known as feature-level fusion. However, in late fusion, the modalities are trained separately from each other and finally the raw predictions are combined using an approach of majority voting, averaging, weighted averaging, product, or sum [32, 36, 67]

Couples emotion recognition

Most of the works in couples emotion recognition have used datasets obtained from lab environments [14,72,74]. These often do not generalize well to real-life emotion recognition [18,52]. A reason not many works use real-life data is because of the challenges and cost of procuring, and labelling [37] such quality data for training machine learning models. In addition, emotions in real life data are have noisy signals and imbalanced emotions. For instance, it is difficult to balance being happy or sad throughout the day, or balance feeling active and tired. Hence, this is the motivation to apply transfer learning and multi-modal approaches to improve the performance of predicting emotions of couples [5,18] with less real world data available.

Transfer learning

Transfer learning in emotion recognition has often be used as a feature extractor [9], or leveraging pre-trained models that have been trained on a different, larger dataset to improve the prediction [47]. Typically, research uses pre-trained datasets that are not pre-trained on emotion recognition tasks. For instance, [47] uses ImageNet for leveraging emotion classification using images. While some research has recently used emotion recognition datasets [27], typically from lab environments in ideal conditions to build transfer learning models. Transfer learning has demonstrated success in the domain of emotion recognition [24, 34, 60]. However, there is a lack of transfer learning models that use data from real life environments, are those that are not scripted or whose data is collected in ideal lab conditions. Hence, this is the motivation of this thesis to develop and leverage transfer learning models that are pre-trained on datasets from real life.

Assessment of emotions

There exist various methodologies for collecting information on emotions from individuals. Some work focuses on participants self-report data [7,9,50], in which they are asked to self-assess their emotions perceived [4,10,29]. A subtle distinction is also made when the individuals are asked to self-assess their emotions. This could be right after a stimulus [50], or during the stimulus [9]. The type of self-assessment of emotions used can influence the quality of the emotions. On the other hand, there exist works that label emotions using external evaluators [31]. In such cases, several evaluators use a consistent methodology to assess the perceived emotions of individuals. The final emotion can be combined from evaluators using averaging, inter-rated reliability scores, weighted averaging or other metrics [30]. The inter-rated reliability agreement can provide information on the confidence of emotion recognition by the evaluators. Each of these methods have their pros and cons. The advantage of self-report data can give the ground-truth of the individuals own emotions [80]. However, if the self-report is after the stimulus, the individual will have to recall from memory the emotions they felt, and if the self-report is during the stimulus, it can become a constant distraction and self-awareness, which can influence their actual emotions. On the other hand, using evaluators can be advantageous to collectively assess perceived emotions [80]. However, evaluators can also have an inherit bias and differences in their perception of emotions [30]. Additionally, perceived emotions can be different from felt emotions by individuals [11,73].

Evaluation metrics

As real life emotion recognition data is often imbalanced, it is important to use a metric that address the issue of imbalance. Using accuracy to assess an imbalanced dataset will lead to a high or low performance without the machine learning model recognizing any patterns. For example, a classifier that predicts happy when the dataset has 99 out of 100 happy emotions will be 99%

accurate. On the other hand, emotion recognition tasks for real life data use the unweighted average recall metric (UAR) [21, 66, 69, 70]. This metric averages the recall on the positive and negative class equally to evaluate the performance of a model on both classes. This metric is calculated as the following:

$$((tp/(tp + fn) + (tn/(tn + fp))/2) \tag{2.1}$$

In which tp denotes true positive, fn denotes false positive, tn denotes true negative, and fp denotes false positive.

Chapter 3

Methodology

This chapter introduces the methodology to investigate the proposed research questions. First, I located public datasets to use for pre-training the transfer learning models in order to investigate RQ 2. Second, I shortlisted the public datasets that are similar to the DyMand dataset. The process of locating and shortlisting is described in the Section 3.1. Each of the datasets selected are further described in detail. Third, these datasets are then pre-trained (i.e, pre-processed, feature extraction, hyperparameter tuning, and evaluation) and eary fusion and late fusion approaches are investigated using the same test set for comparison. This would investigate RQ 1. Fourth, the DyMand dataset is trained and evaluated without transfer learning for the baseline approach. The DyMand dataset is also investigated for early fusion and late fusion approaches using the same test set for comparison to investigate RQ 1. Fifth, the pre-trained datasets are used for transfer learning on the DyMand dataset by investigating different fine-tuning layers to investigate RQ 2. The DyMand dataset using transfer learning is then evaluated using the same test set, both with and without transfer learning for comparison. Finally, the results are reported in the Section 4. The source code used in this thesis is also hosted on here. The following section will describe each dataset in-depth to understand the process of data collection for these datasets, followed by pre-training approach, early fusion and late fusion approaches and transfer learning approaches.

3.1 Datasets

The following datasets described in the thesis were a result of search and exploration to find public datasets that met a certain criteria to be used as pre-training datasets for the DyMand target dataset. The DyMand dataset is used as the target dataset as it is the motivation of this thesis to investigate multi-modal fusion techniques (RQ 1) and transfer learning by fine-tuning (RQ2) with the goal to improve the classification of emotions for a in-the-wild dataset.

The criteria was the following: the datasets must have dyadic interaction, the datasets must participants with who engage in natural/unscripted and spontaneous interactions, the datasets must have emotion annotations in the arousal-valence dimension, the datasets must have the modalities present in DyMand (acoustic, lexical, and sensor data) and the acoustic/lexical data must be in the Swiss-German or German language and come from the same source. These criteria help in finding a dataset that is similar to the DyMand dataset in terms of context of data collection, the emotion annotations dimension and the modalities present in the DyMand dataset for effective transfer learning.

The search led to 12 public acoustic and linguistic datasets as shown in Table 7.1, and 6 sensor datasets as shown in Table 7.2. A summary of these datasets is shown in the Appendix A. A majority of the acoustic and lexical datasets found were in English and some did not have their respective transcriptions. Only a subset of the acoustic and lexical datasets was in German, had both acoustic and lexical from the same source, and the emotions were in the arousal-valence dimensions. This includes the VAM dataset and the CMU-MOSEAS data. However, since the CMU-MOSEAS dataset was not available during the course of the work, the VAM dataset was finalized for the acoustic and lexical pre-training. For the sensor datasets, a majority did not have the accelerometer modality or ambient light. At the end, only two candidates were the closest match to the criteria aforementioned, having both accelerometer and heart rate sensors; the WESAD dataset [63] and the K-EmoCon dataset [50]. However, since the WESAD dataset was not in-the-wild/spontaneous and only a subset of the dataset was dyadic, it was not finalized, and rather the K-EmoCon dataset was finalized due to it meeting the most criteria and have a large number of participants and sensor data.

3.1.1 VAM

The Vera-am-mittag (VAM) [31] dataset is a public dataset that was created from a German TV show by same name, Vera-am-mittag. This TV show was based on a group of guest speakers (2-5) who were invited to the show in each episode to discuss personal topics such as "friendship crises, fatherhood questions, or romantic affairs [31]" with the program host and moderator, Vera.

The motivation of creating the VAM dataset was to provide audio-visual data with "authentic and spontaneous interaction" of participants and have their accompanying annotated emotion data. The authors decided to use this TV show in particular, due to several reasons. First, the concept of the TV show revolved around real guest speakers who discussed highly emotional topics. Second, the participants were not aware that their dialogues would be used for emotion analysis research. Third, there is a reasonable amount of speech available per participants.

The dataset is based on 12 broadcasts of Vera-am-mittag aired between December 2004 and February 2005 [31]. The authors used 10 out of 12 of the broadcasts for further processing into an audio-visual dataset. The 2 broadcasts that were not used were cited having "not relevant [..] affective content [31]." From the 10 broadcasts, the authors extracted videos per discussion, and obtained 45 videos. The videos contained both audio and visual signals. Afterwards, the dialogues were segmented into utterances, or statements, which are roughly the length of a sentence. The audio and visual data was stored separately, and the audio data was later transcribed in German.

As another processing step, the speakers in the dataset were classified into categories which described the usability of data from that person. These categories were *very good, good, usable* and *not usable* [31]. The *very good* speakers were those who had many utterances and showed a wide range of emotions in the utterances. Similarly, the *good* speakers were categorized as those who had many utterances but showed a limited amount of emotions. The *usable* speakers had less utterances, while still showing emotions and *non usable* speakers had limited utterances, and showed little to no emotions.

To create the audio and transcribed dataset of VAM, there were two iterations. In the first iteration, the authors only considered participants classified as *very good*. These were 19 participants that had a high number of utterances and a wide range of emotions. Each of the utterances in the *very good* speakers were evaluated by 17 different human evaluators who listened to the utterances and labeled them with emotions using SAMs. The emotions are annotated on a range between -1 and 1. The authors noticed that the emotion labels were highly imbalanced, and thus in the second iteration, their introduced the *good* speakers as well, which had 28 participants. However, the utterances from the *good* speakers were only evaluated by 6 different human evaluators. The emotions were from the evaluators were averaged using an approach of evaluator weighted estimator [30] which attempts to find the average rating while accounting for evaluator bias using correlations between evaluator ratings and their standard deviation.

VAM Dataset Properties	
Subjects	47
Age	16 to 69
Gender	36 female, 11 male
# of Audio data	947
# of Transcribed data	947
# of Emotion data	947

Table 3.1: VAM Dataset Properties

In total, the VAM dataset has 47 participants (ages between 16 to 69 years) with 947 utterances. The average duration of an utterance is 1.01 seconds per speaker, as found during data exploration. Each of these utterances has a corresponding German transcription. A summary of the properties of the VAM dataset is seen in Table 3.1. The data is further explored in Section 3.2.1.

3.1.2 K-EmoCon

The K-EmoCon dataset [50] is a publicly available (with prior approval from data owner) multimodal emotion dataset capturing naturalistic dyadic interactions. Naturalistic datasets can be defined as datasets that have spontaneous and unscripted interactions. The dataset is created to address the lack of naturalistic interaction data that has accompanying labelled emotions. The data was collected from a field study in South Korea (from January 2019 to March 2019) in which two participants are randomly chosen to debate on the social issue of the Jeju Yemini refugee crisis for 10 minutes. Prior to the debate, the participants were sent 4 news articles by email, giving perspectives on a for-favor, not for-favor and neutral views on this topic. A total of 32 participants (ages between 19 to 36) were recruited after filtering for an inclusion criteria of either living in an English-speaking country for a minimum of 3 years or achieved a score above a certain threshold in a standardized English exam such as a minimum of level 7 in speaking in the IELTS exam. This ensures the level of English was adequate for the debate and the participants did not struggle with the language barrier, as the debate was in English.

For the data collection, two rooms were allocated that had controlled temperature and lighting for all of the debates. The participants were asked to sit on the ends of the table facing each other. At the center of the table, there were two tripod mounted smartphones which recorded the video (audio and visual) of the respective participants. In addition, the participants also wore 4 devices that collected multi-modal sensor data. These include an *Empatica E4 wristband*, *Polar H7 Bluetooth Heart Rate Sensor*, *NeuroSky MindWave Headset* and a *LookNTell Head-Mounted Camera*. Relevant for the pre-training stage, the *Empatica E4 wristband* collected a 3-axis acceleration data, and photoplethysmography (PPG) which was used to derive the heart rate data. The collected sensor data from each of the wearable devices is synchronized by a Unix timestamp. The heart rate data from the *Polar H7 Bluetooth Heart Rate Sensor* was not used as it was another device than the one used to collect the accelerometer and secondly it was not placed on the wrist (such as the E4) but rather on the chest, which is different to the DyMand heart rate data collection as seen in Section 3.1.3. Additionally, the audio data was not relevant as the language was in English.

The sensor data in the dataset is a clipped version, which removes the recording period before and after the debate. However, there is still a period roughly 2 minutes before the actual start of the debate that is in the dataset. This period is the baseline for measuring the neutral state. This period is later removed when doing the actual data processing in the data-processing step in Section 3.2.1.

Distinct from other datasets, the K-EmoCon dataset has emotion annotations for 3 perspec-

K-EmoCon Dataset Properties	
Subjects	32
Age	19 to 36
Gender	12 female, 20 male
# of Sensor data	varies between 4,017 to 4,159
# of Emotion data	4,159

Table 3.2: K-EmoCon Dataset Properties

tives. The self-rated emotions by the participant, the emotions rated by the partner of the debate, and 5 external raters. In addition, each of the emotions annotated range from the arousal-valence dimension to categorized emotions such as angry, happy, nervous, sad etc. The emotions in the arousal-valence dimension are annotated in the range of 1 to 5, which 1 denoting very low, 3 denoting neutral, and 5 denoting very high. The emotions were annotated by each participant (and the partner) after the debate. The participants had a break of 15 minutes after the debate and were assigned a PC where they watched two 2nd person point-of-view recordings; one of themselves and another of the debate partner. The participants then annotated emotions every 5 non-overlapping seconds since the start of the debate to the end of the debate.

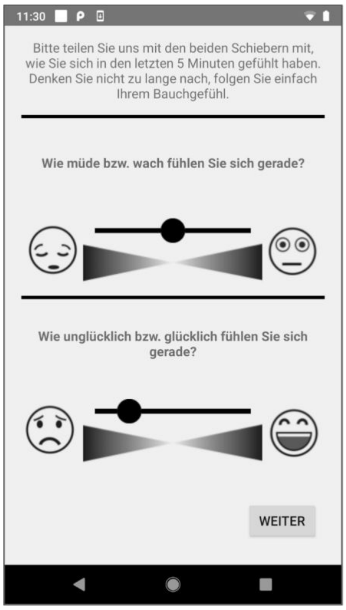
The dataset is also accompanied by meta data which gives information about potential outliers, the completeness of the data, the duration, the start of data collection, and the start and end times of debate for each participant. A summary of the properties of the dataset can be seen in Table 3.2. The data exploration can be seen in Section 3.2.1.

3.1.3 DyMand

The DyMand dataset [8] was created in a field study to collect objective data on the dyadic management of diabetes (DyMand) project. The field study was conducted between 2019 and 2021 with 13 heterosexual couples (N=26; ages 47 to 81 years) from the German-speaking part of Switzerland in which one partner was managing type-2 diabetes. The duration of the field study for each couple was 7 consecutive days. Each participant was given a Polar M600 smartwatch and a Nokia 6.1 smartphone for the duration of their field study. The devices were marked with black covers (for the supporting partner) and white covers (for the patient managing type-2 diabetes) for the devices not to get mixed up over the course of the study. Before starting the experiment, research assistants from the study helped the participants to set up the smartwatch and smartphone paring in order to check the equipment and familiarize the participants with the devices. The participants also received instructions to always keep the devices with them from waking up to going to bed for 7 consecutive days. Additionally, the couples together decided a common data collection period. For the morning hours, they could select a period between the hours of 4am to 11am and for the evening hours, they could select a period between the hours of 4pm and 11pm. During the weekend, they could only select a period between 6am and 10pm. The authors claim this reduces the burden of the participants who have to also self-report their emotions with every audio and sensor data collected and addresses privacy concerns by reducing the number of audio recordings [8].

During the study, the data was collected from the couples during their pre-defined data collection period. There are several criteria that needs to be met for the data collection process to start and complete. First, the data collection process can only start during the pre-defined hours set by the couples. Second, data collection only starts when the couples who are wearing their smartwatches are physically close to each other, which is estimated by the smartwatch app that uses the Bluetooth signal strength between the smartwatches to determine proximity. Third, the app uses a voice activity detection machine learning algorithm to distinguish between speech and other noise. In case there is speech, and the criteria has been fulfilled, then data collection starts by recording the audio and sensor data (ambient light, gyroscope, accelerometer, and heart rate) for 5 consecutive minutes. In the case where speech, or physically closeness was not detected in the data collection period, then a backup recording is collected by the app in the last 15 minutes of the hour.

After the 5-minute recording ends, the smartwatch notifies the participant by vibrating and then opens the self-report Affective slider on the smartphone for the participant to report their emotions. The Affective slider ranges from 0 to 100. The participant, however, does not see the value of the slider. To help the participant understand the range, there is an emoticon depicting the emotion (tired to attentive) for arousal and (upset to elated) for valence as shown in Figure 3.1. The smartwatch and smartphone communicate with each other to determine whether the participant has started the self-report. The smartwatch waits for 2 minutes for the participant to start filling the self-report. In case the participant has not started filling the self-report in this time, the smartwatch sends another notification by vibrating and waiting for another 2 minutes. In case the participant is still not able to start the self-report in this period, the smartwatch app deletes the collected audio and sensor data and attempts data collection again during the hour. Thus, the participant has a total of 4 minutes to start the self-report after data has been collected. In addition, “the app ensures at least 20 minutes between the subsequent data collection [8].”



Bitte teilen Sie uns mit den beiden Schiebern mit, wie Sie sich in den letzten 5 Minuten gefühlt haben. Denken Sie nicht zu lange nach, folgen Sie einfach Ihrem Bauchgefühl.

Wie müde bzw. wach fühlen Sie sich gerade?

Wie unglücklich bzw. glücklich fühlen Sie sich gerade?

Figure 3.1: DyMand Affective Slider Interface Source: [8]

In total, there are 1,017 5-minute recordings of audio data collected (85 hours) in which 797 contain speech. From these, 564 audios contain speech between the couple in which both the female and male couple spoke. Additionally, the audio data is later manually acoustic into German that compliments the audio modality (acoustic modality). Additionally, the audio data is manually speaker-diarized which gives specific information about when the male or female participant is speaking in the audio. However, only 380 audio samples have a corresponding self-report for both arousal and valence emotions where both the male and female couple are speaking. It is

DyMand Dataset Properties	
Subjects	26 (13 couples)
Age	47 to 81, mean 68, std 9
Gender	13 female, 13 male
# of Audio data	1017
# of Sensor data	1017
# of Emotion data	608
# of Audio data dyadic	564
# of Sensor data dyadic	564
# of Emotion data dyadic	380

Table 3.3: DyMand Dataset Properties

important to have data where both male and female couple are speaking to get the dyadic context of their interaction. The emotion data is pre-processed to binarized emotion values for arousal and valence. As the original self-report emotion ranges between 0 and 100, it is binarized to 0 if the value is less than or equal to 50 and 1 if the value is greater than 50. In addition, the raw data is accompanied with a meta-data file which gives information about the audio file, its corresponding self-report emotion, and other details such as if it contains speech. A summary of the data can found in table 3.3. Further data exploration is conducted in Section 3.3.1.

3.2 Pre-training Models

This section includes the pre-training methodology. Pre-training is an important step for transfer learning. The pre-trained models are built using the data from the pre-trained datasets. In this phase, the VAM dataset is used for pre-training the acoustic and linguistic early fusion and late fusion models, and K-EmoCon is used for pre-training the early fusion and late fusion heart rate and accelerometer models.

3.2.1 Data exploration and pre-processing

VAM dataset

In the VAM dataset, each participant has a folder which contains their respective utterances. The utterances are saved as wav files, and named according to "Satz" + participant ID + utterance number. For each utterance the average evaluator emotion label and the individual evaluator emotion labels can be found in two respective files with the same filename, but with different extensions. The transcriptions for the utterances are located in another folder structure which contains a Microsoft Excel spreadsheet of the File name, phonetics, transcription, and columns for the averaged emotion labels, and their respective standard deviations per utterance.

As a pre-processing step, I write a bash script that gathers all the audio files into one folder. This reduces the execution time of the python script to recursively go through each directory individually. The bash script finds all wav files in the dataset and copies them to one folder.

In another pre-processing step, I binarize the emotion labels to resemble the format of the emotion labels in the DyMand dataset. Emotion labels with values less than 0 are defined as 0 and with values greater than or equal to 0 are defined as 1. Although the 0 indicates neutral, it is important to not exclude it from the binarization, otherwise it becomes a 3-class problem.

For data exploration, it is important to understand the data distribution of emotions. Figure 3.2 summarizes the binary arousal values for each participant in the dataset. In the arousal dimension, participants with ID 18, 37, 44, 46, 47 do not have either a low or high arousal labels. Additionally, there are participants who have a large imbalance between the binary arousal such as participant ID 13, 19, 22, and 39. Besides these participants, the arousal dimension has a reasonably balanced low and high arousal classes.

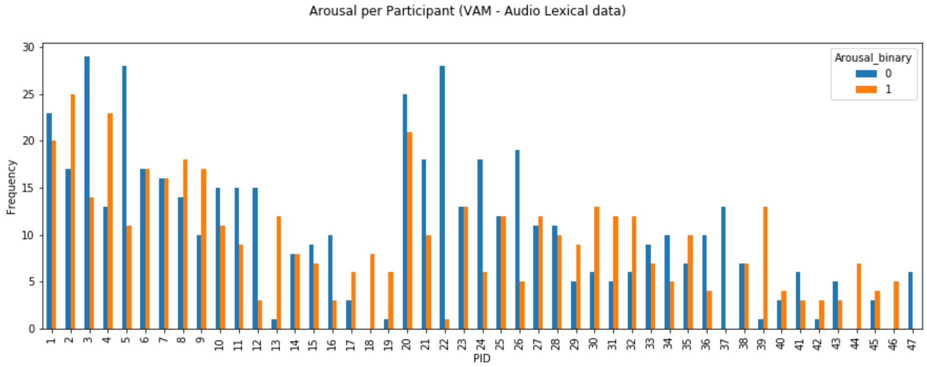


Figure 3.2: VAM Arousal per participant

However, the class imbalance increases even more for the valence dimension. Many participants do not have a positive valence as seen in Figure 3.3. These include participants with ID 2, 4-6, 8, 9, 13, 15-19, 34, 40, 44-47. An explanation could be the context of the data. As the VAM data revolves around highly personal and emotional topics such as romantic affairs or fatherhood questions, it is reasonable to expect not as many positive valence data within participants.

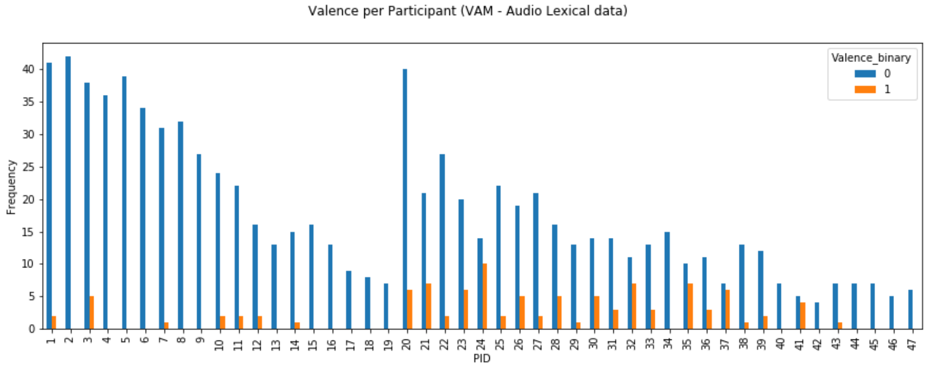


Figure 3.3: VAM Valence per participant

When plotting the arousal and valence dimensions together per participant, the class imbalance becomes more clear between arousal and valence dimensions as seen in Figure 3.4. It is seen that the valence dimension is more imbalanced than the arousal dimension. Some participants are missing 1-2 classes of low arousal, negative valence, high arousal or positive valence, such as participants with ID 18, 44, 46, and 47.

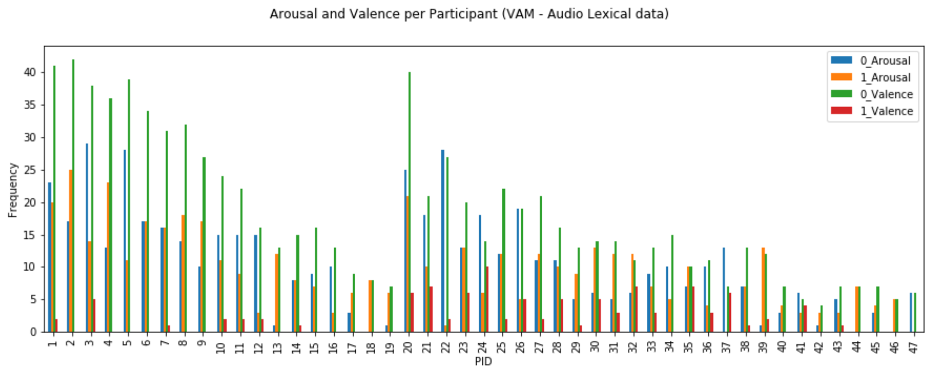


Figure 3.4: VAM Arousal and Valence per participant

In addition, a distinction between the first and second iteration of the VAM audio dataset can be seen. The participants with IDs 1-19 were from the *very good* category of speakers who had a high number of utterances and a wide range of emotions, and IDs 20-47 were from the *good* category of speakers with many utterances, but less range of emotions. However, analyzing Figure 3.4 it can be seen that some participants in the *good* category could have been placed in the *very good* category. Also, a majority of relatively balanced valence labels are from the *good* category as seen in Figure 3.3.

Another pre-processing step for the VAM dataset is to remove the participants from the emotion dimension who had missing classes in either the low arousal, negative valence or high arousal, positive valence classes. Removing these participants is important as training, validation or test sets could have these participants and they could change the behavior of the machine learning models. For example, a highly imbalanced participant for the valence dimension, such as participant 4, could have a very high accuracy if the machine learning model only predicted 0s for valence.

K-EmoCon dataset

The K-EmoCon dataset is structured similar to the VAM dataset where each participant has their own folder. In this case, there are two folders; one for the E4 data, and another for the NeuroSky and Polar data. Relevant for this thesis is the E4 data. Inside this folder contains csv files for the sensor modalities (accelerometer, heart rate, EDA, IBI, BVP, and temperature). These sensor modalities vary in terms of content, but typically have the timestamp (in Unix) the participant ID, the raw sensor data, device serial, device number, and entry time. Participants IDs 29, 30, 31, and 32 have two device serials instead of a unique serial. The authors of the dataset mention to only use one of the device serials and also give suggestions. The reasoning is not explained. For participant IDs 29, and 31 it is recommended to use "A013E1" as the device serial and for IDs 30, 32, it is recommended to use "A01A3A" as the device serial. By filtering samples to these devices serials, the raw data reduces to roughly half for these participants.

Sensor data is also susceptible to missing values or outliers caused by different and unpredictable reasons. For example, non-contact of the wearable device, malfunctioning of the device, or loosing connection to the device. Raw sensor data measurements can also have noise which can be due to their highly-sensitive nature, such as accelerometer data which measures the acceleration of an object. For pre-processing the heart rate data, a common approach is to apply a minimum and maximum range. For example, a human heart rate can not be below 30 beats per minute or higher than 300 beats per minute according to several studies [39,79,81]. These values of 30 beats per minute as the minimum and 300 beats per minute as the maximum were used. This eliminates raw heart rate data values not in the threshold. For pre-processing the accelerometer data, an approach is to apply a filter that removes data that is 1, 2 or 3 standard deviations away from the mean. This filter removes the sudden movements that are captured by the accelerometer sensor. As seen on Figure 3.5, after removing data points more than 1 standard deviation away from the mean, the information on the movements is not lost, but rather is more uniform.

In the K-EmoCon public dataset, the raw sensor data was clipped to remove the samples before the start of the debate and after the end of the debate. However, there is a period of 1-3 minutes that varies per participant, which includes the baseline measurement of their neutral state when the debate has not begun. This baseline period needs to be removed when merging the emotion labels. Although the each sensor data contains information about the timestamp for each sample which helps in synchronizing the sensor modalities, this is not the case for the emotion labels. It is important to point that there are several emotion labels for K-EmoCon. As described in the related work, K-EmoCon has self-reported emotion labels, debate partner-reported emotion labels, and external-reported emotion labels. This work uses the self-reported emotion labels as the focus of this work is on felt emotions, as explained earlier. These self-reported emotion labels are annotated every 5-second non-overlapping without information about the respective timestamps. As the authors describe in the paper that the emotions were rated by the individuals, the partner of the debate and external raters by watching their audio-visual recordings of the debate and it did not specify if they rated the baseline period, it is assumed that the emotions were rated at the start time of the debate. Therefore, to merge the sensor modalities with the respective emotion labels, there are several pre-processing steps required. First, the sensor data is cropped to include samples from the start time of the debate to the end time of the debate. The information about the start and end time of debates is different for each participant is found in the metadata information file provided. Second, the sensor modalities have different sampling rates. For example, the accelerometer sensor has a sampling rate of 32Hz, whereas the heart rate has a sampling rate of 1Hz. However, the sampling rate of the emotion labels are every 5 seconds. In order to merge the sensor data with the emotion labels, they must be converted to the same sampling rate.

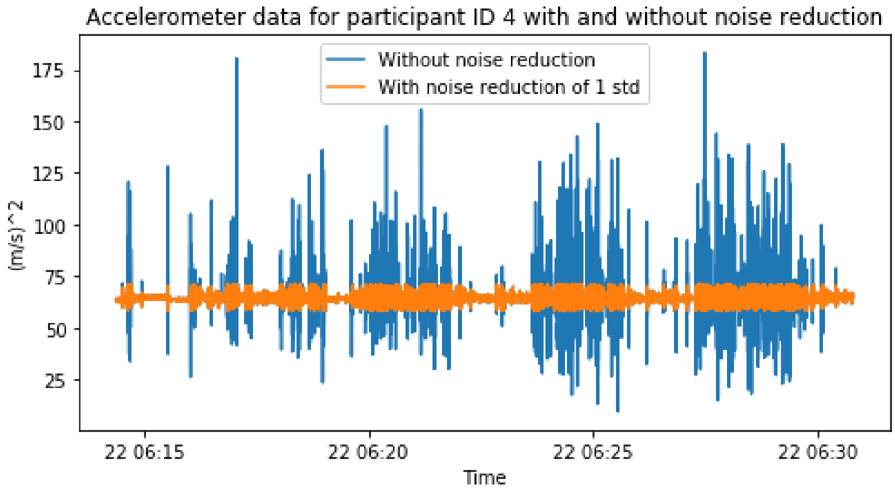


Figure 3.5: Accelerometer data pre-processing example with participant ID 4. Removing data 1 standard deviation away from the mean for noise reduction.

There are several ways to re-sample sensor data for them to align to a target sampling rate. This usually involves up-sampling or down-sampling the sampling rate. Up-sampling increases the sampling rate of the source data by interpolating between points. Whereas, down-sampling decreases the sampling rate of the source data by only selecting points at the target sample rate. However, there are disadvantages to this approach, and in particular, to the current problem. By applying an down-sampling approach to the heart rate sensor to match the emotion labels, the heart rate variability over the 5 seconds would be lost and only 1 heart rate beats per minute value will aligned to the emotion label. Likewise, for down-sampling the accelerometer to match the sampling rate of the emotion label will result in a loss of movement information. To avoid this loss of information, I use the concept of grouping data into non-overlapping windows of 5 seconds. In each of these windows, I compute features of the sensor data which describe the movement or heart rate variations during those 5 seconds. These features are described in detail in Section 3.2.2

The emotion labels in K-EmoCon are represented on a Likert scale. These values range from 1 to 5, with 1 being very low and 5 being very high. The authors denote 3 as being neutral. To binarize the emotion labels, I define values less than 3 as 0, and values more than or equal to 3 as 1. This is in accordance to other research that uses the K-EmoCon dataset [1,76].

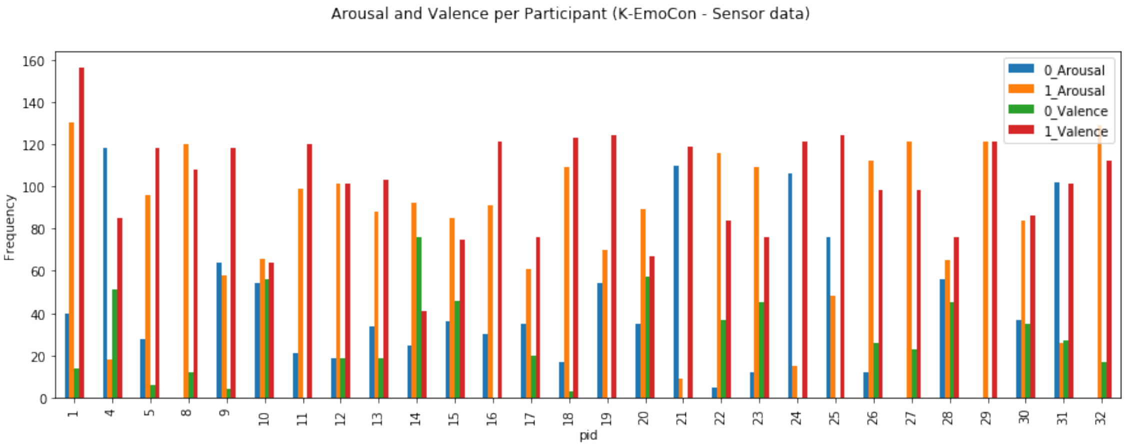


Figure 3.6: K-EmoCon Arousal and Valence per participant

Similar to the VAM dataset, the K-EmoCon dataset has participants who do not have data for all the emotion classes of high arousal, positive valence, low arousal or negative valence. These participant's need to be removed from the training, validation and test sets in order to not bias the models. The Figure 3.6 illustrates the arousal and valence binary classes per participant. The participants with IDs 8,27,29,32 do not have a low arousal class, and 11,16,19,21,24,25,29 do not have a negative valence class.

3.2.2 Feature extraction

VAM dataset

The VAM dataset has the acoustic modality and the linguistic modality. Both of these modalities are distinct from each other, yet their hold the same information. Both of these modalities can give insight into the emotions portrayed. For example, the tone of voice can be an indicator of the arousal emotion (low vs high) and the semantics of the transcripts can give information about the the valence emotion (negative or positive). In certain cases, they can also indicate vice-versa if the emotions are portrayed on very high or very low ranges.

In the field of natural language processing, many linguistic feature extraction techniques and models exist that are suitable for the task of semantic similarity. The goal of these techniques and models are to use the words, sentences, language structure and other linguistic rules to understand the semantic meaning. BERT [17] is a pre-trained transformer model that has gained popularity in the domain of natural language processing for its ability to achieve state-of-the-art results in several tasks. The last layer of BERT can be fine-tuned to linguistic datasets, and prior research has shown improvements in performance in several tasks. However, in recent years, newer models have been developed that optimize the architecture of BERT that can perform better on certain tasks. The Sentence-BERT model [57] modifies the architecture of BERT, by using a siamese networks [15] to compute sentence embeddings such that semantically similar are close in the vector space. Sentence-BERT has shown improvements and outperforms BERT in semantic similarity tasks. As the linguistic data in VAM is in German, the German BERT *cased* model is used, which has been trained on approximately 12GB of text data from Wiki, OpenLegalData and the News [46]. The German BERT *uncased* model is not used as it pre-processes all the text to lowercase which looses semantics in German.

The SentenceTransformer package for Python is used to create the sentence embeddings with the German BERT *cased* model. The maximum sequence length used is 512, which corresponds roughly to the maximum number of words the model will take as input at a time. For the German BERT *cased* model. 512 is the maximum sequence length it supports. Sentence inputs longer than this gets truncated. For generating the output, the mean pooling setting was used, this results in a 768-dimensional feature vector for each input.

For the acoustic modality feature extraction, several features exist that are fit for different tasks of acoustic emotion recognition. A widely used feature set is the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [21] which contains 88 features that combines low-level-descriptors, frequency descriptions, spectral descriptors, and cepstral parameters to describe the acoustics. This feature set has shown improvements in both arousal and valence balance accuracy metrics, and has also been used for prior experiments with the VAM dataset and other commonly used datasets for emotions. The eGeMAPS feature set was extracted using the Python package OpenSmile [22], which the authors note as their official publicly available implementation [21].

K-EmoCon dataset

The K-EmoCon dataset has several sensor modalities, with the accelerometer data and the heart rate data being the relevant to the task of transfer learning to the DyMand dataset. As these sensor modalities have different sampling rates, and the emotion labels has a much lower sampling rate, in the pre-processing section, I grouped them into 5 second windows. To preserve information from these 5 second windows of the heart rate variability and the acceleration movements, statistical features are highly relevant. Prior work using sensor data has given attention to statistical features information during the time window. Similar to prior work, this work uses the mean, maximum, minimum, range, first quartile, second quartile, third quartile, standard deviation, skewness and kurtosis [26, 33, 45, 58].

These statistical features were computed after grouping data in 5-second non-overlapping windows and using the Pandas library and its respective statistical functions to extract the features. These feature vectors were then grouped for all of the windows to create the sensor dataset.

3.2.3 Early fusion approach

Multi-modal data contains different perspectives on the phenomena occurring during a period of time. Often a challenge in multi-modal fusion is the methodology of fusing together signals. As previously seen, signals can have varying sampling rates, or in raw data, they might not be compatible, such as fusing acoustic and linguistic data. However, the pre-processing steps and feature extraction approaches enable us to represent the modalities such that they hold the semantic representation of the modalities and these feature vectors can be fused.

Early fusion is a multi-modal approach of combining several modalities into a unified space. The features from several modalities are fused together to create a larger dimensional feature space which is used as input before the training process. Humans also communicate in a multi-modal manner, which entails combining information from acoustics and from visual senses to understand each other more effectively. The early fusion approach enables the model to use the combined features of the modalities and find inter-relations between the modalities for its final prediction. An overview of the approach can be seen in Figure 3.7.

VAM early fusion

There are 947 samples in the acoustic modality and linguistic modality. After extracting the relevant features for each modality, there are 88 features for the acoustic modality and 768 for the linguistic modality. When merging the features for early fusion, there are 947 samples and 856 feature vectors. For early fusion, there are two models that needs to be created; one for predicting the arousal emotions, and another for predicting the valence emotions. As previously noted, there exist an participants who have missing emotion classes within the arousal or valence dimensions. After removing these participants from each of the emotions, the early fusion dataset used to predict arousal has 908 samples and 856 features. Whereas, the early fusion valence dataset used to predict valence now has 624 samples with 856 features. An overview of the early fusion approach for acoustic and linguistic data can be seen in Figure 3.7

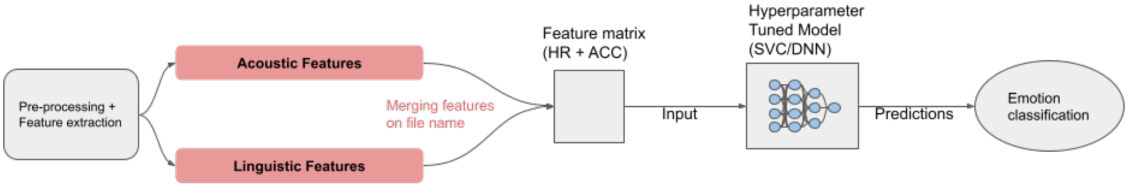

Pre-processing + Feature extraction
Acoustic Features
Linguistic Features
Merging features on file name
Feature matrix (HR + ACC)
Input
Hyperparameter Tuned Model (SVC/DNN)
Predictions
Emotion classification

Figure 3.7: Early fusion approach for acoustic and linguistic data

Prior to training the models, a strategy for splitting the training, validation and test data is important. The training set will be used to train the respective models, the validation set is used by the model to prevent over-fitting, and provide information on the performance of the trained model, and the test set is used to evaluate the model. The model is never aware of the test set during training or hyperparameter tuning, and thus, can serve as a good indicator of the model on unseen data. By performing a split based on the number of samples, for example 60-20-20, for the training, validation and test set, respectively, is not the correct approach as it causes data leakage. For instance, the data of one participant can be spread in the training set, and in the test set. This can cause the model to perform better since it has seen the vocal features, the types of words and expressions used by the participant in the training set. Hence, in order to build a generalizable model that performs well on unseen data, a different strategy is required to split the data into these three sets. An option is to specify the groups in the split, such that data from one group (represented by data from one participant) stays in either the training, validation or test set and does not leak to other sets. However, it is not sufficient to only define groups. As seen from the data exploration of VAM, the emotions within each participant differ widely from each other. The StratifiedGroupKFold is a cross-validation set splitting function from the scikit-learn library in Python that creates the training and test sets such that the proportion of target classes in the training and test set are roughly balanced. For example, if a training set has a ratio of 10:2 positive to negative valence, then the test set will approximately resemble this ratio. An exact match is highly difficult as not only the emotions expressed, but also the number of samples differ within each participant. For deciding the ratio of the number of participants to be in the training set vs the test set, the *n_splits* parameter can be controlled.

For VAM, the train and test split was done using the StratifiedGroupKFold with *n_splits* as 10 and selecting only the indexes from the first fold. Since StratifiedGroupKFold is a cross-validation function, the number of splits indicate the number of folds for cross-validation. Based on this, the function allocates the ratio of participants in the train and test sets in order to meet the number of folds. For the arousal training set, the following participants were included: [1, 3-12, 14-27, 29-32, 34-45]. For the arousal test set, the following participants were included: [2 13 28 33]. The training and testing sets also stay fixed for the late fusion approach for the arousal data in order to enable comparability. Whereas, for the valence training set, the following participants were included: [1, 3, 7, 10-12, 14, 20, 22-27, 29-33, 35, 36, 38, 39, 41-43]. For the valence test set, the following participants were included: [21, 28, 37]. This training and testing set also remains the same for the late fusion approach for valence data for comparability.

A Support-Vector Machine (SVM) and a 3-hidden layer neural network was trained for classifying both arousal and valence. First, an SVM was hyperparameter tuned using GridSearchCV, a hyperparameter tuning function from scikit-learn, that searches across all the specified hyper-parameters while also applying cross-validation. The search space included 192 different combinations of hyper-parameters and cross-validated for 10 folds, resulting in 1,920 times the SVM model was run to find the best hyper-parameters. The best hyper-parameters found by GridSearchCV was returned, and used to evaluate on the test test.

Second, for the 3-hidden layer neural network, a validation is required for the model to prevent over-fitting on the training data. The validation set should also be similar to the class ratios of the training and test sets, and thus it requires the StratifiedGroupKFold split. However, as the StratifiedGroupKFold does not allow to create 3 sets, it was not possible to make these in the beginning. The validations set can be created as a subset from the training set. The StratifiedGroupKFold is applied with *n_splits* as 6, and it returns the validation set as [3, 11, 25, 42]. The training set then reduces as these participants go into the validation set. The neural network is then hyper-parameter tuned using Keras Tuner, a hyperparameter optimization library, and the hyperband class algorithm, which efficiently finds the best hyper-parameters using a combination of early stopping, randomized search, and successive halving. To evaluate the best hyper-parameters, I first use StandardScaler from the sci-kit learn library, to scale the training and validation set to have a mean of 0 and standard deviation of 1. This scaler is then used to transform the test set using the function it used to scale the training and validation set. Then the scaled training and validation set is used to train the neural network with the best found hyper-parameters and evaluate on the test set.

The metrics used to evaluate the performance of the models are not accuracy but rather balanced accuracy or unweighted average recall (UAR). In highly imbalanced datasets, such as the VAM, K-EmoCon and DyMand dataset where the emotion classes are imbalanced across the dataset, a better measure is the balanced accuracy which takes into account the weights of the classes and thus scores poorly if the model is not able to classify the minority classes.

K-EmoCon early fusion

In K-EmoCon, there is the heart rate sensor modality and the accelerometer sensor modality having 3,450 samples and 3,451 samples, respectively. The accelerometer sensor modality has one more sample than the heart rate modality due to it being recorded 5 seconds longer for participant ID 17. After extracting the statistical features of both modalities, there are 10 dimensional features for each modality. When merging the features for early fusion, there are 3,450 samples with 20 dimensional features. The next step is to create one model for the arousal emotion and another for the valence emotion. However, as K-EmoCon has participants who do not have either a low arousal, negative valence, or high arousal, and positive valence class, these participants are removed so they do not bias the model. After removing these participants for early fusion arousal, there are 2,959 samples and 20 features. Whereas, after removing these participants for early fusion valence, there are 2,601 samples and 20 features. An overview of the early fusion approach can be seen in Figure 3.8

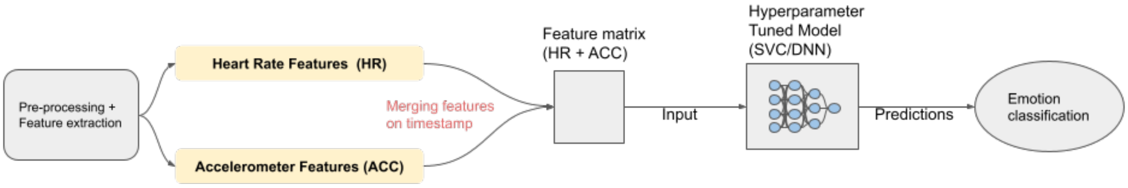


Figure 3.8: Early fusion approach for sensor data

The train and test set split was done using StratifiedGroupKFold with *n_splits = 8*. The number of splits was less than in VAM due to K-EmoCon having less participants. The participants in the train set for early fusion and late fusion arousal are the following: [1, 4, 5, 9 10, 11, 13-17, 19-22, 24-28, 30, 31]. The participants in the test set for early fusion and late fusion arousal are the following: [12, 18, 23]. For the valence train set, these were the participants for early fusion and late fusion valence [4, 5, 8-10, 12-15, 18, 20, 22, 26, 27, 28, 30-32]. For the valence test set, these were the participants for early fusion and late fusion valence [1, 17, 23].

Similar to the VAM early fusion, a SVM and 3-hidden layer neural network was trained to classify emotions for both arousal and valence. First the SVM was hyperparameter tuned with GridSearchCV with 6 splits and then evaluated with the test set. Second, a validation set was created for hyperparameter tuning the neural network. Using StratifiedGroupKFold, it stratified the set to create a validation set from the training set with similar emotion balance as in the training set. The hyperparameter tuning was done using the Keras Tuner which found the number of neurons in each hidden layer, the dropout rates after each hidden layer, and the learning rate. The training and validation set are combined again and passed on the StandardScaler function which scales it to have a mean of 0 and standard deviation of 1. The scaler is used to transform the test set. This scaled training and validation set and the best hyper-parameters are used in the neural network and evaluated using the test set.

3.2.4 Late fusion approach

Late fusion is also a multi-modal approach of combining several modalities to improve the performance of a model. In late fusion, each of the modalities are trained separately to find the best models for them and the predictions from each of these models are fused together. Late fusion is also known as decision level fusion for this reason. There are several ways of combining the predictions from each model. One common approach is to take the average of the predictions [32,48]. The methodology for the late fusion approach similar for both VAM and K-EmoCon in order to standardize the methodology across the pre-trained datasets. An overview of the late fusion approach is shown in Figure 3.9 for the Acoustic and Linguistic Data.

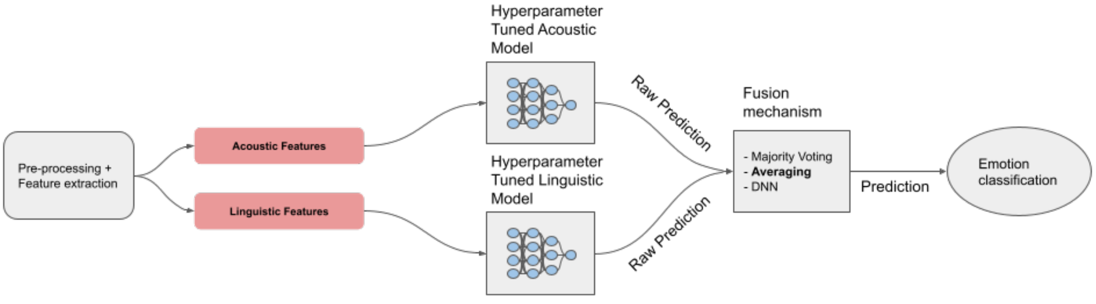


Figure 3.9: Late fusion approach for Acoustic and Linguistic Data

VAM late fusion

In VAM late fusion, the acoustic and lexical modalities are trained separately. For the acoustic modality, an SVM and a 3-hidden layer neural network was trained for both arousal and valence dimensions. These models were hyperparameter tuned and evaluated using the same train, validation and test sets as in VAM early fusion. Similarly, for the linguistic modality, an SVM and a 3-hidden layer neural network was trained for both arousal and valence dimensions. These models were also hyperparameter tuned and evaluated using the same train, validation and tests sets as in VAM early fusion. The raw predictions from the best hyperparameter tuned models were combined by averaging. Each of the modalities have raw predictions from the DNN which range in values from 0 to 1 as a result of the last prediction layer in the DNN being a sigmoid function. After averaging these predictions, they are converted into binary values where values less than or equal to 0.5 are 0 and greater than 0.5 are 1. This final prediction is then evaluated on the test set.

K-EmoCon late fusion

In K-EmoCon late fusion, the accelerometer sensor modality and heart rate sensor modality are trained separately. For both the modalities, an SVM and a 3-hidden layer neural network was

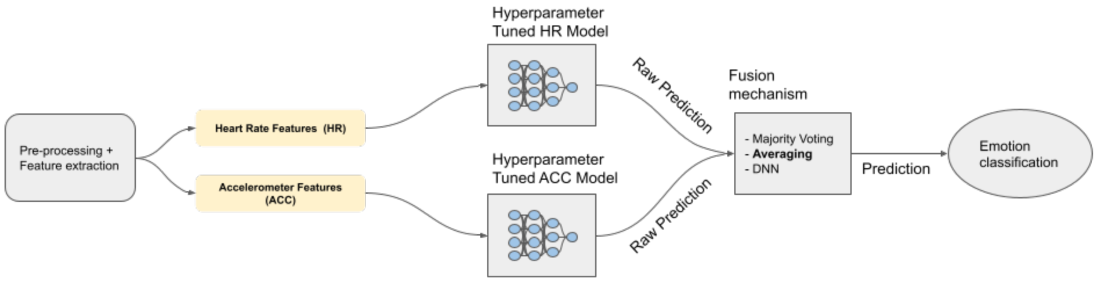


Figure 3.10: Late fusion approach for Accelerometer and Heart Rate Data

trained for both arousal and valence dimensions. These models were hyperparameter tuned and evaluated on the same train, validation and test sets as in K-EmoCon early fusion to enable comparability. The predictions from the best hyperparameter tuned models were combined by averaging the raw predictions to keep it standardized and then binarized using the approach aforementioned in VAM late fusion. The final prediction is evaluated on the test set.

3.3 Transfer Learning on DyMand

3.3.1 Data exploration and pre-processing

The DyMand dataset is organized in several folders. There is a folder for the smartwatch data, which has sub-folders on the patient (P) and the partner (Z). The IDs for the couples are also uniquely identified by the prefix of P or Z which denotes the patient or the partner, respectively. Inside of these folders, are sub-folders which have data for each day for that participant. Within those folders, the data is grouped per hour in each folder, and nested inside those folders can be more folders which represent the data taken in that hour. For the smartwatch directory, the nested folders hold data on the smartwatch sensors such as audio, accelerometer, heart rate, light and gyroscope. Whereas for the smartphone directory, the folders are nested in the same way, but the folders hold data on the raw arousal and valence self-report emotions from the Affective Slider on the smartphone app. Apart from this, the transcripts for the respective audios are stored in the transcripts folder, which currently does not group data into folders in the nested structure aforementioned.

There is a meta-data information file which gives information on the attributes of each of the audio files. Such as, the name of the audio file, whether the audio file has speech, whether both male and female couples are speaking, does it have a matching self-report, does it have the respective transcript, and more. The full list of attributes the DyMand data describes is available in the Appendix. The meta-data is used to filter audio samples that have speech from both the male and female participant. In addition, the condition was applied to have samples that have both arousal and valence self-ratings, and respective transcripts. The meta-data contains information on 1,017 audio data. After filtering for these conditions, there are 380 audio files. These conditions were applied to have data from couples from a dyadic context, and for the data to have all of the modalities of acoustic, linguistic, and sensor. After filtering the samples in the meta-data, the audio filenames were used and transformed to create the paths to the respective folders and sub-folders where the data is located. Once the data is located, Python was used to read from each file, and append to a larger Pandas DataFrame object.

While reading each file, there were certain data pre-processing steps applied. First, the sensor data that had less than 50% of contact with the skin during the field study was discarded. This information was obtained from sensor data, which has a column that describes the percentage estimate of contact. Second, the outliers from the sensor data were removed similar to the K-EmoCon pre-processing methodology. For the heart rate sensor data, values that were less than 30 and greater than 300 were removed. While for the accelerometer data, values that were more than 1 standard deviation away from the mean were removed. The heart rate sensor data was resampled to match the sampling rate of 1 Hz, and the accelerometer sensor data was re-sampled to match the sampling rate of 20 Hz. In case the sampling rate of the raw data was less than the sampling rate specified, those samples are interpolated to match the sampling rate. As the sampling rate for the emotions in 5 minutes in DyMand, and the total duration of the sensor data for each data collection sample is also 5 minutes, the data does not need to be further grouped into non-overlapping time windows as was the case in the K-EmoCon dataset as the DyMand data is already grouped into 5 minute periods. The audio data is pre-processed by creating the paths from the filenames to the respective locations of the audio files. In addition, each of the audio files have a separate folder which contains the manual speaker diarization for the audio files. These give information about the speech segments start and stop times for each of the male and female participants speech in the audio file. The paths for these annotations are also created. Due to not all speaker diarization of audio data being present during the experiment, the number of samples for the acoustic and linguistic data used (358) are less than the sensor data (380).

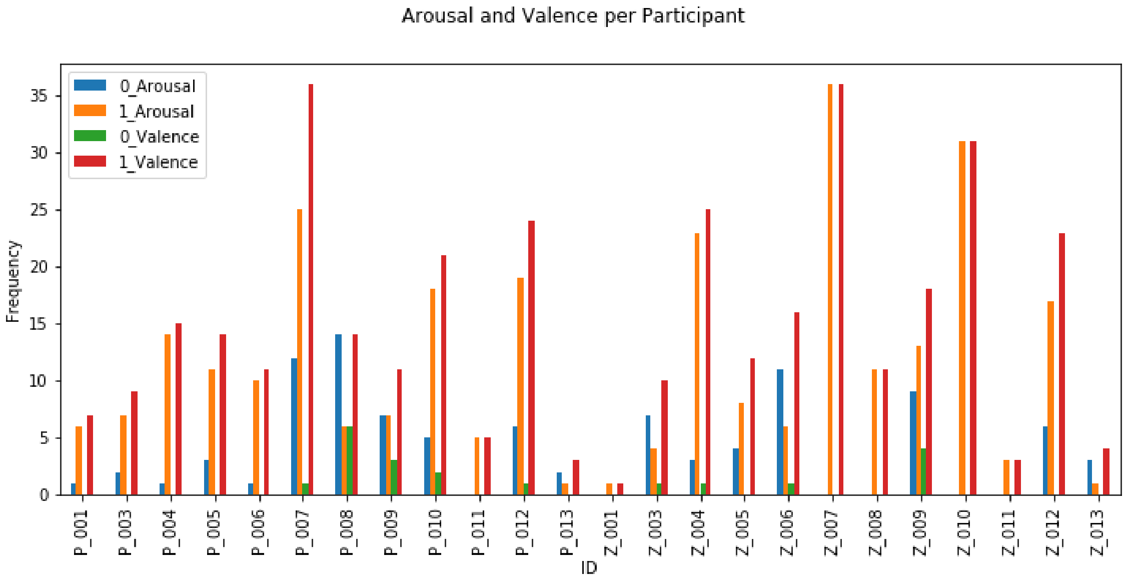


Figure 3.11: DyMand Arousal and Valence per participant

Similar to the pre-trained datasets, the DyMand dataset also has class imbalances in the emotions, as expected for the in-the-wild datasets. This is visualized in Figure 3.11. First, the participants who do not have either a low arousal, negative valence, or high arousal, and positive valence class are removed from the arousal or valence datasets for DyMand to prevent the model from being biased on the training or the evaluation. For the arousal dataset, the following participant with IDs are removed: [P_011, Z_001, Z_007, Z_008, Z_010, Z_011]. On the other hand, for the valence dataset, the following participants with IDs are removed: [P_001, P_003, P_004, P_005, P_006, P_011, P_013, Z_001, Z_005, Z_007, Z_008, Z_010, Z_011, Z_012, Z_013]. As all of the modalities in the DyMand data have the same emotion labels, these participants were removed from the arousal and valence datasets respectively for all of the modalities.

3.3.2 Feature extraction

For the sensor data modalities (accelerometer and heart rate), the 10 statistical features were extracted from the raw data. These are the same features as extraction above for the K-EmoCon dataset. Whereas, for the acoustic data, eGeMAPS feature set were used to extract the 88 features similar to VAM. Also, for the linguistic feature extraction, S-BERT was used to extract the 768-dimensional sentence embeddings.

Sensor data

The sensor data (accelerometer and heart rate) was read from the respective folders, applied preprocessing steps aforementioned, and the features were extracted for each file. These features include the 10 statistical feature sets identified in the K-EmoCon dataset. These features were also extracted using the Pandas library and the SciPy library. The features from each file, and from each participant were collected to create a complete dataset in the form of a Pandas DataFrame object.

Acoustic and linguistic data

The acoustic data feature extraction is more challenging than the VAM data acoustic feature extraction. First, the DyMand acoustic data is in audio segments of 5 minutes, which not only contain speech, but also pauses, other speakers that are not the participant, and other noises. Therefore, only the audio segments where the participant in focus is speaking needs to be extracted. The audio annotations are located in another folder structure, hence the paths for those were first found and then simultaneously, a Python script was written to align the information from the audio and its respective annotations file. The annotations file is a text file which contains the start and stop times of the male or the female participant. The audio segments from these time periods were first extracted. Afterwards, if an audio file contains one or more speech segments, then these are combined together. After combining the speech segments, the eGeMAPS feature set is used to extract the 88 features similar to the VAM data feature extraction.

Another Python script was written to extract the features from the linguistic data. The transcripts of the respective audio files are located in different folders than the audio file, which is why it could not processes the transcripts in parallel. These transcripts are formatted in Microsoft Word Docx and are not read typically as other text files. The Python package docx is used to read and modify contents of Microsoft Word documents, specifically, with the Docx extension. Additionally, the transcripts contain symbols such as "//" or "XX" and "XY" which are placeholders written by the manual transcribers who do not understand what is being said. Before linguistic feature extraction, these are removed by replacing these symbols with empty strings. Additionally, other structural information such as paragraph spacing, indentation, are removed. After filtering for these conditions, the data is saved in an array. All of the data is collected and transformed, before being saved as a Pandas DataFrame Object. This DataFrame is then passed onto the S-BERT model which creates the 768-dimensional sentence embeddings for each of the samples.

3.3.3 Early fusion approach

The DyMand dataset has 4 modalities (acoustic, linguistic, accelerometer, and heart rate) that can be combined together for early and late fusion. However, the pre-trained models only have either acoustic and linguistic, or accelerometer and heart rate. Therefore, not all 4 of DyMand's potential modalities are combined since the effect of transfer learning will not be studied.

In early fusion, the acoustic and linguistic features are combined to create a 856-dimensional dataset. The dataset is combined using the filename which also indicates the day and the hour of the recording from both the modalities. Separately, the accelerometer and heart rate sensor modalities are combined to create the early fusion sensor dataset with 20-dimensional features. For both the early fusion datasets, they are trained for both arousal and valence classification. Similar to previous early fusion approaches, an SVC model and a 3-hidden layer neural network architecture is used. The SVC is hyperparameter tuned using the GridSearchCV cross-validation, whereas the neural network is hyperparameter tuned using the Keras Tuner's Hyperband class. The best hyperparameters found are used to evaluate the performance using the test set. The results are reported in Section 4.

3.3.4 Late fusion approach

In late fusion, the acoustic, linguistic, accelerometer and heart rate sensor datasets for both arousal and valence are trained separately. There are two models; SVC and 3-hidden layer neural network that are used as the model architecture. Similar to previous approaches, the SVC is hyperparameter tuned using GridSearchCV and the neural network is hyperparameter tuned using Keras Tuner's Hyperband class. The best hyperparameter tuned model is used to evaluate the performance for each modality using the same test set as in early fusion to enable comparability between early fusion and late fusion. The best neural network model found for acoustic and linguistic have their raw predictions averaged. The raw predictions are created by the final prediction layer which uses a sigmoid function to produce values between 0 and 1. Thus, these values from both the modalities are averaged and then binarized (less than or equal to 0.5 is 0 and greater than 0.5 is 1) to create the final prediction and thus fused at the decision level. Similarly, the best neural network models for the accelerometer and heart rate have their raw predictions averaged. The evaluation is done using the using the same test set as early fusion to enable a comparison of methodology on the performance.

3.3.5 Fine-tuning approach

For fine-tuning, the best neural network models from the pre-training datasets need to be saved. This is done for both early and late fusion models found in K-EmoCon and VAM. The models are saved using the Keras save function and they are given the extension as "h5" which stores the model weights, model learning rates, model structure and more in a hierarchical data format.

Early fusion

For early fusion - acoustic and linguistic, the best model from the early fusion VAM dataset is loaded. An experiment is conducted by freezing no layer, freezing the first two layers (layer 1 and dropout 1), freezing the first four layers (layer 1, dropout 1, layer 2, dropout 2), freezing the first 6 layers (layer 1, dropout 1, layer 2, dropout 2, layer 3, dropout 3) and finally freezing all the layers (layer 1, , dropout 1, layer 2, dropout 2, layer 3, dropout 3, and the final prediction layer). By freezing the layer, the layer becomes not trainable, and thus the weights from the pre-trained model remain fixed. By evaluating the experiment in such a way gives an indication of freezing which layers, no layers, or all layers has the most impact on transfer learning. The test set from DyMand acoustic and linguistic data is used to as input to the model from VAM after freezing the aforementioned layers, and thus it evaluates the performance of transfer learning when freezing different layers.

Similarly, for early fusion - accelerometer and heart rate sensor, the best model from the early fusion K-EmoCon dataset is loaded. The exact experiment aforementioned is conducted (by freezing different sets of layers). The performance is evaluated using the test set from the DyMand accelerometer and heart rate sensor data, and is used as input to these models and experiment.

Late fusion

For late fusion - acoustic and linguistic, the best models from the late fusion VAM dataset for acoustic and linguistic datasets are loaded. The experiment is conducted similarly by freezing the different sets of layers. The test set from the DyMand acoustic data is used to evaluate transfer learning on the acoustic model, and the test set from the DyMand linguistic data is used for evaluate transfer learning on the linguistic model. The best transfer learning models from acoustic and linguistic are used again and their predictions for the test set are merged together using the averaging approach (as used before in late fusion) to create the late fusion.

Similarly, for the late fusion - accelerometer and heart rate sensor, the best models for late fusion accelerometer and late fusion heart rate were loaded from the K-EmoCon dataset. The experiment is conducted similar to the aforementioned late fusion approach, where different layers are frozen from training. The test set for each of the respective modalities are used to evaluate the performance of the transfer learning. Finally, the best transfer learning models from the accelerometer and heart rate are used to combine their predictions using the average approach for evaluating the late fusion.

Chapter 4

Results

This section presents the experimental results to investigate the research questions. In the first section, early fusion and late fusion multimodal approaches are compared to evaluate RQ 1 for the pre-training datasets (VAM and K-EmoCon) and on the DyMand dataset. Further, experiments using only unimodals are also presented which enable a comparison between unimodal and multimodal approaches. In the second section of this chapter, transfer learning is applied using the pre-trained datasets that have the same deep neural network architecture as without transfer learning. This section evaluates RQ 2, which aims to investigate the performance of transfer learning by fine-tuning on the DyMand dataset. The results presented in this chapter are from deep neural networks, which are used for transfer learning. Similar results are reported in the Appendix for early fusion approaches for all the datasets using an SVM machine learning model. This model does not contribute to the research questions as it cannot be used apply transfer learning by fine-tuning. Hence the results are provided in the Appendix for baseline comparison purposes. These results are explained and connections are drawn between multimodal fusion and transfer learning approaches. A further in-depth discussion of the results is presented in chapter 5.

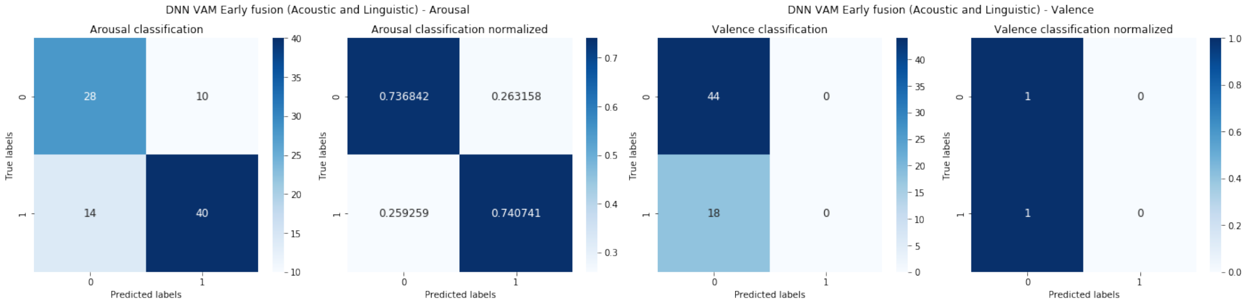
4.1 Multi-modal Fusion without Transfer Learning

4.1.1 VAM

VAM (DNN model)		
Modality	Arousal (UAR %)	Valence (UAR %)
Acoustic	73.3	49.3
Linguistic	52.9	53.6
Early fusion (Acoustic and Linguistic)	73.9	50
Late fusion (Acoustic and Linguistic)	58.0	50

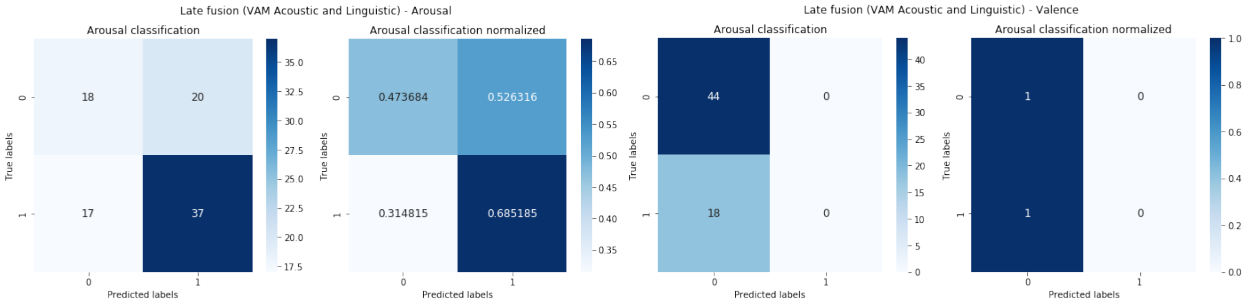
Table 4.1: Experimental results on the VAM dataset for comparing across uni-modal and multimodal fusion techniques.

Among the unimodals, the acoustic modality performs significantly better at classifying arousal than the linguistic modality, as observed in Table 4.1. However, for classifying valence, the linguistic modality outperforms the acoustic modality by a slight margin. These results can give



(a) VAM Early Fusion DNN Arousal (b) VAM Early Fusion DNN Valence

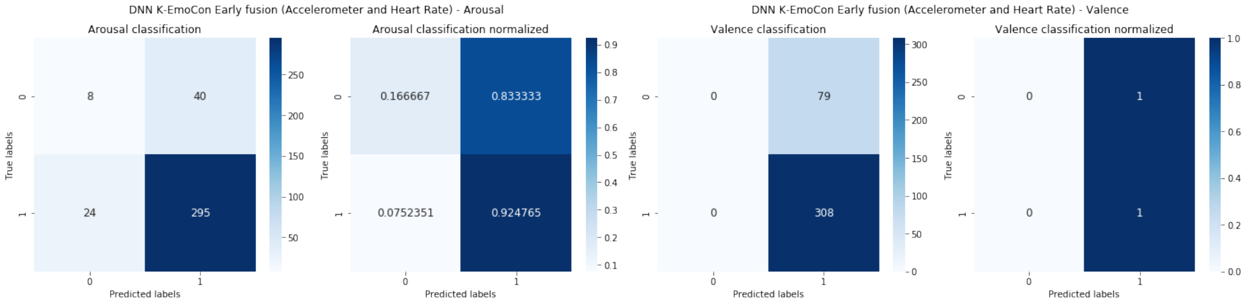
Figure 4.1: VAM Early Fusion DNN for Arousal and Valence



(a) VAM Late Fusion DNN Arousal (b) VAM Late Fusion DNN Valence

Figure 4.2: VAM Late Fusion DNN for Arousal and Valence

an insight on the performance of early fusion and late fusion approaches. The early fusion approach combines the acoustic and linguistic modalities at feature level. The performance of the early fusion model for VAM is closely related to the performance of the acoustic unimodal, with only a small gain in classification. This could indicate the early fusion model leveraged features from the acoustic modality more than the linguistic modality. The late fusion approach performs relatively poorly when compared to the early fusion approach for VAM. Although the acoustic unimodal performed better than the late fusion approach, it is likely the averaging approach to combine the predictions had resulted in a lower performance due to the linguistic modality. For both early and late fusion, the model is unable to discern valence and classifies all samples as negative valence as seen in Figure 4.1b and 4.2b. Overall, early fusion performed better than late fusion for the acoustic and linguistic modality in the VAM dataset. Additionally, multi-modal fusion approaches outperform the unimodals.



(a) K-EmoCon Early Fusion DNN Arousal

(b) K-EmoCon Early Fusion DNN Valence

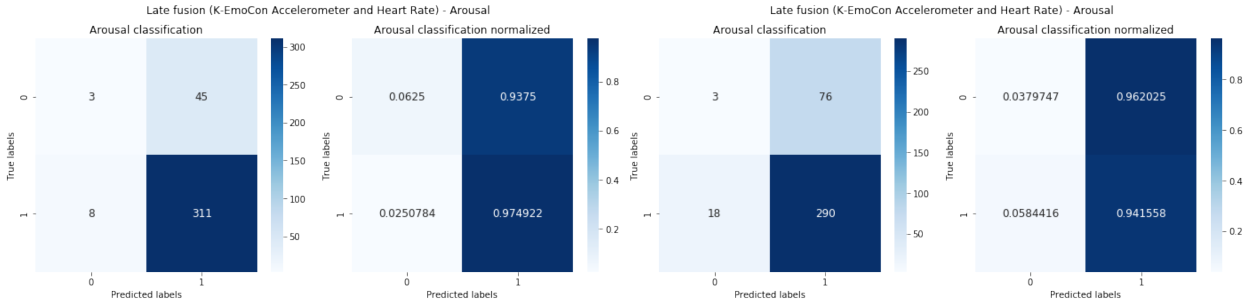
Figure 4.3: K-EmoCon Early Fusion DNN for Arousal and Valence

4.1.2 K-EmoCon

K-EmoCon (DNN model)		
Modality	Arousal (UAR %)	Valence (UAR %)
Accelerometer	51.3	47.9
Heart Rate	55.6	49.7
Early fusion (Accelerometer and Heart Rate)	54.6	50
Late fusion (Accelerometer and Heart Rate)	51.8	48.9

Table 4.2: Experimental results on the K-EmoCon dataset for comparing across uni-modal and multi-modal fusion techniques.

Among the unimodals in K-EmoCon, the heart rate modality outperforms the accelerometer modality by a slight margin in both arousal and valence classification. This result can be used to indicate the performance of the early fusion approach. In early fusion, the accelerometer and heart rate modalities are combined at feature level for the model to leverage the interdependence of these modalities. The early fusion approach outperforms the unimodal for accelerometer, however, it has a slightly less balanced accuracy (UAR%) than the heart rate modality. For late fusion, each of the modalities are trained separately and the results are combined using averaging. This results in the performance of the late fusion approach to be similar to the accelerometer unimodal, although it uses the raw predictions of the heart rate modality as well to outperform the accelerometer unimodal in both arousal and valence, as seen in Table 4.2. Overall, the early fusion approach outperforms the late fusion approach for this sensor dataset in both arousal and valence dimensions. However, multi-modal fusion is not able to perform better than the heart rate unimodal for the arousal emotion. This will be discussed in chapter 5.



(a) K-EmoCon Late Fusion DNN Arousal

(b) K-EmoCon Late Fusion DNN Valence

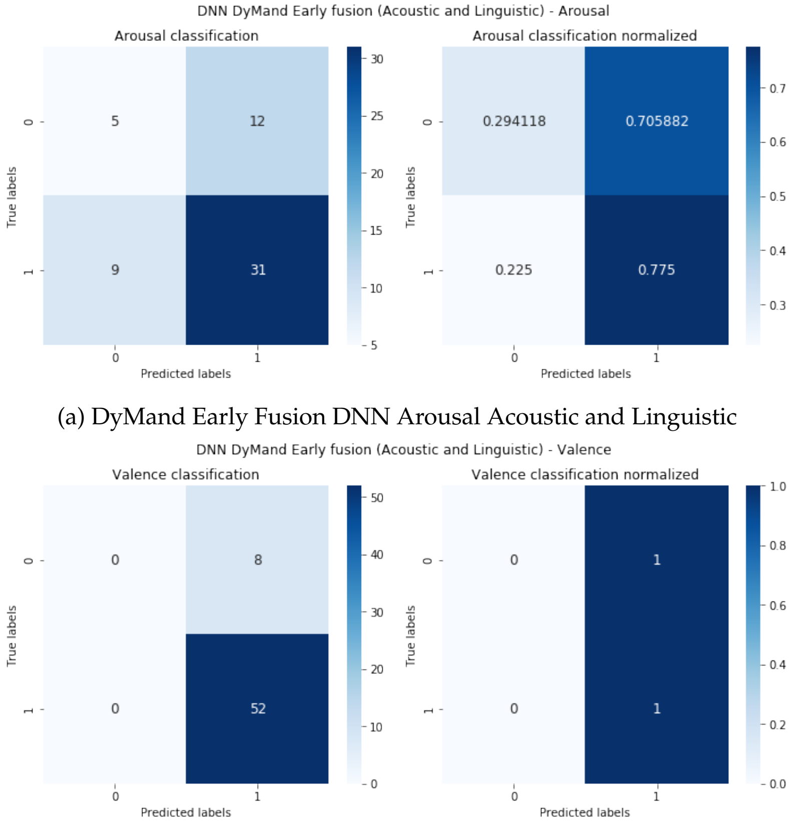
Figure 4.4: K-EmoCon Late Fusion DNN for Arousal and Valence

4.1.3 DyMand

DyMand - without transfer learning (DNN model)		
Modality	Arousal (UAR %)	Valence (UAR %)
Acoustic	55.1	50.0
Linguistic	50.4	49.0
Accelerometer	48.7	50.0
Heart Rate	48.6	50.0
Early fusion (Acoustic and Linguistic)	53.4	50.0
Early fusion (Accelerometer and Heart Rate)	55.4	50.0
Late fusion (Acoustic and Linguistic)	56.7	50.0
Late fusion (Accelerometer and Heart Rate)	50.0	50.0

Table 4.3: Experimental results, without transfer learning, on the DyMand dataset for comparing across uni-modal and multi-modal fusion techniques.

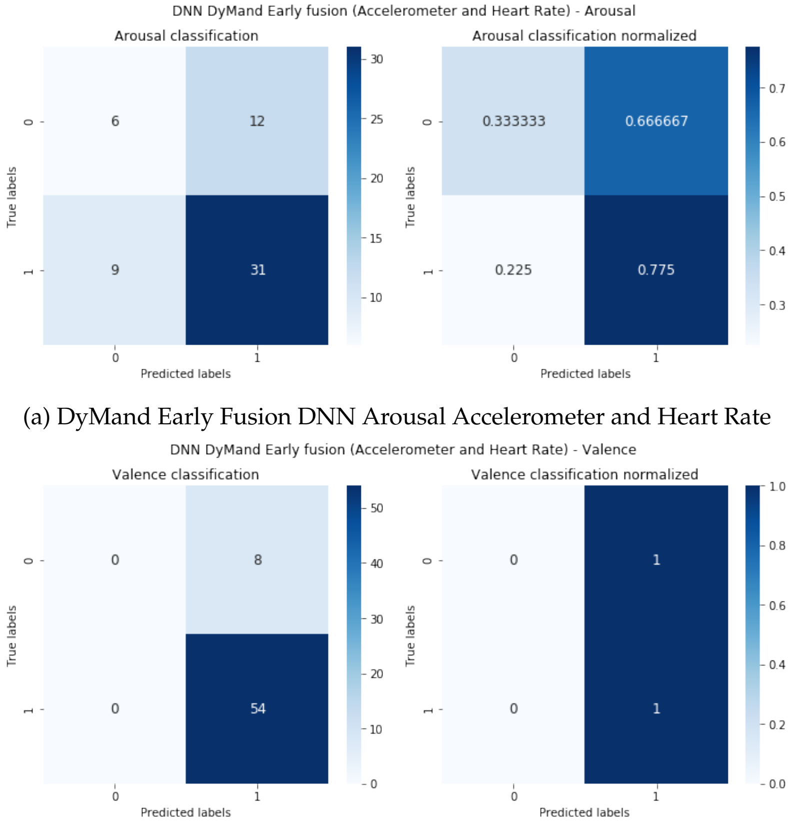
In the DyMand experiments, the acoustic modality outperforms the other unimodals of linguistic, accelerometer and heart rate as seen in table 4.3. These results are similar the VAM dataset, which also showed the acoustic unimodal to outperform the linguistic unimodal. While the performance of the accelerometer and heart rate unimodals are relatively similar. In the DyMand dataset, there are two early fusion and two late fusion approaches applied. This is due to the pretrained models of VAM (having only acoustic and linguistic modalities) and K-EmoCon (having only accelerometer and heart rate modalities), thus transfer learning can only occur for these multimodal fusion approaches. The early fusion approach for acoustic and linguistic sensor data performs relatively poorly than the early fusion approach for the accelerometer and heart rate sensor. While in the unimodals the acoustic and linguistic modalities outperformed the accelerometer and heart rate modalities, in early fusion, this change is not reflected. It is likely the early fusion modal for accelerometer and heart rate was able to leverage the interdependence of these modalities to perform better on the test set. Another pattern is discovered where the late fusion approach outperforms the early fusion approach, however, only for the acoustic and linguistic modalities. The late fusion for the accelerometer and heart rate was not able to distinguish samples between low or high arousal or negative or postive valence, as seen in the confusion matrix in Figure.



(a) DyMand Early Fusion DNN Arousal Acoustic and Linguistic

(b) DyMand Early Fusion DNN Arousal Acoustic and Linguistic

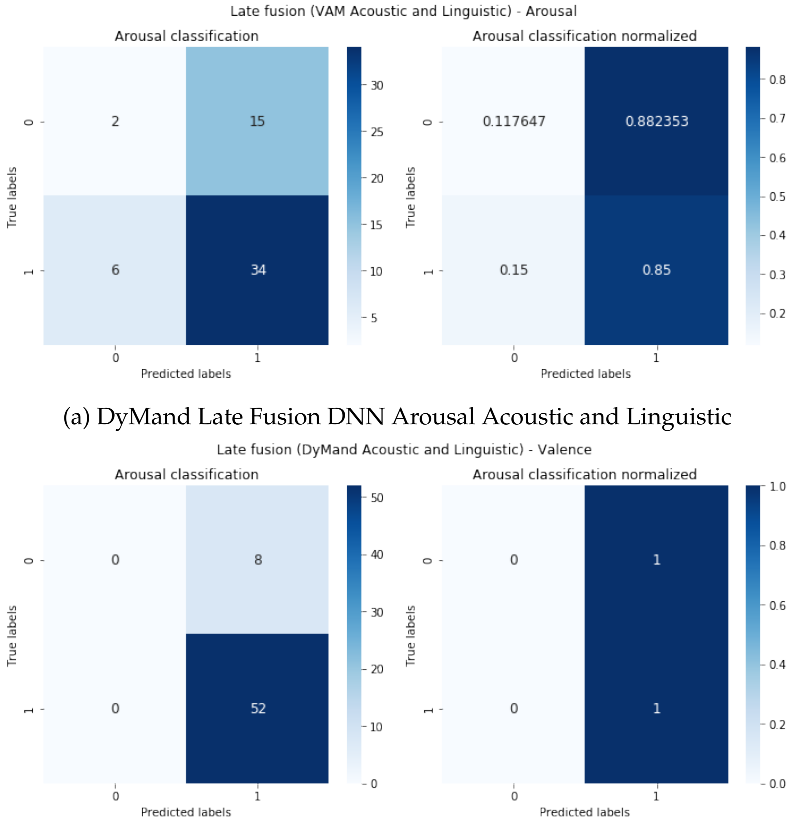
Figure 4.5: DyMand Early Fusion DNN (Acoustic and Linguistic) for Arousal and Valence



(a) DyMand Early Fusion DNN Arousal Accelerometer and Heart Rate

(b) DyMand Early Fusion DNN Valence Accelerometer and Heart Rate

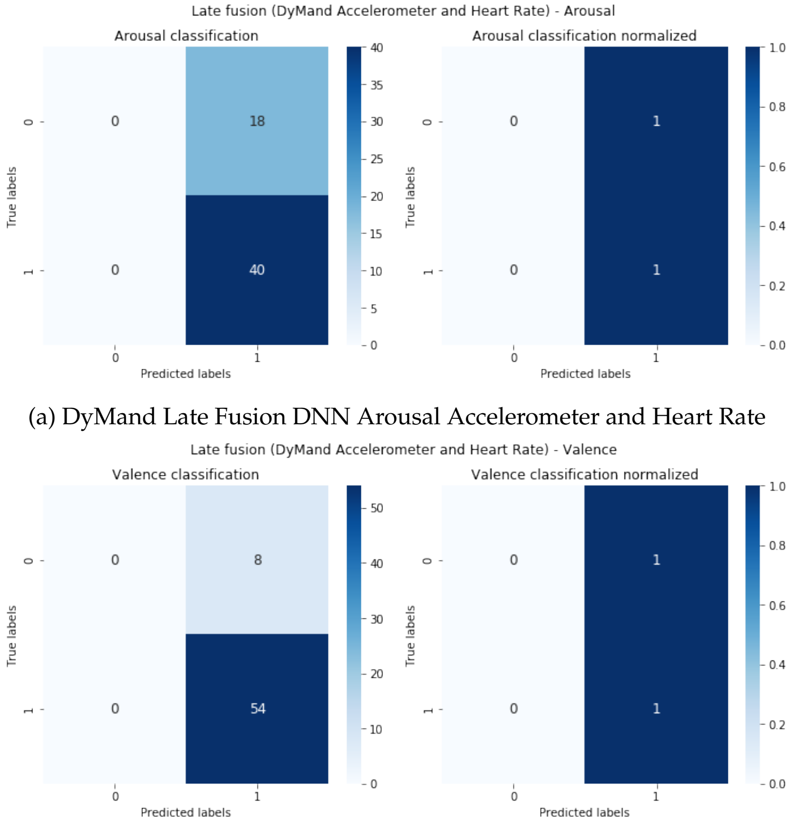
Figure 4.6: DyMand Early Fusion DNN (Accelerometer and Heart Rate) for Arousal and Valence



(a) DyMand Late Fusion DNN Arousal Acoustic and Linguistic

(b) DyMand Late Fusion DNN Valence Acoustic and Linguistic

Figure 4.7: DyMand Late Fusion DNN (Acoustic and Linguistic) for Arousal and Valence



(a) DyMand Late Fusion DNN Arousal Accelerometer and Heart Rate

(b) DyMand Late Fusion DNN Valence Accelerometer and Heart Rate

Figure 4.8: DyMand Late Fusion DNN (Accelerometer and Heart Rate) for Arousal and Valence

4.2 Multi-modal Fusion with Transfer Learning

4.2.1 DyMand

DyMand - with transfer learning (DNN model)		
Modality	Arousal (UAR %)	Valence (UAR %)
Acoustic	56.7	50.0
Linguistic	62.6	50.0
Accelerometer	50.5	50.0
Heart Rate	50.2	50.0
Early fusion (Acoustic and Linguistic)	55.1	57.7
Early fusion (Accelerometer and Heart Rate)	52.3	56.2
Late fusion (Acoustic and Linguistic)	67.7	54.3
Late fusion (Accelerometer and Heart Rate)	52.1	50.0

Table 4.4: Experimental results, with transfer learning, on the DyMand dataset for comparing across uni-modal and multi-modal fusion techniques.

The results for the DyMand unimodal and multimodal approaches with transfer learning can be seen in Table 4.4. The improvement or loss of performance after applying transfer learning is in table 4.5. First, across all unimodal and multimodal approaches except for early fusion accelerometer and heart rate, transfer learning was able to improve the performance of the arousal emotion recognition. A significant improvement in performance is observed for the DyMand linguistic modality and late fusion acoustic and linguistic modality as seen in Table 4.5 for the arousal emotion. Whereas, for valence, the most significant improvement is observed in the early fusion acoustic and linguistic modality, followed by the early fusion accelerometer and heart rate modality.

With transfer learning, early fusion and late fusion approaches had the most significant changes in both arousal and valence emotions. The effect of transfer learning is more observable in the late fusion approaches for the arousal emotions. On the other hand, transfer learning influenced more early fusion valence emotion recognition. A discussion of these results can be found in Chapter 5.

After applying transfer learning on the DyMand dataset, late fusion is observed to perform better than early fusion for the arousal emotions across all modalities. Whereas, early fusion still performs better than late fusion for the valence emotions across all modalities. It is also observed that similar to the VAM dataset, the acoustic and linguistic modalities still outperform the accelerometer and heart rate modalities in all multimodal fusion approaches and unimodal approaches. This indicates that acoustic and linguistic modalities are more suitable for emotion recognition for the DyMand dataset. For recognizing arousal emotions, early fusion is a better approach. Whereas, for recognizing valence emotions, late fusion is a better approach. Overall, transfer learning was able to improve all modalities, multimodal fusion and unimodals except for early fusion accelerometer and heart rate. This indicates that, similar to related work, transfer learning is able to improve classification accuracy on emotion recognition tasks, even with couples emotion recognition in the-the-wild.

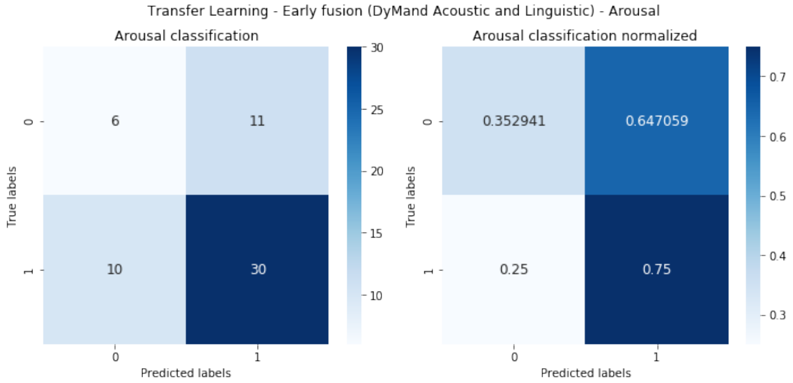
The transfer learning by fine-tuning approach is detailed in Table 4.6. The table indicates the layers that were frozen (i.e, the layers whose weights were not trainable) to obtain the performance of transfer learning as seen in Table 4.4 and 4.5. For early fusion and late fusion transfer learning in acoustic and linguistic modalities, only the first or second layers were frozen to achieve the best performance. This is in contrast to the accelerometer and heart rate modalities for both early and late fusion as it required even the later layers to be frozen (up to the final prediction layer). This indicates that for the accelerometer and heart rate, the DyMand dataset was not able to update the weights and relied on the transfer learning model to achieve the improvement in performance or loss in performance. This is likely due to the sensor data being too noisy and limited in samples for the DyMand dataset to recognize emotions without transfer learning. The K-EmoCon dataset the transfer learning model was pre-trained on for the accelerometer and heart rate modalities, had 10 times more samples, which could have helped in transfer learning. For the valence emotions, the acoustic and linguistic modalities achieved the best performance with transfer learning by only freezing one layer in the transfer learning models. While the accelerometer and heart rate sensor required more to freeze more layers more achieving the best performance in the late fusion approach, with the exception in the early fusion approach. The relation to the layers frozen and the improvement in the valence emotion is not conclusive. As seen in Table 4.5, early fusion accelerometer and heart rate, while only freezing one layer, was able to achieve the second best performance improvement for the valence emotions. These results are further discussed in chapter 5.

DyMand - changes with transfer learning (DNN model)		
Modality	Arousal change (UAR %)	Valence change(UAR %)
Acoustic	+1.6	0.0
Linguistic	+ 12.2	+ 1.0
Accelerometer	+ 1.8	0.0
Heart Rate	+ 1.6	0.0
Early fusion (Acoustic and Linguistic)	+ 1.7	+ 7.7
Early fusion (Accelerometer and Heart Rate)	- 3.1	+ 6.2
Late fusion (Acoustic and Linguistic)	+ 11.0	+ 4.3
Late fusion (Accelerometer and Heart Rate)	+ 2.1	0.0

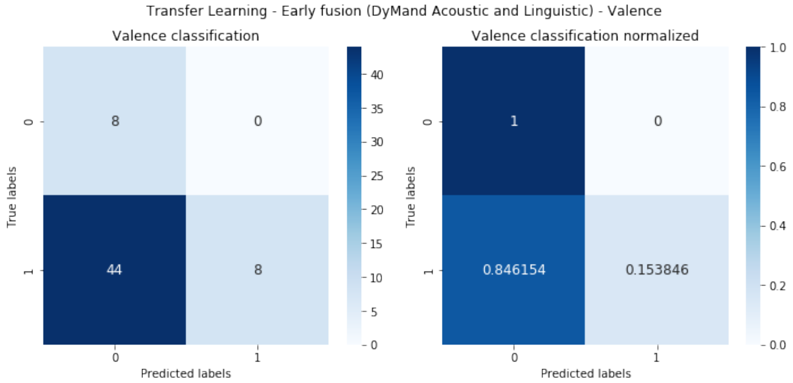
Table 4.5: Experimental results, for changes in UAR, with apply transfer learning on the DyMand dataset. Comparing across uni-modal and multi-modal fusion techniques.

DyMand - transfer learning best layers to freeze (DNN model)		
Modality	Arousal freezing layers	Valence freezing layers
Acoustic	1,2	1
Linguistic	1	1
Accelerometer	1,2,3,4	1,2
Heart Rate	1,2,3	1,2,3
Early fusion (Acoustic and Linguistic)	1,2	1
Early fusion (Accelerometer and Heart Rate)	1,2,3,4	1
Late fusion (Acoustic and Linguistic)	Acoustic.: 1,2; Linguistic,: 1	Acoustic: 1; Linguistic: 1
Late fusion (Accelerometer and Heart Rate)	Acc.: 1,2,3,4; HR: 1,2,3	Acc.: 1,2; HR: 1,2,3

Table 4.6: Experimental results, for the best layers to freeze for transfer learning which lead to the most improvement in UAR across uni-modal and multi-modal fusion techniques.

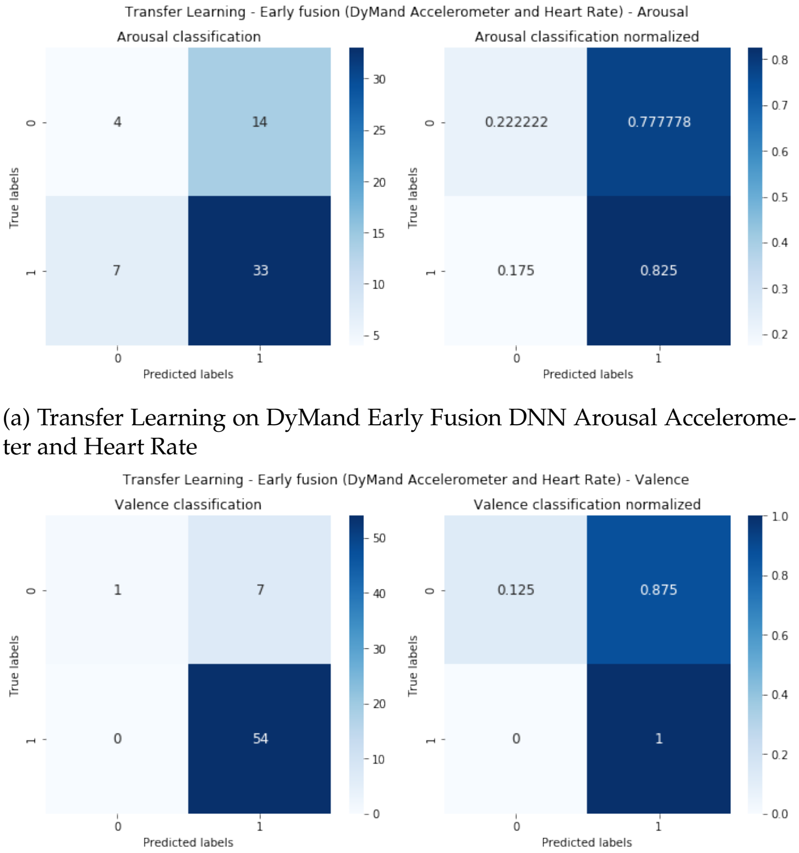


(a) Transfer Learning on DyMand Early Fusion DNN Arousal Acoustic and Linguistic



(b) Transfer Learning on DyMand Early Fusion DNN Valence Acoustic and Linguistic

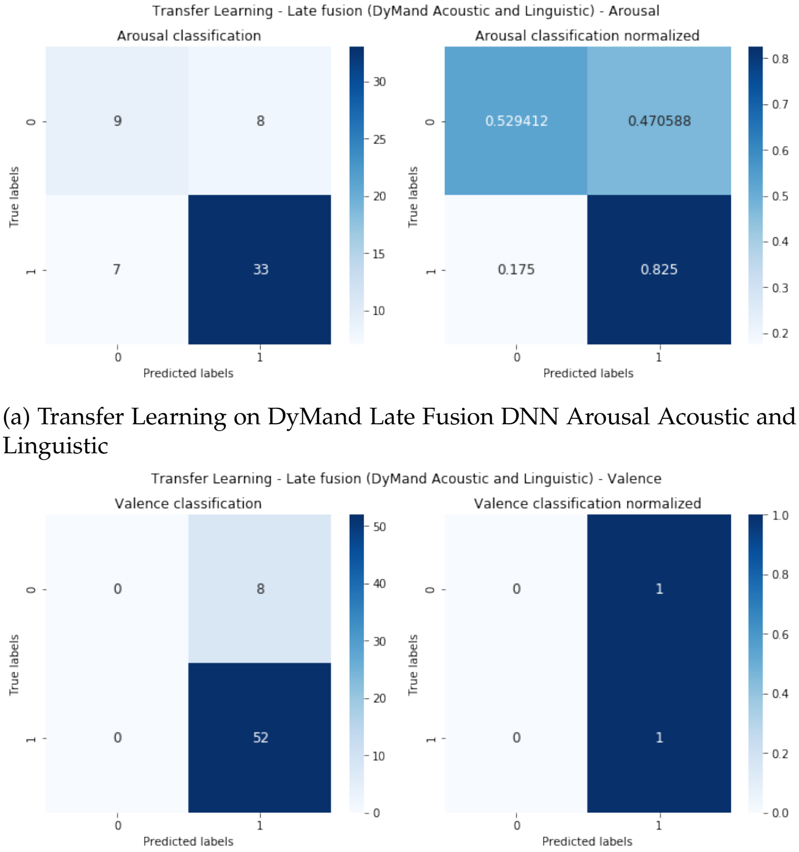
Figure 4.9: Transfer Learning on DyMand Early Fusion DNN (Acoustic and Linguistic) for Arousal and Valence



(a) Transfer Learning on DyMand Early Fusion DNN Arousal Accelerometer and Heart Rate

(b) Transfer Learning on DyMand Early Fusion DNN Valence Accelerometer and Heart Rate

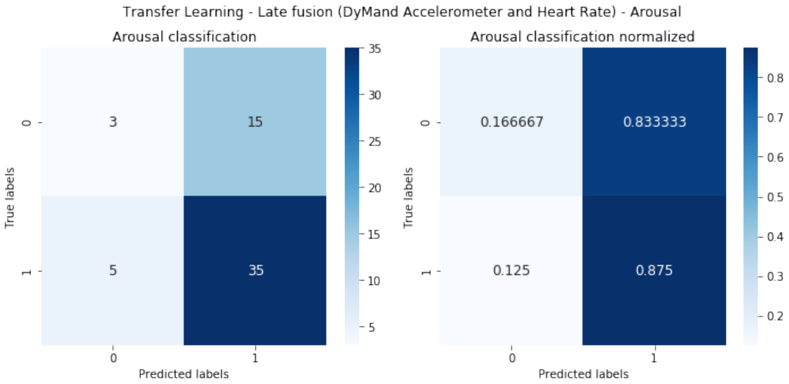
Figure 4.10: Transfer Learning on DyMand Early Fusion DNN (Accelerometer and Heart Rate) for Arousal and Valence



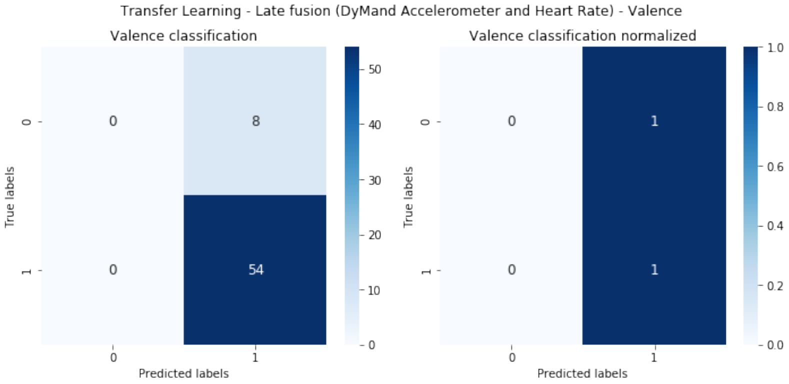
(a) Transfer Learning on DyMand Late Fusion DNN Arousal Acoustic and Linguistic

(b) Transfer Learning on DyMand Late Fusion DNN Valence Acoustic and Linguistic

Figure 4.11: Transfer Learning on DyMand Late Fusion DNN (Acoustic and Linguistic) for Arousal and Valence



(a) Transfer Learning on DyMand Late Fusion DNN Arousal Accelerometer and Heart Rate



(b) Transfer Learning on DyMand Late Fusion DNN Valence Accelerometer and Heart Rate

Figure 4.12: Transfer Learning on DyMand Late Fusion DNN (Accelerometer and Heart Rate) for Arousal and Valence

Chapter 5

Discussion

The first research question (RQ 1) aimed to investigate the performance of early multi-modal fusion with late multimodal fusion. In addition, the contributions of this thesis with respect to RQ 1 aimed to investigate the different modalities the effect of different modalities (unimodal) and multimodal fusion approaches aforementioned and combining which modalities leads to better recognition of emotions. The experimental results from the VAM dataset in Table 4.1 indicates that early fusion outperforms late fusion for the acoustic and linguistic modalities. A similar result can be observed in the DyMand dataset (without transfer learning) that observes early fusion of acoustic and linguistic modalities outperforms late fusion. In addition, we can compare the different modalities in the DyMand dataset, which also indicates that acoustic and linguistic modalities (with or without multimodal fusion) outperforms accelerometer and heart rate data as seen in Table 4.3. For the K-EmoCon dataset, the experimental results also indicate that early fusion performs better than late fusion for accelerometer and heart rate modalities as observed in Table 4.2. Similarly, in the DyMand dataset, these modalities perform better for early fusion rather than late fusion as seen in Table 4.3. A likely reason that early fusion outperforms late fusion in all three datasets could be due to the neural network model being able to use the interdependence of the modalities. Furthermore, in all three datasets (except K-EmoCon) early fusion or late fusion was able to perform better than the unimodals, providing validation that multi-modal fusion can improve emotion recognition on these datasets. In K-EmoCon the exception is due to the heart rate sensor being able to perform marginally better by 1% UAR than early fusion for recognizing arousal. This could likely be that noise from the accelerometer modality in early fusion influenced the early fusion model which could not recognize arousal as well as only using the heart rate modality. However, in all the other datasets, early fusion is able to perform better than late fusion, which indicates the neural network is able to use interdependence of modalities at feature level. Overall, using the three datasets, the experiments found that early fusion outperforms late fusion in recognition emotions.

When comparing the performance of DyMand modalities for arousal and valence, it is observed that the acoustic and linguistic modalities perform better than accelerometer and heart rate modalities when predicting these emotions. This is also observed in the results section for early fusion and late fusion approaches. While the VAM and the K-EmoCon datasets are different from each other, the same model architectures of neural network also performed better on acoustic and linguistic data (VAM) rather than on accelerometer and heart rate data (K-EmoCon). A reason for this could be the difficulty of assessing emotions through these sensor data. The accelerometer sensor measures the acceleration, and thus movement of a participant. While the heart rate sensor measures the heart rate of participant. As these sensors assess the physical characteristics of the body, it could be that they are not related to the emotions expressed by a participant. Rather the acoustics (i.e, how they say) and the linguistics (i.e, what they say) may have a stronger power in describing the emotions that are self-reported by participants. Additionally, the self-reports were filled after the participants had finished the data collection. Another reason why there is a stronger link between the acoustic and linguistic aspects could be that the participants had to recall after 5 minutes of data collection (in DyMand); what they said, and perhaps they remember how they said it. Using this information could or could not have biased their self-reports. Similarly for K-EmoCon, the participants self-reported their emotions after finishing the data collection (debate of approximately 10 minutes) and later viewed their audio-visual footage to self-report their emotions. Although they did not have to recall from memory how they were feeling, they had an audio-visual guide that again described what they said, and how they said it, which could bias their self-reports. Since they could not also see information about the sensor data when rating the emotions, they were biased to use the audio-visual information.

For transfer learning on the DyMand dataset, the results for the DyMand early fusion and late fusion improve considerably. Table 4.5 indicates the performance improvement after transfer learning. The experimental results in Table 4.4 indicates a contrasting observation from earlier, that late fusion outperforms early fusion considerably for the acoustic and linguistic modalities for arousal emotion recognition. However, early fusion for the accelerometer and heart rate modalities still marginally performs better by 0.2% than late fusion for arousal emotions. For valence, the observation that early fusion performs better than late fusion is still valid in the DyMand dataset after transfer learning as well.

As indicated in Table 4.6, for acoustic and linguistic modalities across early and late fusion, the configuration of freezing the first two layers, thus fine-tuning the last layers had the best performance. This observation is for both arousal and valence emotions. A reason for this result could be the DyMand dataset only needs to train on the last layers of the transfer learning as it the initial weights from the pre-training model are able to to extract the lower level details from the transfer learning model. The earlier layers of a neural network typically contain the low-level features of the task [40]. Whereas, the later layers typically are more concentrated on the task of the neural work. For the accelerometer and heart rate sensors, it is observed that they require more layers to be frozen to be able to improve the performance of arousal and valence classification. For instance, in late fusion with the accelerometer and heart rate modalites, the accelerometer modality retains all the layers of the pre-trained model, and the heart rate modality retains upto the first three layers of the pre-trained model. Using this configuration the model is able to improve the performance of the arousal by 2.1% UAR as seen in Table 4.5. Given this observation that the modalities influence the number of fine-tuning layers in transfer learning and not the early fusion or late fusion approaches, could help in extending the transfer learning model.

The limitations in chapter 6 addresses the shortcomings of this work and the threats to validity. It also discusses which limitations can influence the results of multimodal fusion and transfer learning.

Chapter 6

Limitations and Future work

The experimental results could have limitations and threats to validity. Limitations could arise from the simplicity of the models used to evaluate the data. The 3-hidden layer neural network architecture used across the study for comparability, could have been a bottleneck for certain datasets and modalities. As the purpose of the study was to compare early fusion approaches with late fusion approaches and the ability of transfer learning on the in-the-wild DyMand dataset, the neural network architecture had to remain the same across the study. However, a more dense neural network architecture, or other architectures such as convolutions neural networks can be explored.

Second, the feature extraction approach used pre-defined hand-crafted features. The sensor data was limited to 10 statistical features, which may have limited the potential of the models to exploit differences in the data. The same statistical features were used to describe the accelerometer data and the heart rate data, which are two different modalities with their own characteristics. Additionally, the sentence embeddings from S-BERT may have a bias as they are pre-trained on OpenLegalData and News articles and Wikipedia. The formal writing style of these pre-trained datasets could be different from the daily conversational words used in the linguistic data. This can also be a reason for the linguistic data to perform worse than the acoustic data, as reported in the experiments in Chapter 4.

Third, the age of the participants in the VAM (overview in table 3.1) and K-EmoCon (overview in table 3.2) were younger than the ages of the DyMand participants (overview in table 3.3).

Fourth, while the VAM and the DyMand dataset were in German, the K-EmoCon data was recorded in South Korea. As the accelerometer and heart rate modalities were used from K-EmoCon, the cultural differences in South Korea and Switzerland may have prevented the pre-trained model for these modalities to learn similar physiological responses.

Fifth, both the VAM dataset (overview in table 3.1) and K-EmoCon (overview in table 3.2) had an imbalance in the male to female participants. The VAM dataset had 36 female participants and 11 male participants, whereas, K-EmoCon had 12 female participants and 20 male participants. The DyMand dataset had 13 female and 13 male participants. As gender was not taken into account in the emotion recognition models, the gender difference imbalance in the pre-trained datasets can be more bias towards females for the acoustic and linguistic modalities for transfer learning and more biased towards male participants for the accelerometer and heart rate modalities transfer learning in the DyMand dataset.

Sixth, for the VAM dataset, the emotions were rated by external evaluators and for K-EmoCon and DyMand the emotions were rated by the participants themselves. As discussed in the related work, there is a difference between the felt emotions (self-reported) and the perceived emotions (using external evaluators). Hence, transfer learning can be impacted by this methodology choice of assessing emotions since the ground-truth is not by the participants themselves in the VAM dataset. In addition, the emotions were rated at an utterance level in the VAM dataset (per sentence) and the emotions were rated every 5 seconds for the K-EmoCon dataset and similarly, the emotions were rated every 5 minutes for the DyMand dataset. The differences in the granularity of rating emotions can also affect the knowledge transfer as the distribution of data varies across the pre-trained datasets.

There are several future work aspects that can done for continuing this work. First, pre-training from larger datasets can be investigated. The CMU-MOSEAS dataset [78] was not prepared for public release during the course of this thesis and could not be used as an acoustic and linguistic German language dataset. This dataset contains 10,000 hours of German language acoustic and linguistic data, collected in-the-wild from public Youtube videos. Second, while this work only considered in-the-wild datasets for transfer learning, an extension of this work can investigated pre-training using acted datasets. Related work has shown acted datasets to perform better than in-the-wild datasets for emotion recognition tasks [14,21,41,74]. However, this was not considered in this study since acted datasets can also provide limitations, as they are recording in a controlled lab environment with no background noises and external factors influencing participants. Third, several datasets (acted or in-the-wild) can be combined for pre-training. These datasets can be concatenated and used to build a generalizable model. Fourth, selective transfer learning can be investigated. This would entail using models that are pre-trained on data that significantly matches the statistical characteristics of the DyMand dataset. Instead of using the whole data, only selected data is trained and these models are used for transfer learning on the DyMand dataset. Fifth, further pre-trained datasets can be found that also have more modalities such as the ambient light, the gyroscope. Sixth, early fusion and late fusion approaches can also be investigated for combining modalities such as the acoustic modality with the accelerometer, or the linguistic modality with the heart rate modality. This would also entail having a pre-trained dataset with these matching modalities from the same source in order to apply transfer learning.

Chapter 7

Conclusion

This thesis aimed to investigate multi-modal fusion approaches and transfer learning for improving the classification of emotions in the DyMand dataset. Several public datasets were identified and shortlisted for pre-training the transfer learning models. From these datasets, VAM was used for pre-training the acoustic and linguistic emotion classification models, and K-EmoCon was used for pre-training the accelerometer and heart rate sensor emotion classification models. Both early fusion and late fusion was investigated in the pre-training datasets. Similarly, for the DyMand dataset, early fusion and late fusion was investigated, giving insights into which modalities and multi-modal fusion techniques improved emotion classification performance. Transfer learning was investigated by fine-tuning the created pre-trained models on the DyMand dataset. The results indicate for all datasets, including the DyMand dataset, early fusion using handcrafted features outperforms late fusion using the same features. Transfer learning by fine-tuning improves the emotion classification performance on the DyMand data for all modalities and especially for the task of binary arousal classification. With transfer learning on the DyMand dataset, the late fusion approach outperformed the early fusion approach on all modalities for both arousal and valence emotion classification. Overall, the results obtained from multi-modal fusion approaches and transfer learning had shown improved emotion classification performance and addresses the research questions posed in this thesis. For instance, the results validated that transfer learning improves emotion recognition on the DyMand dataset. Additionally, the results validated that early fusion outperforms late fusion using three datasets. With transfer learning, the late fusion outperforms the early fusion on the DyMand dataset. However, future work is needed to address the limitations and further improvements to the models may be useful before it can be used as an automatic emotion recognition model.

Bibliography

[1] ALSKAFI, F. A., KHANDOKER, A. H., AND JELINEK, H. F. A comparative study of arousal and valence dimensional variations for emotion recognition using peripheral physiological signals acquired from wearable sensors. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (2021), IEEE, pp. 1104–1107.

[2] ANH, V. H., VAN, M. N., HA, B. B., AND QUYET, T. H. A real-time model based support vector machine for emotion recognition through eeg. In *2012 International conference on control, automation and information sciences (ICCAIS)* (2012), IEEE, pp. 191–196.

[3] AUGUST, K. J., ROOK, K. S., FRANKS, M. M., AND PARRIS STEPHENS, M. A. Spouses' involvement in their partners' diabetes management: Associations with spouse stress and perceived marital quality. *Journal of Family Psychology 27*, 5 (2013), 712.

[4] BETELLA, A., AND VERSCHURE, P. F. The affective slider: A digital self-assessment scale for the measurement of human emotions. *PloS one 11*, 2 (2016), e0148037.

[5] BOATENG, G. Towards real-time multimodal emotion recognition among couples. In *Proceedings of the 2020 International Conference on Multimodal Interaction* (2020), pp. 748–753.

[6] BOATENG, G., FLEISCH, E., AND KOWATSCH, T. Emotion recognition among couples: A survey. *arXiv preprint arXiv:2202.08430* (2022).

[7] BOATENG, G., LÜSCHER, J., SCHOLZ, U., AND KOWATSCH, T. Emotion capture among real couples in everyday life. In *1st Momentary Emotion Elicitation & Capture workshop (MEEC 2020, cancelled)* (2020), ETH Zurich, Department of Management, Technology, and Economics.

[8] BOATENG, G., SANTHANAM, P., FLEISCH, E., LÜSCHER, J., PAULY, T., SCHOLZ, U., AND KOWATSCH, T. Development, deployment, and evaluation of dymand–an open-source smartwatch and smartphone system for capturing couples' dyadic interactions in chronic disease management in daily life. *arXiv preprint arXiv:2205.07671* (2022).

[9] BOATENG, G., SELS, L., KUPPENS, P., HILPERT, P., AND KOWATSCH, T. Speech emotion recognition among couples using the peak-end rule and transfer learning. In *Companion Publication of the 2020 International Conference on Multimodal Interaction* (2020), pp. 17–21.

[10] BRADLEY, M. M., AND LANG, P. J. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry 25*, 1 (1994), 49–59.

[11] BUSSO, C., AND NARAYANAN, S. S. The expression and perception of emotions: Comparing assessments of self versus others. In *Ninth annual conference of the international speech communication association* (2008).

[12] CARSTENSEN, L. L., GOTTMAN, J. M., AND LEVENSON, R. W. Emotional behavior in long-term marriage. *Psychology and aging 10*, 1 (1995), 140.

[13] CHEN, Y.-C., CHANG, L.-C., LIU, C.-Y., HO, Y.-F., WENG, S.-C., AND TSAI, T.-I. The roles of social support and health literacy in self-management among patients with chronic kidney disease. *Journal of Nursing Scholarship 50*, 3 (2018), 265–275.

[14] CHENCHAH, F., AND LACHIRI, Z. Speech emotion recognition in acted and spontaneous context. *Procedia Computer Science 39* (2014), 139–145.

[15] CHICCO, D. Siamese neural networks: An overview. *Artificial Neural Networks* (2021), 73–94.

[16] CURTIS, R., GROARKE, A., COUGHLAN, R., AND GSEL, A. The influence of disease severity, perceived stress, social support and coping in patients with chronic illness: a 1 year follow up. *Psychology, Health & Medicine 9*, 4 (2004), 456–475.

[17] DEVLIN, J., CHANG, M.-W., LEE, K., AND TOUTANOVA, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[18] D'MELLO, S. K., AND KORY, J. A review and meta-analysis of multimodal affect detection systems. *ACM computing surveys (CSUR) 47*, 3 (2015), 1–36.

[19] EGGER, M., LEY, M., AND HANKE, S. Emotion recognition from physiological signal analysis: A review. *Electronic Notes in Theoretical Computer Science 343* (2019), 35–55.

[20] EKMAN, P., AND FRIESEN, W. V. Constants across cultures in the face and emotion. *Journal of personality and social psychology 17*, 2 (1971), 124.

[21] EYBEN, F., SCHERER, K. R., SCHULLER, B. W., SUNDBERG, J., ANDRÉ, E., BUSSO, C., DEVILLERS, L. Y., EPPS, J., LAUKKA, P., NARAYANAN, S. S., ET AL. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE transactions on affective computing 7*, 2 (2015), 190–202.

[22] EYBEN, F., WÖLLMER, M., AND SCHULLER, B. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia* (2010), pp. 1459–1462.

[23] FAYEK, H. M., LECH, M., AND CAVEDON, L. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks 92* (2017), 60–68.

[24] FENG, K., AND CHASPARI, T. A review of generalizable transfer learning in automatic emotion recognition. *Frontiers in Computer Science 2* (2020), 9.

[25] FRIJDA, N. H., ET AL. *The emotions*. Cambridge University Press, 1986.

[26] GARCIA-CEJA, E., OSMANI, V., AND MAYORA, O. Automatic stress detection in working environments from smartphones' accelerometer data: a first step. *IEEE journal of biomedical and health informatics 20*, 4 (2015), 1053–1060.

[27] GIDEON, J., KHORRAM, S., ALDENEH, Z., DIMITRIADIS, D., AND PROVOST, E. M. Progressive neural networks for transfer learning in emotion recognition. *arXiv preprint arXiv:1706.03256* (2017).

[28] GOTTMAN, J. M. *What predicts divorce?: The relationship between marital processes and marital outcomes*. Psychology Press, 2014.

[29] GRIMM, M., AND KROSCHEL, K. Evaluation of natural emotions using self assessment manikins. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.* (2005), IEEE, pp. 381–385.

[30] GRIMM, M., KROSCHEL, K., MOWER, E., AND NARAYANAN, S. Primitives-based evaluation and estimation of emotions in speech. *Speech communication 49*, 10-11 (2007), 787–800.

[31] GRIMM, M., KROSCHEL, K., AND NARAYANAN, S. The vera am mittag german audio-visual emotional speech database. In *2008 IEEE international conference on multimedia and expo* (2008), IEEE, pp. 865–868.

[32] GUNES, H., AND PICCARDI, M. Affect recognition from face and body: early fusion vs. late fusion. In *2005 IEEE international conference on systems, man and cybernetics* (2005), vol. 4, IEEE, pp. 3437–3443.

[33] HASHMI, M. A., RIAZ, Q., ZEESHAN, M., SHAHZAD, M., AND FRAZ, M. M. Motion reveal emotions: identifying emotions from human walk using chest mounted smartphone. *IEEE Sensors Journal 20*, 22 (2020), 13511–13522.

[34] HAZARIKA, D., PORIA, S., ZIMMERMANN, R., AND MIHALCEA, R. Conversational transfer learning for emotion recognition. *Information Fusion 65* (2021), 1–12.

[35] ISSA, D., DEMIRCI, M. F., AND YAZICI, A. Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control 59* (2020), 101894.

[36] JIN, Q., LI, C., CHEN, S., AND WU, H. Speech emotion recognition with acoustic and lexical features. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (2015), IEEE, pp. 4749–4753.

[37] KERIG, P. K., AND BAUCOM, D. H. *Couple observational coding systems*. Taylor & Francis, 2004.

[38] KOOLAGUDI, S. G., AND RAO, K. S. Emotion recognition from speech: a review. *International journal of speech technology 15*, 2 (2012), 99–117.

[39] KOSTIS, J. B., MOREYRA, A., AMENDO, M., DI PIETRO, J., COSGROVE, N., AND KUO, P. The effect of age on heart rate in subjects free of heart disease. studies by ambulatory electrocardiography and maximal exercise stress test. *Circulation 65*, 1 (1982), 141–145.

[40] KOZMA, R., ILIN, R., AND SIEGELMANN, H. T. Evolution of abstraction across layers in deep learning neural networks. *Procedia computer science 144* (2018), 203–213.

[41] LI, Q., AND CHASPARI, T. Exploring transfer learning between scripted and spontaneous speech for emotion recognition. In *2019 International Conference on Multimodal Interaction* (2019), pp. 435–439.

[42] LÜSCHER, J., KOWATSCH, T., BOATENG, G., SANTHANAM, P., BODENMANN, G., SCHOLZ, U., ET AL. Social support and common dyadic coping in couples' dyadic management of type ii diabetes: protocol for an ambulatory assessment application. *JMIR research protocols 8*, 10 (2019), e13685.

[43] MILLER, D., AND BROWN, J. L. Marital interactions in the process of dietary change for type 2 diabetes. *Journal of nutrition education and behavior 37*, 5 (2005), 226–234.

[44] MIRSAMADI, S., BARSOUM, E., AND ZHANG, C. Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)* (2017), IEEE, pp. 2227–2231.

[45] MOTTELSON, A., AND HORNBÆK, K. An affect detection technique using mobile commodity sensors in the wild. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2016), pp. 781–792.

[46] N.D. Open sourcing german bert model. https://www.deepset.ai/german-bert. Accessed: 2022-03-10.

[47] NG, H.-W., NGUYEN, V. D., VONIKAKIS, V., AND WINKLER, S. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction* (2015), pp. 443–449.

[48] ORTEGA, J. D., SENOUSSAOUI, M., GRANGER, E., PEDERSOLI, M., CARDINAL, P., AND KOERICH, A. L. Multimodal fusion with deep neural networks for audio-video emotion recognition. *arXiv preprint arXiv:1907.03196* (2019).

[49] PAN, S. J., AND YANG, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering 22*, 10 (2009), 1345–1359.

[50] PARK, C. Y., CHA, N., KANG, S., KIM, A., KHANDOKER, A. H., HADJILEONTIADIS, L., OH, A., JEONG, Y., AND LEE, U. K-emocon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations. *Scientific Data 7*, 1 (2020), 1–16.

[51] PARROTT, W. G. *Emotions in social psychology: Essential readings*. psychology press, 2001.

[52] PICARD, R. W., VYZAS, E., AND HEALEY, J. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE transactions on pattern analysis and machine intelligence 23*, 10 (2001), 1175–1191.

[53] PLUTCHIK, R. The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist 89*, 4 (2001), 344–350.

[54] PORIA, S., CAMBRIA, E., BAJPAI, R., AND HUSSAIN, A. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion 37* (2017), 98–125.

[55] POSNER, J., RUSSELL, J. A., AND PETERSON, B. S. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology 17*, 3 (2005), 715–734.

[56] RAD, G. S., BAKHT, L. A., FEIZI, A., AND MOHEBI, S. Importance of social support in diabetes care. *Journal of education and health promotion 2* (2013).

[57] REIMERS, N., AND GUREVYCH, I. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[58] RUENSUK, M., OH, H., CHEON, E., OAKLEY, I., AND HONG, H. Detecting negative emotions during social media use on smartphones. In *Proceedings of Asian CHI Symposium 2019: Emerging HCI Research Collection* (2019), pp. 73–79.

[59] RUSSELL, J. A. A circumplex model of affect. *Journal of personality and social psychology 39*, 6 (1980), 1161.

[60] SAHOO, S., KUMAR, P., RAMAN, B., AND ROY, P. P. A segment level approach to speech emotion recognition using transfer learning. In *Asian Conference on Pattern Recognition* (2019), Springer, pp. 435–448.

[61] SÁNCHEZ-LOZANO, E., LOPEZ-OTERO, P., DOCIO-FERNANDEZ, L., ARGONES-RÚA, E., AND ALBA-CASTRO, J. L. Audiovisual three-level fusion for continuous estimation of russell's emotion circumplex. In *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge* (2013), pp. 31–40.

[62] SCHERER, K. R. What are emotions? and how can they be measured? *Social science information 44*, 4 (2005), 695–729.

[63] SCHMIDT, P., REISS, A., DUERICHEN, R., MARBERGER, C., AND VAN LAERHOVEN, K. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction* (2018), pp. 400–408.

[64] SCHULLER, B., MÜLLER, R., LANG, M., AND RIGOLL, G. Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensemble. In *Proc. of Interspeech 2005-Proc. Europ. Conf. on Speech Communication and Technology, Lisbon, Portugal* (2005).

[65] SCHULLER, B., RIGOLL, G., AND LANG, M. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In *2004 IEEE international conference on acoustics, speech, and signal processing* (2004), vol. 1, IEEE, pp. I–577.

[66] SCHULLER, B., STEIDL, S., BATLINER, A., HIRSCHBERG, J., BURGOON, J. K., BAIRD, A., ELKINS, A., ZHANG, Y., COUTINHO, E., EVANINI, K., ET AL. The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language. In *17TH Annual Conference of the International Speech Communication Association (Interspeech 2016), Vols 1-5* (2016), pp. 2001–2005.

[67] SCHULLER, B., ZHANG, Z., WENINGER, F., AND RIGOLL, G. Using multiple databases for training in emotion recognition: To unite or to vote? In *Twelfth Annual Conference of the International Speech Communication Association* (2011).

[68] SCHULLER, B. W. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM 61*, 5 (2018), 90–99.

[69] STUHLSATZ, A., MEYER, C., EYBEN, F., ZIELKE, T., MEIER, G., AND SCHULLER, B. Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (2011), IEEE, pp. 5688–5691.

[70] SUN, T.-W. End-to-end speech emotion recognition with gender information. *IEEE Access 8* (2020), 152423–152438.

[71] THOMPSON, E. R. Development and validation of an internationally reliable short-form of the positive and negative affect schedule (panas). *Journal of cross-cultural psychology 38*, 2 (2007), 227–242.

[72] TIAN, L., MOORE, J. D., AND LAI, C. Emotion recognition in spontaneous and acted dialogues. In *2015 international conference on affective computing and intelligent interaction (ACII)* (2015), IEEE, pp. 698–704.

[73] TRUONG, K. P., VAN LEEUWEN, D. A., NEERINCX, M. A., AND JONG, F. Arousal and valence prediction in spontaneous emotional speech: felt versus perceived emotion.

[74] VOGT, T., AND ANDRÉ, E. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In *2005 IEEE International Conference on Multimedia and Expo* (2005), IEEE, pp. 474–477.

[75] WEISS, K., KHOSHGOFTAAR, T. M., AND WANG, D. A survey of transfer learning. *Journal of Big data 3*, 1 (2016), 1–40.

[76] YANG, K., TAG, B., GU, Y., WANG, C., DINGLER, T., WADLEY, G., AND GONCALVES, J. Mobile emotion recognition via multiple physiological signals using convolution-augmented transformer. In *Proceedings of the 2022 International Conference on Multimedia Retrieval* (New York, NY, USA, 2022), ICMR '22, Association for Computing Machinery, p. 562–570.

[77] ZAD, S., HEIDARI, M., JAMES JR, H., AND UZUNER, O. Emotion detection of textual data: An interdisciplinary survey. In *2021 IEEE World AI IoT Congress (AIIoT)* (2021), IEEE, pp. 0255–0261.

[78] ZADEH, A., CAO, Y. S., HESSNER, S., LIANG, P. P., PORIA, S., AND MORENCY, L.-P. Cmumoseas: A multimodal language dataset for spanish, portuguese, german and french. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing* (2020), vol. 2020, NIH Public Access, p. 1801.

[79] ZAVORSKY, G. S. Evidence and possible mechanisms of altered maximum heart rate with endurance training and tapering. *Sports medicine 29*, 1 (2000), 13–26.

[80] ZHANG, B., ESSL, G., AND MOWER PROVOST, E. Automatic recognition of self-reported and perceived emotion: Does joint modeling help? In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (2016), pp. 217–224.

[81] ZHANG, G. Q., AND ZHANG, W. Heart rate, lifespan, and mortality risk. *Ageing research reviews 8*, 1 (2009), 52–60.

Appendix A

Methodology - datasets shortlisted for pre-training

Candidates for Pre-training - Acoustic and Linguistic Emotion Recognition Datasets				
Dataset	Modalities	Emotions	Access type	Stimulus/Environment
EmoDB	Speech data: German	Categorical emotions	Public	Lab
EU - Emotion voice database	Speech data: English, swedish, hebrew	Categorical emotions (20 different)	Public with restrictions	Lab
K-EmoCon	Physiological sensor data, audiovisual footage, Speech data: English	Arousal, Valence and categorical emotions	Public with restrictions	Naturalistic
MELD	Speech and Linguistic data: English	Categorical emotions	Public	Lab/Acted
CMU-MOSEAS	Speech, Linguistic and Visual data: English and German	Categorical emotions	Public with restrictions	Naturalistic
VAM	Speech and Linguistic: German	Activation (arousal), valence, dominance	Public	Naturalistic
RECOLA	Acoustic, Physiological, Visual: French	Arousal and Valence	Public with restrictions	Acted
SEMAINE	Acoustic, Visual: English	Arousal and Valence and Categorical	Public with restrictions	Spontaneous
IEMOCAP	Acoustic, Linguistic, Visual: English	Arousal and Valence and Categorical	Public with restrictions	Lab
RAVDESS	Acoustic and Visual: English	Categorical	Public	Acted
MSP-Improv	Acoustic and Visual: English	Categorical	Public with restrictions	Spontaneous + Lab
SEWA	Acoustic and Visual: English	Arousal and Valence	Public with restrictions	In-the-Wild

Table 7.1: Candidates for pre-training dataset - Acoustic and Linguistic

Candidates for Pre-training - Sensor Data Emotion Recognition Datasets				
Dataset	Modalities	Emotions	Access type	Stimulus/Environment
WESAD	Accelerometer, Blood value pulse, electrodermal activity, temperature	Arousal and Valence, Categorical emotions	Public	Lab
MAHNOB-HCI	Camera, microphone, eye gazer, EEG, ECG, respiration, temperature	Arousal and Valence	Public with restrictions	Lab
DECAF	Brain signals (MEG), EEG, near-infrared facial videos, horizontal Electrooculogram (hEOG),	Arousal, Valence and Dominance	Public with restrictions	Lab
ASCERTAIN	ECG, EDA, EEG, facial activity data	Arousal and Valence and Categorical emotions	Public with restrictions	Lab
DEAP	ECG, EDA, EEG, EMG, EOG, respiratory, temperature, video of face	Arousal and Valence	Public with restrictions	Lab
K-EmoCon	Acceleration, Skin temperature, Heart rate, EDM, EEG	Arousal and Valence and Categorical emotions	Public with restrictions	Naturalistic

Table 7.2: Candidates for pre-training dataset - Sensor Data

SVM models - VAM

Early Fusion - VAM	
Emotion	SVM (UAR %)
Arousal	73.1
Valence	50.0

Table 7.3: Early Fusion on VAM Data; Arousal and Valence

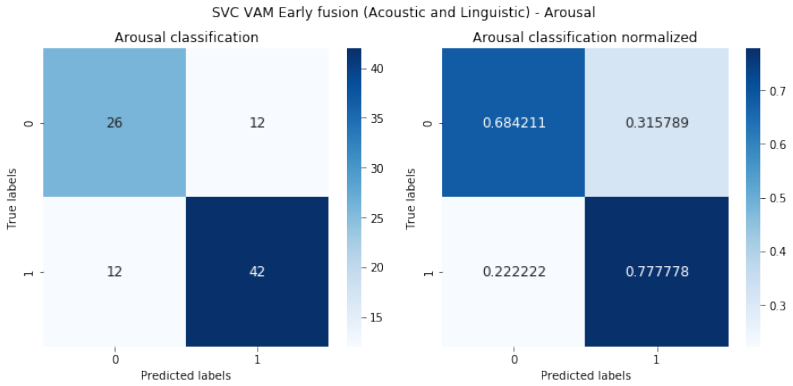


Figure 7.1: VAM Early Fusion SVM Arousal

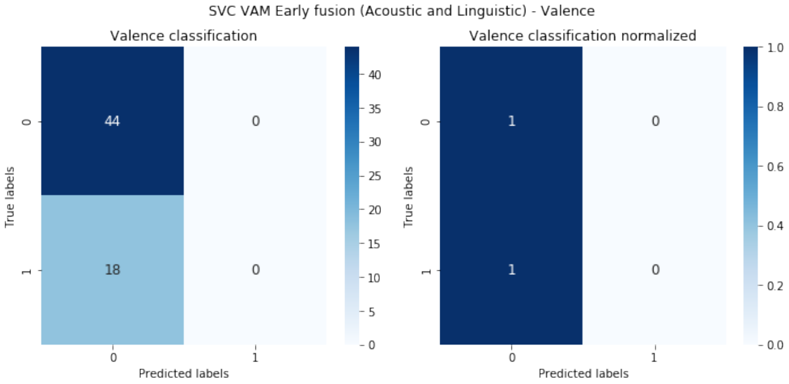


Figure 7.2: VAM Early Fusion SVM Valence

SVM models - K-EmoCon

Early Fusion - K-Emocon	
Emotion	SVM (UAR %)
Arousal	59.7
Valence	43.2

Table 7.4: Early Fusion on K-EmoCon Data; Arousal and Valence

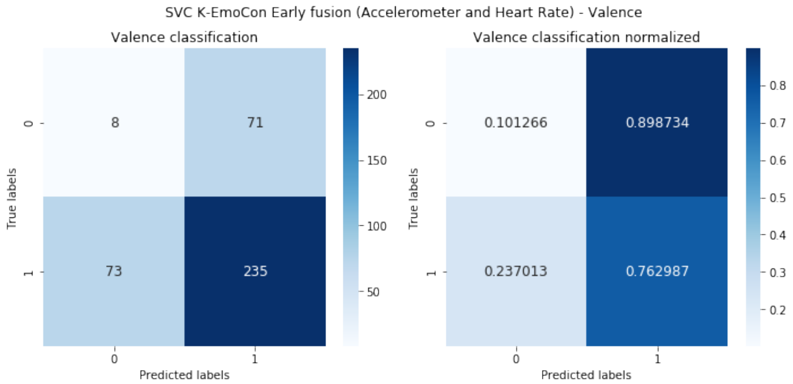


Figure 7.3: K-EmoCon Early Fusion SVM Valence

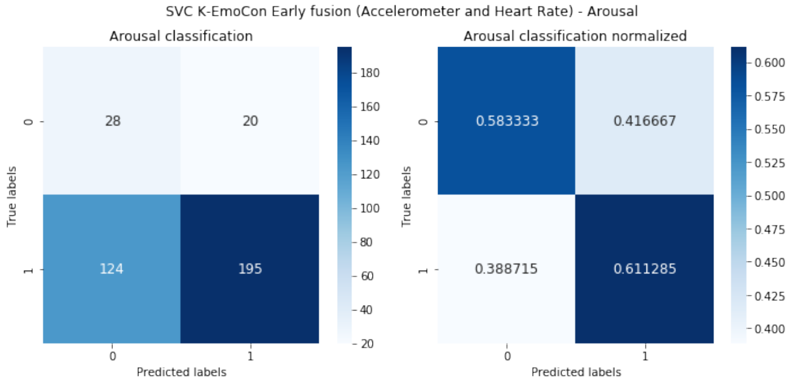


Figure 7.4: K-EmoCon Early Fusion SVM Arousal

SVM models - DyMand

Early Fusion - DyMand (Acoustic and Linguistic)	
Emotion	SVM (UAR %)
Arousal	53.4
Valence	49.5

Table 7.5: Early Fusion on DyMand Data (Acoustic and Linguistic) ; Arousal and Valence

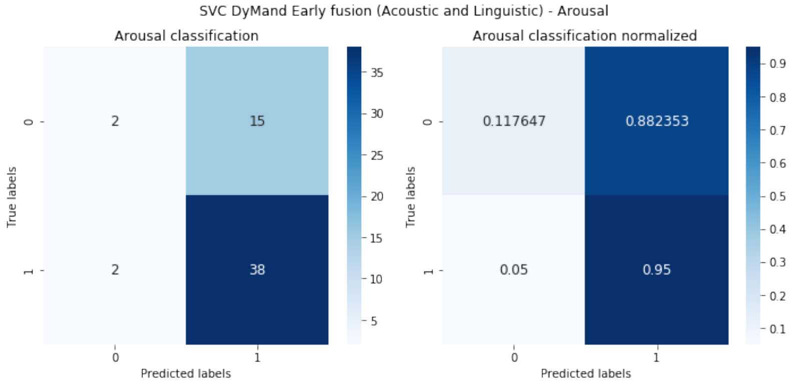


Figure 7.5: DyMand Early Fusion SVM Arousal Acoustic and Linguistic

Early Fusion - DyMand (Accelerometer and Heart Rate)	
Emotion	SVM (UAR %)
Arousal	56.4
Valence	65.9

Table 7.6: Early Fusion on DyMand Data (Accelerometer and Heart Rate) ; Arousal and Valence

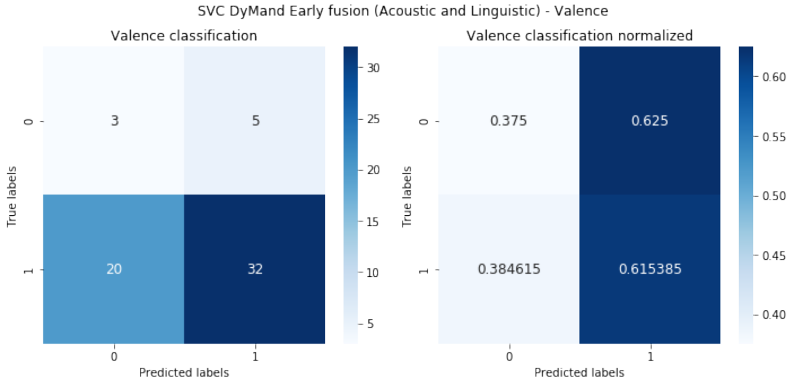


Figure 7.6: DyMand Early Fusion SVM Valence Acoustic and Linguistic

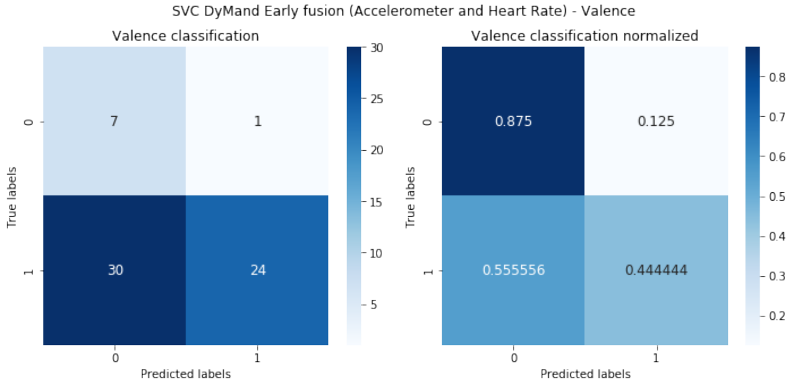


Figure 7.7: DyMand Early Fusion SVM Valence Accelerometer and Heart Rate

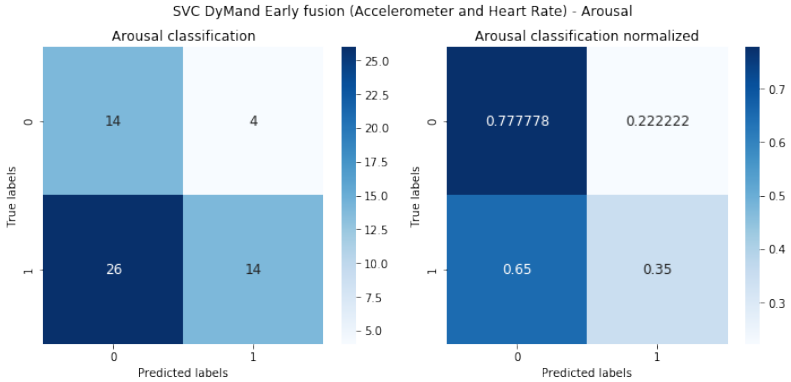


Figure 7.8: DyMand Early Fusion SVM Arousal Accelerometer and Heart Rate