DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

The Role of Physiological Signals in Multimodal Emotion Recognition Solutions in the Era of Autonomous Driving

Simge Özcan

DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master′s Thesis in Informatics

The Role of Physiological Signals in Multimodal Emotion Recognition Solutions in the Era of Autonomous Driving

Die Bedeutung von Physiologischen Signalen in Lösungen zur multimodalen Emotionserkennung in Zeiten des autonomem Fahrens

Author:	Simge Özcan
Supervisor:	Prof. Dr. Alois Christian Knoll
Advisor:	Sina Shafaei, M.Sc.
Submission Date:	14.07.2021

I confirm that this master's thesis in informatics is my own work and I have documented all sources and material used.

Munich, 14.07.2021 Simge Özcan

Acknowledgments

First and foremost I would like to thank my thesis advisor Sina Shafaei for his motivation, support, and constructive working atmosphere. His guidance helped me throughout the research and writing of this thesis.

I owe my deepest gratitude to my family for their unconditional love and support throughout my educational career. This journey would not have been possible without them, and I dedicate this milestone to them.

Lastly, I am grateful to my friends and colleagues for their never ending support and for keeping me sane during this journey.

Abstract

Emotions are complex reaction patterns involving observational, behavioral, and physiological elements. Affect and emotion are fundamental components of a human being, and they play an important role in everyday lives, for example, in computer-related interaction. With the rapid development of autonomous systems and artificial intelligence, more elaborate human-machine interfaces are becoming prominent. In particular, it is crucial for an autonomous system to understand and react to human emotions. One such application is an intelligent assistant that can react to the driver's negative emotional states and adjust the driving behavior according to the driver's emotional state. This thesis aims to evaluate and study the overall role and effects of the physiological signals in an in-vehicle multimodal emotion recognition system. To this end, we implemented three end-to-end deep learning architectures specialized for time-series classification and compared the experimental results with the raw physiological signals. The best performing architecture, Spectro-Temporal Residual Network, reached the accuracy of 66% on classifying the anger, sadness, contempt, and fear emotional states. Later, this network was fused into a multimodal emotion recognition model. By fusing physiological modality with behavioral and facial modalities, we were able to improve the accuracy of the multimodal system by 23%. We implemented a preprocessing pipeline for the utilized database, which can be reused in future research in this field.

Contents

1 Introduction

1.1 Introduction and Motivation

Emotions are known to play a vital role in our daily lives, such as interpersonal communication, decision making, motivation, or driving. A significant part of our lives is occupied by daily commuting to the workplace and back home, which often leads to emotional states like anger[2] or anxiety[3]. These emotional states are triggered by the feeling of lack of control, delays, and potential or occurred accidents. People who frequently commute as a part of their jobs, such as taxi drivers, commonly encounter these triggers.[4] Having too many negative emotions resulting from these triggers may negatively impact driving performance and overall well-being. Therefore, vehicles that can monitor and react to the emotional status of the driver and the passengers are vital for not only improving road safety but also the mental health of the drivers.[5] Furthermore, the recent advancements in the field of computer science enabled machines and humans to engage in more dynamic ways. In order to improve the quality of interaction, it is vital that machines have an understanding of human emotional states.

Affective Computing is a field of study which covers technical aspects of emotional state handling. It is a multidisciplinary field that focuses on better understanding how humans recognize, interpret and simulate emotional states. Figure 1.1 depicts a high-level pipeline of a generic affect recognition application. Affective Computing has applications in many areas such as healthcare, automotive, and education. We focus particularly on autonomous driving and specifically to improve safety and performance through an automated driving assistance system (ADAS).

In autonomous driving research, developing a more intuitive and more efficient connection between the driver and the vehicle is becoming a priority for autonomous vehicles[6] and has many applications. Improving safety is one of the most important applications, which can be done by monitoring the emotional states of the driver or the passengers and accordingly adjusting the driving behavior or driving environment. The system can be capable of warning the user if they are sleepy, unconscious, or unhealthy to drive or lowering the speed, or stopping the car if necessary, which would lead to a more safe and secure driving experience. The importance of emotions and in-cabin comfort are also one of the valid reasons to integrate human emotions with

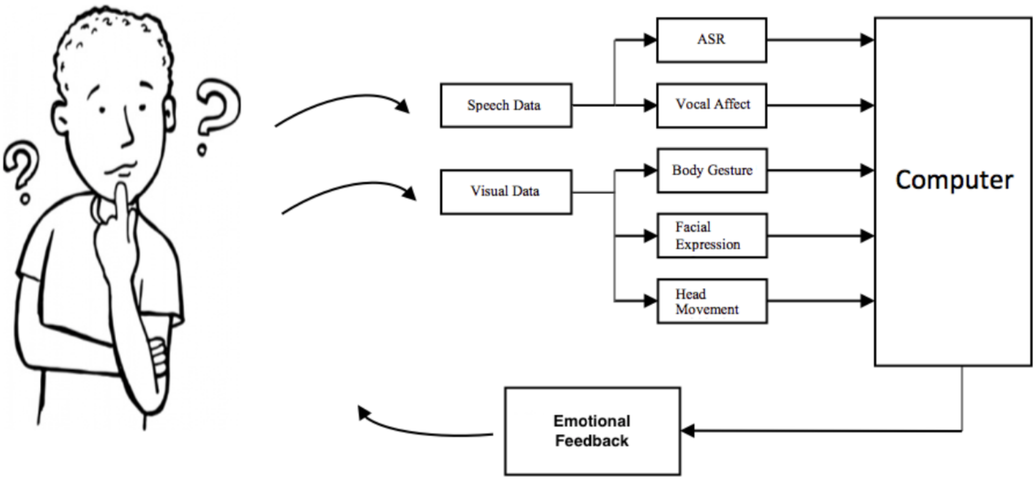

Figure 1.1: A typical multimodal affect analysis framework[1]

the autonomous vehicle. Human beings are prone to change their attitude based on the emotions felt at that moment. Humans realize these changes and adjust their actions, speeches, and attitudes. Similarly, an autonomous vehicle can also offer different interfaces to provide customized settings or take control of the tertiary driving tasks such as air-conditioner, seat heater, and radio based on the emotions it detects. In this way, driving pleasure, comfort, and quality can increase and be a unique to drivers and passengers.[7] Comfort is also an important issue that can be evaluated and adjusted by human emotions.

Several techniques are used to recognize the emotional state of an individual. The most common techniques used in autonomous driving are analysis of facial expressions, voice recognition, driver behavior, eye tracking, and analysis of physiological signals.[8] All these techniques depend on the data collected from the driver of the car. This data is then processed by a Machine Learning(ML)-based model to predict the right emotion status. Therefore, continuous data collection is required to have a robust emotion recognition system.

For accurate and robust detection of emotions, having a recognizer that processes signals from multiple sensors is vital.[1] For example, a modality that depends purely on the camera signal could easily be deceived when the camera vision is obstructed or when the camera is malfunctioning. Thus, dependence on sensor data from one device is not reliable even though the accuracy could be high. Multimodal recognizers which combine two or more of the individual approaches tackle this problem. The use of several modalities requires a multi-sensor dataset and a multimodal sensor fusion to analyze the information coming from different sensors in a meaningful way.

There is not a clear and fixed definition of emotions. However, they can be defined as positive or negative experiences that can be linked to a particular pattern of physiological activity. Emotions induce different physiological, behavioral, and cognitive changes. Therefore, physiological signals play a crucial role when it comes to emotions. Emotions are not only expressed by visible inputs such as facial expressions or gestures, but also invisible inputs such as heart rate, breathing rate, body temperature, and sweating level. However, it was found to be difficult to correlate emotion changes with a single physiological signal. Establishing a clear relationship between emotion changes and physiological changes in terms of different sensor signals still remains an open research question.[9] Therefore, studying emotion recognition using multiple physiological signals is significant in both research and real applications. Acquiring these signals in a cabin environment is difficult due to the complexity of the hardware.[6] Furthermore, a majority of sensors used to collect physiological signals have an intrusive way of detecting emotions. Therefore, the presence of such tools could also affect the current emotions of the driver. For accurate and reliable collection of data, sensors that use those signals which can be acquired in the least intrusive way and are applicable to a cabin environment are needed.

1.2 Structure of the Thesis

Having section 1.1 in mind, we first implement several Deep Learning (DL)-based approaches to learn from multiple physiological signals to comparatively investigate how well they perform on physiological data that was collected on a real-life-like simulation setup. Then, we investigate the role of physiological signals in a multimodal emotion recognition system. In chapter 2, we introduce a multimodal recognition system that was developed as a baseline model and establish our research questions. In chapter 3, we give an extensive background research and literature review on emotion recognition focusing on physiological signals and the automotive field. In chapter 4, we present how we attempt to tackle the established research questions with proposed three different Convolutional Neural Network (CNN) based architectures and a multimodal fusion system. In chapter 5, we present the details and results of the experiments. Lastly, in chapter 6, we summarize the conclusions and discuss the open questions and future work.

2 Problem Statement

Advancements in the fields of artificial intelligence (AI) and autonomous systems have enabled a combination of technologies to form advanced AI applications, as opposed to AI technology to be a stand-alone system. With the emergence of such systems, human-machine interaction becomes more crucial. Recognition of human emotions is the key to securing a safe interaction between humans and intelligent systems, especially in the context of driverless cars. As mentioned in many papers[10, 11, 12], the best emotional state for safety which has less probability of leading to accidents, is being in a good mood. Happy drivers are more likely to achieve an error-less driving experience[6]. On the other hand, being aggressive and angry affects not only the drivers themselves but also the other traffic participants[13]. Therefore, knowing the driver's emotional state, especially when it is towards the negative side, helps to adapt or improve the situation. Having this in mind, we focus on detecting this kind of abnormal emotion in the vehicle environment.

In order to recognize emotions accurately, it is important to process information from different sensors, just like human beings naturally do by recognizing facial expressions and changes in speech or behavioral patterns. We hypothesize that the emotion recognition problem is a multimodal learning task, and utilizing more sensor modalities would make it more robust and reliable.

Although facial or vocal expressions have been leading in the research of emotion recognition for decades, the idea of recognizing emotions from physiological signals is relatively new. Studies have shown that change of pattern in physiological signals is inevitable and detectable in different types of emotions[14]. This is solely because it is in the control of sympathetic nerves of the Autonomic Nervous System (ANS) of our body.[15] This results in increased heart rate, high respiration rate, an increase in blood pressure, and more such physiological changes. Therefore, we hypothesize that taking physiological signals into account would make a multimodal recognition system more robust and reliable.

However, multimodal fusion is a challenging task on various levels, from acquiring the multimodal data to linking the underlying relationships between different modalities. Especially, processing the physiological signals modality, which considers the in-vehicle environment, is particularly challenging due to environmental uncertainties and noise.

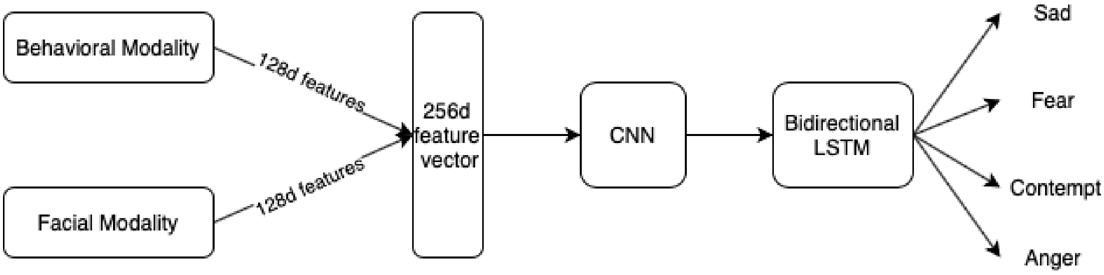

Figure 2.1: Baseline Fusion Model

2.1 Multimodal Approach

Our baseline model is the multimodal fusion approach that had previously been developed in the Chair of AI, Robotics, and Real-Time Systems at TUM. The setup, as depicted in Figure 2.1, includes two convolutional neural networks (CNNs) which extract features from the respective modalities. In particular, the modalities are behavioral and facial video data. Behavioral data includes speed, steering wheel, braking, and acceleration signals from the car. These two CNNs are trained on each respective modality. Once trained, the features are extracted from the corresponding modality data and concatenated into an n-dimensional feature vector. The concatenated feature vector is fed into a final CNN-long short-term memory (LSTM) network that predicts the final sentiment result as either anger, sadness, fear or contempt. Details of this approach and the results are presented later in section 5.4.

This multimodal fusion setup includes only two modalities, namely behavioral and facial. Although it is a multimodal emotion recognition approach, it is only using modalities of human physical signals such as facial expressions and response variables, namely acceleration, speed, steering, and brake force signals. However, the reliability of these modalities cannot be guaranteed, as these are the signals that are relatively easy for people to control. For example, a person might seem to be smiling even if they are in a negative emotional state, and similarly, they might also control the driving behavior. Therefore, to increase the reliability of the overall fusion architecture, we study the emotion recognition from physiological signals, which are also known to be internal signals that are involuntarily activated and cannot be easily controlled by people.

2.2 Research Questions

In this work, we will explore multimodal emotion recognition focusing on physiological signals with a purpose of applicability in an in-vehicle environment. We set our research direction to attempt to answer the following primary research questions:

- Which physiological signals could be used for in-cabin environment in order to recognize emotions in a non-invasive way?
- How to map the features of the physiological signals to fit the negative emotion states?
- How can the input signals be grouped to achieve better performance (w.r.t. accuracy and robustness)?
- How preprocessing phase on data could enhance the deep learning performance when applied before model training?
- What common characteristics do DL architectures that perform well on physiological data have?
- How is the accuracy of DL approaches for emotion recognition from physiological signals, especially when combined with the signal preprocessing techniques before model training?
- How does fusing camera-based and behavioral approaches with physiological signals enhance the performance and what are the robust fusion solutions?

3 Background and Related Work

Since we will investigate the role of physiological signals in multimodal emotion recognition, it is important to understand this field in a broader perspective. In this section we present an overview of the techniques used to study the emotions, and modalities proposed to recognize emotions with a focus on applicability in the autonomous driving domain.

3.1 Emotion Models

In order to classify emotions, it is important to understand how humans perceive and interpret emotions. There are two unique models suggesting perceptions on how emotions are represented by the human mind: the categorical model and the dimensional model. Each model helps to convey the features of human emotions, and assess the real emotional statuses of a person.

The categorical model describes the emotions by using distinct emotional categories, mostly making use of six basic emotion classes namely anger, disgust, fear, joy, sadness, and surprise[16] or domain-specific expressive classes such as boredom, confusion. Although most research in Affective Computing has concentrated on the aforementioned six basic emotions, there are both significant and unrelated emotions in this model.[17]

The dimensional models classify emotions in a continuous fashion and denotes the affects in a dimensional form. These kind of models attempt to conceptualize the human emotions by defining where they lie in dimensional space. Two dimensional models including valence and arousal or three dimensional models including valence, arousal, and intensity are the proposed approaches. The valence dimension defines the positive or negative emotions, while the arousal dimension defines the level of excitement that an emotion depicts. Emotions ranges from unpleasant feelings to pleasant feelings on the valence dimension, representing the sense of happiness. Similarly, emotions ranges from sleepiness or boredom to wild excitement on the arousal dimension. Finally the degree of intensity is defined on the influence dimension such as a sense of control over the emotion. Several dimensional models of emotion have been proposed by the researches. However, there are just a few that remain as the prominent models currently accepted by most[18]. One of the most dominant valence-arousal models are the circumplex model, introduced by psychologist James Russel. Russel suggests that

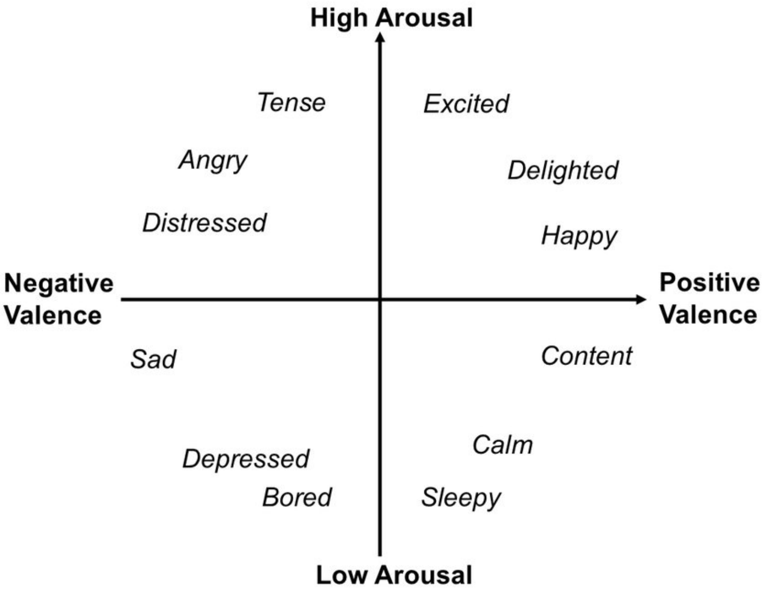

Figure 3.1: Circumplex Model of Emotions

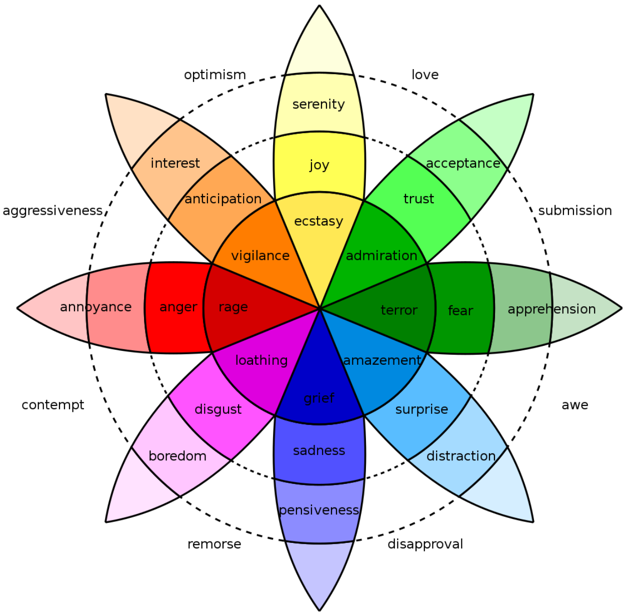

Figure 3.2: Plutchik’s Wheel of Emotions Model

emotion-related terms are organized in a circumplex shape, as depicted in Figure 3.1 and the numerical data are obtained from the corresponding location of the points in two dimensions. Another noteworthy model was suggested by Robert Plutchik, combining both the categorical and dimensional models.[19] Plutchik's model indicates that emotions are based on eight primary emotions. It arranges emotions in circles where inner circles are the primary ones and outer circles more complex and formed by blending the inner circle emotions. The model is called the "emotion wheel," which is depicted in Figure 3.2.

3.2 Emotion Elicitation and Database Collection

After selecting the representation model, next step towards studying emotions is to elicit some emotions in the subject and reliably annotate them. There are many approaches and different kinds of stimuli to provoke emotions. One approach is to ask subjects to recall their specific memory that causes them a determined feeling. Conducting prepared tasks has also been presented as an useful way to elicit emotions. Guided imagery is a particular prepared task approach, in which the subject listens to a narrative that guides them into imagining a situation where a particular emotion would be elicited. This approach has been successfully used to elicite emotions. These studies can either be performed under laboratory conditions in which external influencing factors can be easily controlled, or in a real-life setting, in which the results may be more representative but the data are usually more difficult to analyze.

To effectively recognize the emotional state of drivers, it is important to obtain reliable annotations of emotions that can be used as a gold standard to train and evaluate the models. Self-reports, external annotators, and experimental context are some approaches that have been used in the literature.

Availability of data is vital for Machine Learning-based applications, as these models require high amounts of data for successful learning process. Therefore, it is important to collect these elicited and annotated observations. Next, we will give an overview of the available datasets that capture information of drivers and their approaches to collect the data.

The UTDrive Classical dataset[20] involved 77 subjects performing real-world driving in urban and highway scenarios. Data was collected in the context of cognitive load and driver stress. Collected signals included audio, video, pedal pressure, and distance with the preceding car. UTDrive Portable dataset[20] was also collected by the same authors and it only relied on smartphone sensors to collect the data. In particular, they collected video of the face, audio, car acceleration, and GPS location.

Healey and Picard[21] collected a dataset in the context of stress recognition. Exper-

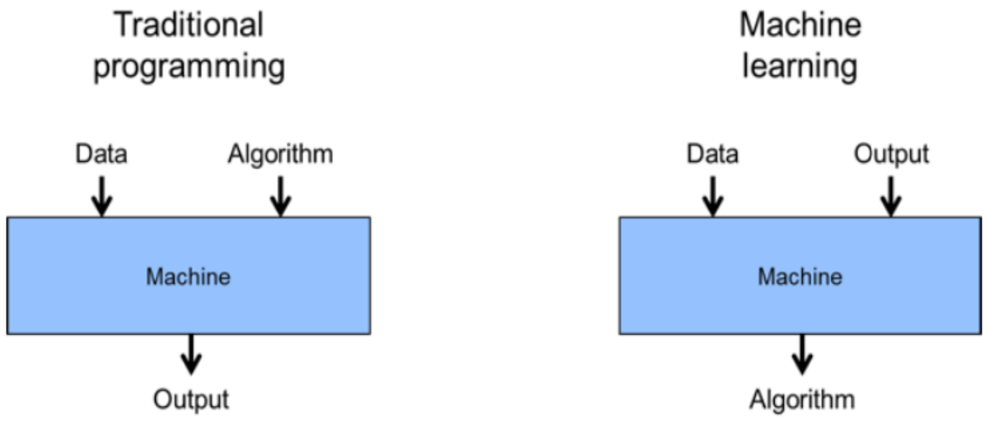

Figure 3.3: Left figure: Development of traditional software algorithms. Right figure: Machine learning application

iments involved 17 participants performing real-life driving for a specific amount of time. Collected data included multiple biophysiological signals (ECG, electromyogram (EMG), EDA, and RESP).

The database in Ma et al.[22] provides data in the context of four driving emotional states, namely happy, bothered, concentrated, confused. Experiments involved 10 participants driving for approximately 24 km each. Collected data includes face videos and the publicly available dataset only contains facial features associated with facial expressions due to privacy reasons.

Finally, the CIAIR dataset[23] includes recordings from real-world driving. Experiments involved more than 500 subjects driving for approximately 60 minutes each. No emotional annotations were provided with the dataset. Collected data included multi-channel video from three cameras, multi-channel audio from 16 microphones, and GPS signals.

3.3 Machine Learning Basics

Machine learning is a subset of artificial intelligence. It is a field that combines the disciplines of computer programming, statistics, and mathematics. An adaptive algorithm tries to learn the underlying probability distribution of the given data samples.[24] The information the algorithm tries to find is hidden inside the distribution of the data. Figure 3.3 illustrates a comparison of the development of a machine learning application and the development of traditional software algorithms. In comparison to traditional programming, a machine learning algorithm gets, in addition to the input data, the expected output and learns the parameters of the algorithm from these two ingredients.

Two types of machine learning exist: supervised and unsupervised learning. For supervised learning, it is known to which class the samples from the input data belong. Hence the algorithms get, in addition to the measured data, also a vector that contains the labels of each sample. For unsupervised learning, the step of assigning a label to the input data by a supervisor is not necessary.

Therefore in the context of this master thesis, the machine learning algorithm retrieves as input various sensor information from the driver and the car and learns from these input data the differences between negative emotional states. In addition, the algorithms also get a vector of labels that contain the class belongings of the sensor information.

To test the predictability of the machine learning algorithm, that is, the performance of the algorithm, the input data are split into a test and training set. Hereby the training set is only used to train the algorithm. Afterward, the performance of the algorithm is measured by the predictability of the previously unseen test data.

3.3.1 Deep Learning

Deep learning is a class of machine learning algorithms that uses multiple layers to gradually extract higher-level features from the raw input. Deep learning methods are based on artificial neural networks (ANNs), which are mostly referred to as neural networks (NNs). ANNs were inspired by the information processing of the nodes in biological systems.[25]

It is called "deep" learning because a neural network typically includes multiple layers and forms a deep structure. Each level learns to convert its input data into a slightly more high-level and abstract representation. For example, for emotion recognition from facial expressions, a matrix of pixels can be the raw input, and the first representational layer can transform the pixels and encode edges. The second layer can create and encode arrangements of these edges. The third layer can encode the nose and eyes, and finally, the fourth layer can recognize the fact that the frame contains a face. Importantly, a deep learning structure can learn and extract the optimal features of each layer on its own. This eliminates the requirement of expert knowledge in a particular domain. However, the need for hand-tuning for hyperparameters still exists, for example, the selection of the number of layers and layer dimension sizes.

3.4 Modalities for Emotion Recognition

To investigate the different methods of emotion measurement or recognition, we begin by providing a broad overview of methodologies that cover unimodal studies. The largest and most researched categories are neuroimaging, ANS, facial expression, and speech.[26] These categories evaluate techniques from multiple areas such as signal processing, machine learning, computer vision, and speech processing. Following subsections focus on studies on the most researched modalities in the automotive field and briefly summarize their methodologies.

3.4.1 Audio-based

Audio-based approaches are one of the most researched and applicable models. Since emotions like anger, irritation, and nervousness are strongly correlated to simple speech features like volume, energy or pitch, audio-based approaches are found to be especially good for the recognition of these driver states[12]. However, driver states such as sleepiness cannot be reliably recognized with the speech signal as a sleepy driver does not tend to talk much. One of the great advantages of the speech modality is low hardware costs and giving less apparent observation feeling to the driver[6]. This makes the audio-based approaches more applicable from the user acceptance point of view. On the other hand, the user can control how much emotion is shown, which of course is a disadvantage for a constant and reliable driver monitoring. Audio information is not continuously present if the driver does not constantly speak. For instance, the previously mentioned sleepiness state is one of the most prominent examples of this situation. Furthermore, studies have shown that open microphone settings and language differences make recognition tasks depending on audio rather challenging[27]. As emotions have found to be highly related to the language spoken, the mean accuracy of the studies varies depending on the language. Therefore, language dependant studies result in better estimations[28].

3.4.2 Behavioral-based

The final approach is the use of driving style as a modality for emotion recognition. It exploits features like steering wheel, acceleration, lane-keeping and reacting. Since the driving style that is relevant to these features and the active state of the driver are highly correlated, it could be utilized for in-vehicle emotion recognition. For level 3 and level 4 autonomous driving, it could seem a bit useless however, the steering wheel can still exist and the driver is likely to drive the car. It still worth mentioning as the results are quite promising. One of the advantages of this approach is that it is quite obvious from the emotional point of view. Further, it is less costly as it does not require any extensive hardware compared to physiological approaches. However, it is not investigated very intensely so far[6].

3.4.3 Camera-based

As the face is considered to be the major input resource for recognizing the emotions, video and image based approaches are the most focused approaches. This approach exploits the features like eye tracking, head movement, and facial expressions. Recognizers which use visual information have found to be good at detecting 6 basic emotions like anger, sadness, happiness, disgust, fear, irritation, and surprise. It is seen to be especially good for detecting the interest level of the driver compared to the audio modalities[6]. In contrast to speech, one advantage of the visual approaches is that the visual data is continuously present. Therefore, emotions cannot be hidden or controlled by the driver. However, this also gives an increased observation feeling to the driver which is a disadvantage from a user acceptance point of view.

3.5 Physiology-based Modality

Besides from a pattern recognition point of view, more extensive approaches are about collecting and analyzing the physiological signals, which include the electroencephalogram (EEG), body temperature (BT), electrocardiogram (ECG), electromyogram (EMG), electrodermal activity (EDA), respiration rate (RSP), etc. One advantage of this approach is that it provides reliable and better accuracy as the data is directly coming from the user's body and is highly correlated to emotions such as excitement, anger, and nervousness. Depending on the type of signal and the type of hardware that will be used for the recognition task, the hardware costs of physiological measurements could be affordable. However, it would be a disadvantage to wearing devices that give an apparent feeling of being watched from a user acceptance point of view.[6]. This consequently also affects the user's current emotions and behaviors. Further, working with these kinds of signals is highly challenging. For instance, EEG data can be noisy and diverse. An adequate amount of research has been done in this field which exploits several different types of physiological data. In the following section, we present some of the previously done work utilizing different physiological signals.

3.5.1 Machine Learning Approaches

So far, most of the studies focusing on physiological signals have applied various feature-based machine learning (ML) approaches which includes extracting the valuable

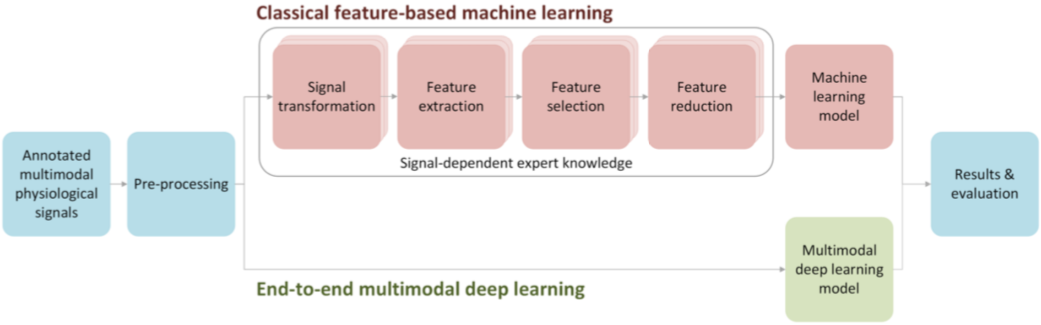
Classical feature-based machine learning
Signal transformation
Feature extraction
Feature selection
Feature reduction
Machine learning model
Signal-dependent expert knowledge
Annotated multimodal physiological signals
Pre-processing
Results & evaluation
Multimodal deep learning model
End-to-end multimodal deep learning

Figure 3.4: Comparison of feature-based ML methods and end-to-end DL methods which is applicable to emotion recognition from physiological signals problem[32]

features from signals and processing those features based on some expert knowledge in order to train classifiers for recognizing emotions. Figure 3.4

Khezri et al.[29] combined EEG with galvanic skin response (GSR) to recognize six basic emotions via k-nearest neighbors (kNN) classifiers. Yin et al.[30] followed an ensemble of autoencoders approach for sentiment recognition from physiological signals. Features from seven different physiological sensors were utilized, namely EEG, EDA, EMG, ECG, blood volume pulse (BVP), and RSP sensors. In general, the methods based on EEG data usually outperform the ones based on other data[31]. It is predicted to come from the fact that EEG provides a more direct channel to one's mind.

In the field of stress detection using physiological sensors and ML are Healey and Picard [21], who proposed a quite accurate stress detection system in 2005. It achieved an accuracy as high as 97% when tested in a constrained real-life scenario, for instance subjects driving a car. The authors utilized ECG, EMG, EDA, and respiration signals to extract features and feed as input to the system. This study had confirmed that stress detection was possible in a real-life scenarios, even though the presented system was intrusive.

All the presented approaches for emotion and stress recognition are based on features extracted from signal data, mostly provided by the wearable sensors. Feature extraction and selection, and domain dependant signal analysis are the most important and challenging steps in the ML-based approaches. Figure 3.4

3.5.2 Deep Learning Approaches

As previously explained, deep learning is a class of ML algorithms that uses multiple layers of nonlinear processing units. A great benefit of the DL models based on

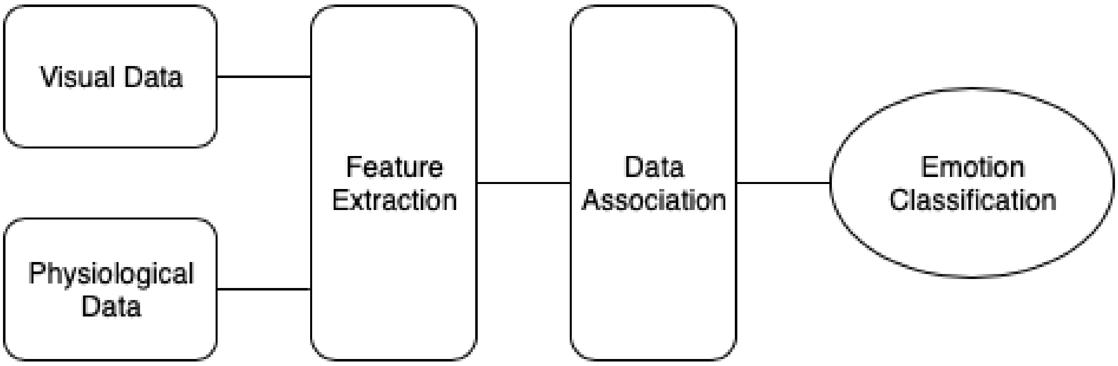

Figure 3.5: Generic pipeline of an early fusion approach

Convolutional Neural Networks (CNNs) and Long Short Term Memory Networks (LSTMs) is that they are able learn directly from the raw data, thus avoiding the need for signal and feature processing. Figure 3.4

In a study of the valance level while walking in the city center, participants filled Self-Assessment Mankin (SAM) questionnaires.[33] The collected data, including heart rate, EDA, body temperature, and motion data, was put into three different DL architectures, including a Multi-Layer Perceptron, a CNN, and a CNN-LSTM. The models respectively achieved an F1-score of 0.63, 0.71, and 0.87. Another approach to end-to-end DL was presented by Li et al.[34]. Firstly, each raw signal was transformed into a spectrogram. These spectrograms later were fed into an attention-based bidirectional LSTM network and a DNN. This approach achieved an F1-score of 0.72 for binary arousal recognition and 0.70 for binary valence evaluation. Qiu et al.[35] proposed Correleted Attention Networks for emotion recognition using bidirectional Gated Recurrent Units (GRUs), a Canonical Correlation Layer, a Signal Fusion Layer, an Attention Layer, and a Classification Layer. Evaluation of this architecture was done on three different datasets: SEED, SEED IV, and DEAP. Their framework achieved higher accuracy than feature-based support vector machine, although details about the features utilized were not provided. Additionally, CNNs were used on the MAHNOB-HCI dataset[36] to achieve better accuracy than those found using the methods based on feature extraction[37].

3.6 Multimodal Fusion

Although approaches exploiting single modalities seem to achieve above baseline accuracy, a single modality could not be fully reliable in the real world environment. In an in-vehicle environment, it is vital to have redundancy by utilizing data from different sensors. This requires processing data from more than one modality and find relations

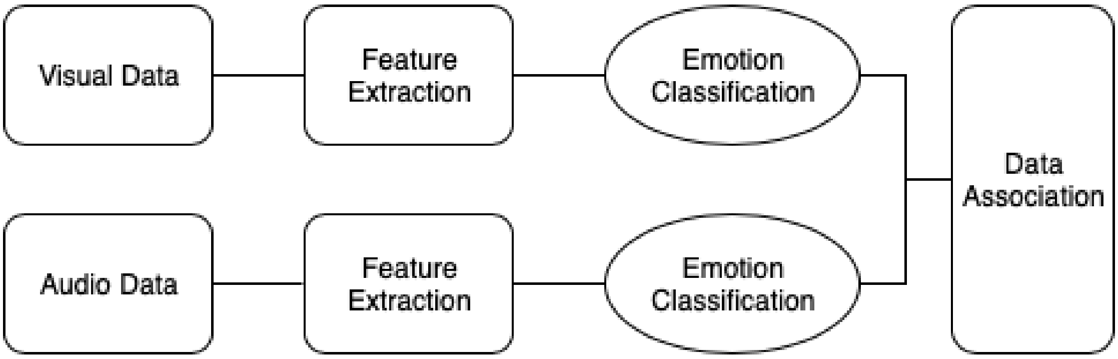

Figure 3.6: Generic pipeline of a late fusion approach

between modalities, just like intelligent beings do in an naturally noisy and multimodal world. Many ML-based fusion approaches have been proposed to handle multimodal information for classification tasks.[38] The most widely used methods are data level fusion (or early integration) and decision level fusion (or late integration).[39] Data level fusion as depicted in Figure 3.5 integrates low-level features from each modality by correlation, which potentially accomplishes an improved task, but has difficulty in temporal synchronization among various input sources. Decision level fusion as presented in Figure 3.6 obtains unimodal decision values and integrates them to acquire the final decision. Although the late fusion ignores some low-level interactions between modality, it allows easy training with more flexibility and simplicity to make predictions when one or more of modalities are missing. The hybrid (or mid-level) fusion attempts to exploit the advantages of both early and late fusion in a common framework[40].

Caridakis et al.[41] proposed a multimodal system that utilizes a Bayesian classifier and performs recognition of eight emotions. The system integrates information from facial expressions, body movement and gestures and speech. The authors separately trained classifiers for facial and behavioral modalities and then performed both feature level and decision level fusion. They observed significant improvement on the recognition rates when the multimodal data was fused, as compared to the individual modalities. Further, the fusion performed at the feature level showed better results than the one performed at the decision level. Tzirakis et al.[42] proposed an end-to-end fusion that operates on the raw signal. To perform an end-to-end spontaneous emotion prediction task from speech and visual data, LSTMs were used. They fused visual and speech modalities. To speed up the training, individual networks were trained separately. The multimodal model greatly outperformed other models. There is scope to improve their work by adding other modalities to the same fusion architecture. Furthermore, according to[43] DL approaches require more modalities in order to perform better. They compared several ML and DL methods exploiting physiological and visual signals, namely, nasal EDA (nEDA), palm EDA (pEDA), heart rate (HR), breathing rate (BR) and eye tracking data.

4 Proposed Methodology

4.1 Proposed Approach and Methodology

In this chapter, we outline the proposed approach and the methodology through which we attempt to address our main research questions. In order to investigate the role of physiological signals in a multimodal emotion recognition system, we propose developing end-to-end DL architectures for the classification task. We then feed these networks the preprocessed raw sensor data to discover useful patterns. Later, the best performing modality is integrated into the baseline multimodal fusion model. Lastly, we find out the appropriate preprocessing methods for physiological signals, the accuracy and flexibility level of multimodal systems for in-cabin environment, and compare the benchmarks and limitations of uni/multimodal solutions focusing on physiological signals.

Main steps required in the development of a ML system for emotion recognition includes:

- Data Collection
- Signal Preprocessing
- Feature Engineering / Dimensionality Reduction
- Classification
- Validation and Evaluation

We follow these steps and also fuse the developed modality into multimodal classification architecture. Next, we elaborate the methodologies we use in these steps according to the proposed approach.

4.1.1 Dataset

For this work, it is important to utilize a database that includes physiological signals as well as other modalities, is sufficiently large, and includes a variety of emotional driving scenarios. Additionally, having physiological signals that were obtained on a driving simulator with non-obstructive sensors is another important criterion.

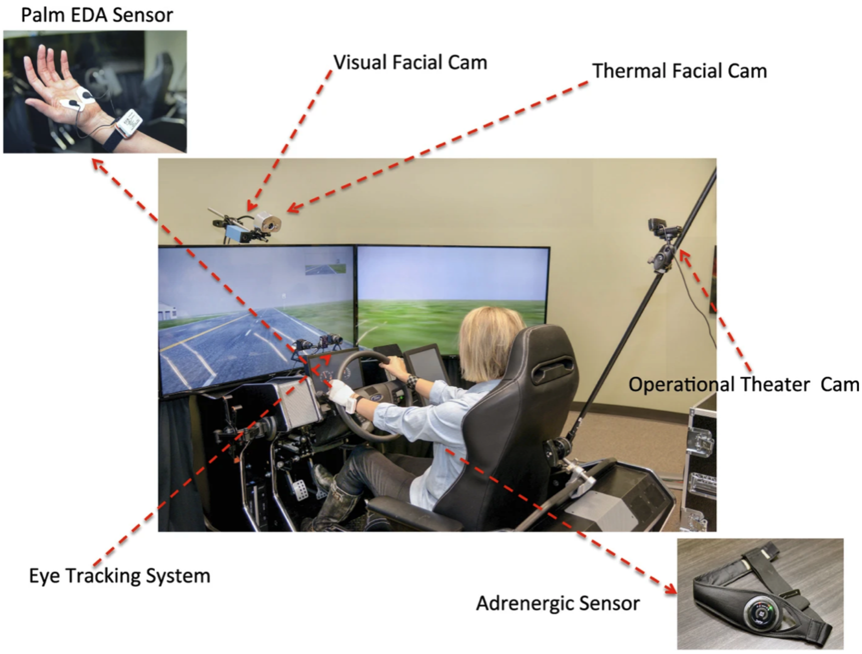

Figure 4.1: Experimental setup of the dataset[44]

We, therefore, utilized a previously collected database due to limitations in hardware and participant acquisition. We selected a dataset called *'A multimodal dataset for various forms of distracted driving'* which met our requirements. The data for the study was acquired from 68 participants using a driving simulator. In Figure 4.1 an example of the driving simulator setup is depicted. The following description elaborates the setup. The experiments were done with the participation of n=68 volunteers that drove the same highway under four different conditions. These conditions were no distraction, cognitive distraction, emotional distraction, and sensorimotor distraction. A special driving session was conducted at the experiment where all subjects experienced a startle stimulus in the form of unintended acceleration; half of them under a mixed distraction and the other half in the absence of a distraction. During the experimental drives, key response variables and several explanatory variables were continuously recorded. The response variables included speed, acceleration, brake force, steering, and lane position signals. The explanatory variables included perinasal electrodermal activity (EDA), palm EDA (pEDA), heart rate (HR), breathing rate (BR), and facial expression signals.[44].

4.1.2 Preprocessing

Since we are focusing on physiological signals, we will be dealing with raw sensor signals which are noisy and exposed to external interference such as subject movement, disconnection, unstable environment temperature, and missing data. The signal preprocessing step is vital for our research, and it will be applied to the raw signals as a first step. Generally preprocessing step consists of: synchronization of the different sensor's signals; removal of data-loss and null values; filtering, noise, and outlier removal. The most common preprocessing techniques used on EDA, HR, and BR signals are Butterworth filter, low-pass filter, smoothing, downsampling, winsorization, and normalization[45]. The selection of the technique and the related parameters depends on the data and sensors used to collect the data. Therefore, at this step, we plan to use the trial and error method to find out the right techniques that work well on our selected dataset.

As we plan to use feature-independent ML methods, we need little to no requirement of handcrafted features. Thus, the feature extraction step is also irrelevant. We only use preprocessing techniques that are not related to the specific domain knowledge and to a corresponding ML technique. This step is also vital for our research as one of our research questions is directly related to analyzing the impacts of preprocessing on DL performance.

4.1.3 Classification

The next step is to develop the physiological modality. To this end, several end-to-end DL architectures that can learn from several sensors simultaneously and also from time-series data are implemented, mostly inspired by Gjoreski et al.[46] They compared ten end-to-end DL architectures that are specialized for time-series classification in their study. Their work provides a solid base for a fair comparison of different DL architectures and analyzing the influence of the preprocessing techniques. We utilize these architectures by considering our chosen dataset and signals. Three best performing DL architectures according to their experimentation on four different datasets are Spectrotemporal Residual Network (Stresnet)[43], Residual Network (Resnet)[47], and Fully Convolutional Network (FCN)[48]. These CNN-based architectures were taken from corresponding cited works and enhanced to learn from multilevel physiological signals[46]. Most accurate results were obtained from the dataset, which focuses on stress detection and with ECG, EDA, BCP, and RSP. As we will use the same signals to classify the emotional states of the driver, utilizing these architectures for our dataset can give promising results. Additionally, Stresnet was originally proposed for driver distraction detection. In the following sections, we will give more details about the architectures and our motives to utilize them.

Fully Convolutional Network

A fully Convolutional Network (FCN) has shown compelling quality and efficiency for semantic segmentation on images. FCN architecture consists solely of locally connected convolution layers, pooling, and upsampling. Each output pixel is a classifier corresponding to the receptive field, and the networks can thus be trained pixel-to-pixel given the category-wise semantic segmentation annotation.[49] Avoiding the use of dense layers reduces the number of trainable parameters, thus making the network faster to train.[50]

FCN architecture proposed by Wang et al.[49] provides a strong baseline for time series analysis. The architecture consists of an input layer, followed by three hidden basic blocks, and finally, a softmax output. The network includes fully convolutional layers as their basic blocks. We utilize the version which was enhanced to learn from multimodal data by horizontally stacking branches dedicated to each signal[46]. In each branch, after the convolutional blocks, the extracted features are fed into a global averaging layer, which reduces the number of weights significantly. Extracted features are then concatenated and fed into the final output layer. Detailed architecture is depicted in Figure 4.2. In our problem setting, FCN could perform as a feature extractor. We replace the GAP layers with flattened layers whose outputs are then concatenated and fed into dense layers in order to get a 128d vector as a feature map. We are interested in exploring how well this lightweight yet powerful model could perform on time series physiological signals data.

Residual Network

Residual Network (Resnet) achieves state-of-the-art performance in object detection and other vision-related tasks. It consists of residual blocks stacked on top of each other to form a network. Residual block has a shortcut connection that provides a reference to the layer inputs, which enable the gradient flow directly through the bottom layers. Instead of hoping every few stacked layers directly fit a desired underlying mapping, Resnet lets these layers fit a residual mapping.[51]

Resnet architecture proposed by Wang et al.[52] provides a strong baseline for time series analysis. It consists of convolutional blocks to build each residual block. It stacks three residual blocks, which are followed by a global average pooling layer and a softmax layer. The architecture excludes any pooling operation to prevent overfitting. We utilize the version which was enhanced to learn from multimodal data by horizontally stacking branches dedicated to each signal[46]. Each branch consists of

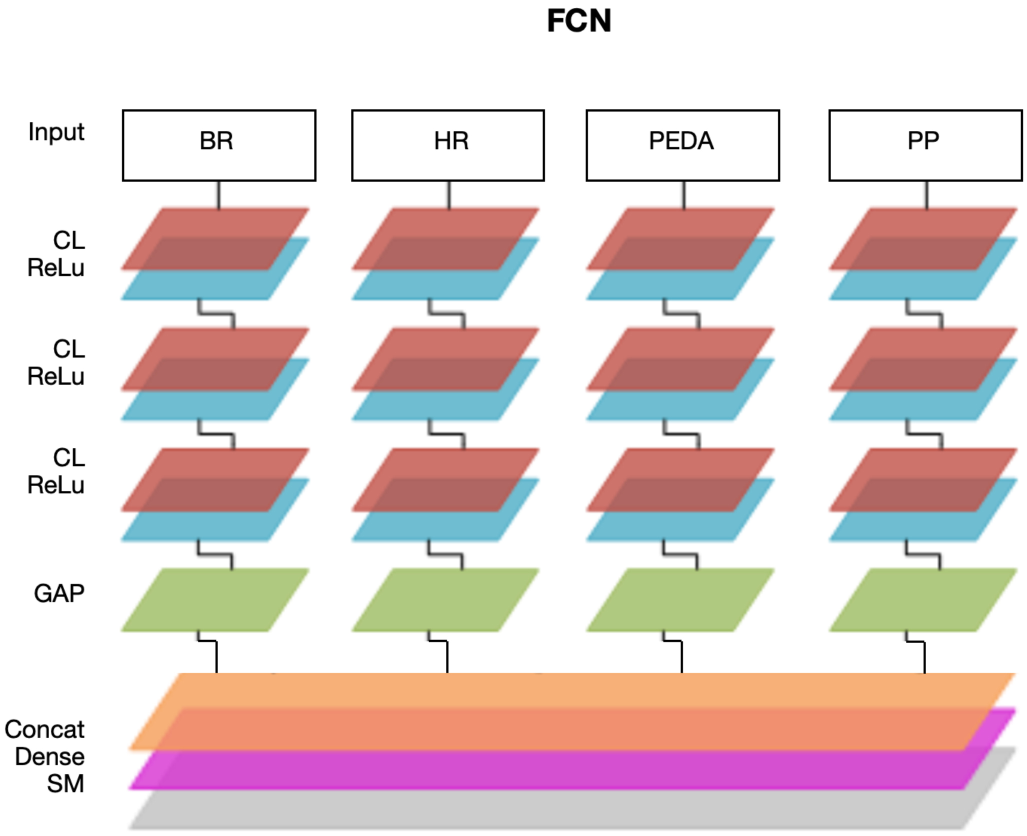

Figure 4.2: FCN Architecture (BR: breathing rate, HR: heart rate, PEDA: palm EDA, PP: perinasal EDA, CL: convolutional layer, ReLu: rectifier layer, GAP: gloabal average pooling, Concat: concatenation layer, Dense: dense layer, SM: softmax activation)

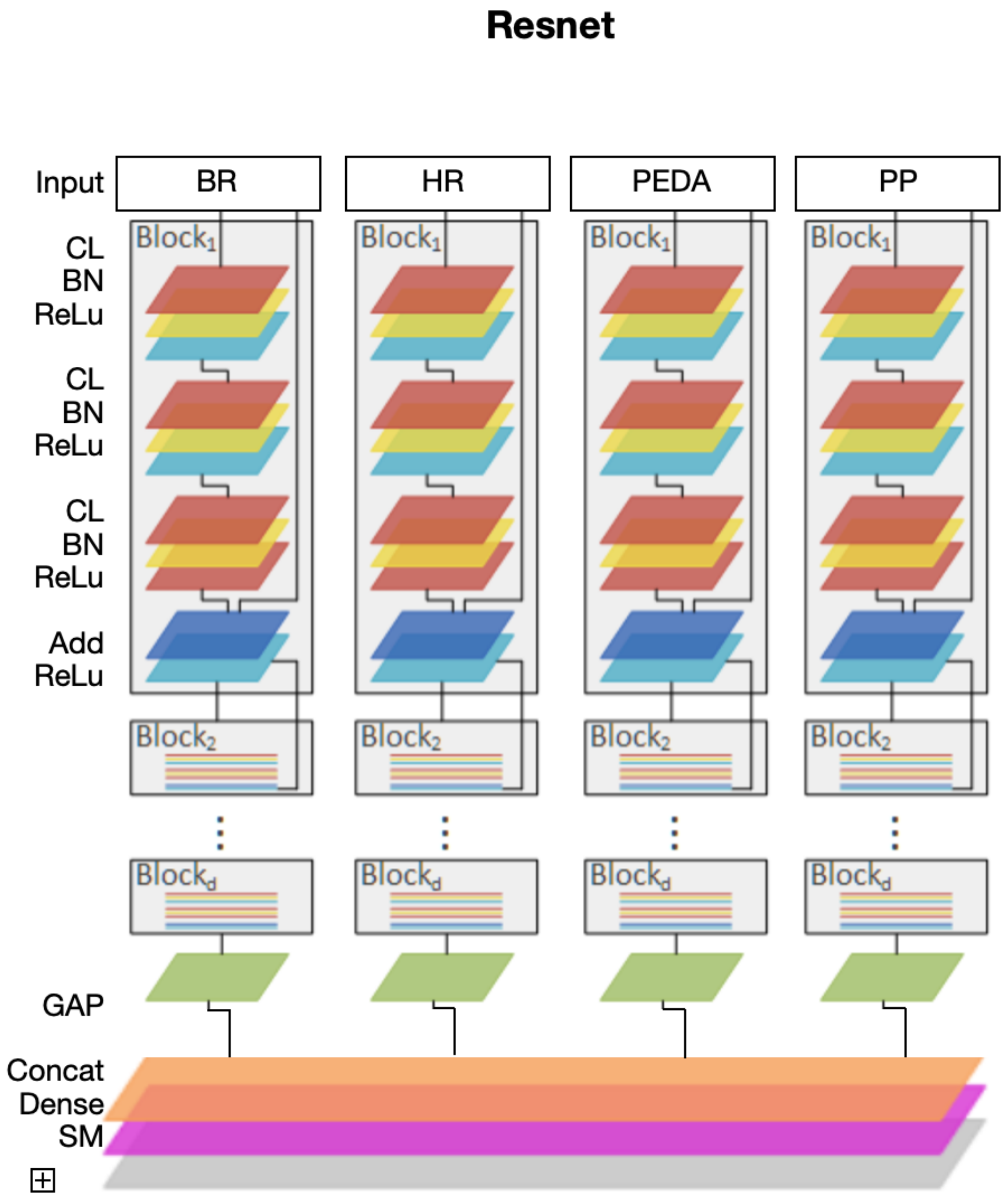

Figure 4.3: Resnet Architecture (BR: breathing rate, HR: heart rate, PEDA: palm EDA, PP: perinasal EDA, CL: convolutional layer, ReLu: rectifier layer,BN: batch normalization layer, GAP: global average pooling, Concat: concatenation layer, Dense: dense layer, SM: softmax activation)

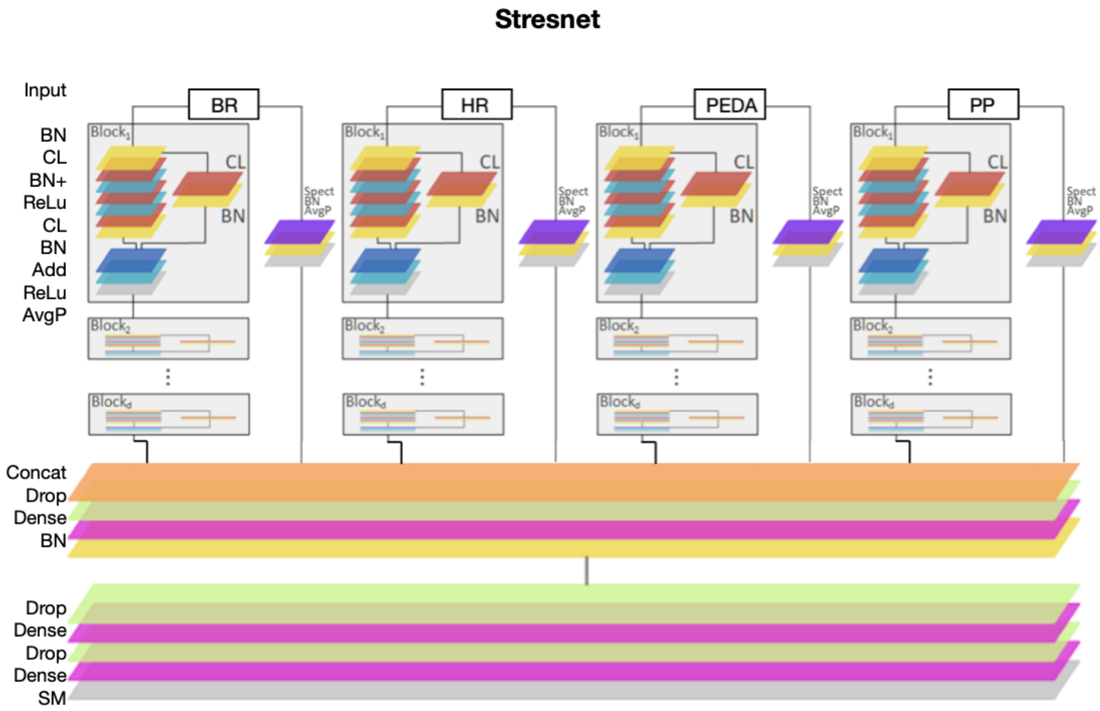

Figure 4.4: Stresnet Architecture (BR: breathing rate, HR: heart rate, PEDA: palm EDA, PP: perinasal EDA, CL: convolutional layer, ReLu: rectifier layer, BN: batch normalization layer, Drop: dropout layer,AvgP: average pooling layer, Spect: spectogram layer, Concat: concatenation layer, Dense: dense layer, SM: softmax activation)

d residual blocks and a global averaging layer. Extracted features are then concatenated and fed into the final output layer. Detailed architecture is depicted in Figure 4.3.

We explore the Resnet structure since we are interested to see how the very deep neural networks perform on the time series data and analyze the pros and cons. We keep the residual block number at most 3 in order to keep the model lightweight. Furthermore, we replace the GAP layers with flattened layers whose outputs are then concatenated and fed into dense layers in order to get a 128d vector as a feature map.

Spectrotemporal Residual Network

The novel neural network architecture Spectro-temporal Residual Network (Stresnet) is an end-to-end network architecture and was originally developed to predict driver distraction from physiological and visual data, and achieved 67% of F1-score[43].

As shown in Figure 4.4, in the Stresnet network, each data signal is associated with two branches: a Resnet that processes the sensor data in the time domain and another branch analyzing a spectral representation of the sensor signal. Each branch consists of layers stacked vertically. The architecture was later enhanced to learn from multimodal data by horizontally stacking branches dedicated to each signal and achieved up to 62% of F1-score[46]. These two branches, namely, the spectral and the temporal ones, are followed by fully a convolutional layer which first merges them and feds them into a dense layer with softmax activation. Each dense layer is followed by a dropout layer to prevent overfitting. We keep the residual block number at most 3 in order to keep the model lightweight. Furthermore, we modify the dense layers in order to get a 128d vector as a feature map.

Since we will be dealing with several physiological signals that were collected with different sensor types in different frequencies, it is important to analyze them not only in the time domain but also in the frequency domain. Particularly because this is an end-to-end learning approach that does not include feature extraction by an expert, this approach will allow the network to extract additional features.

4.1.4 Validation and Evaluation

In the validation and evaluation step, we will compare and identify the characteristics of DL architectures that can learn better from physiological signals by conducting several experiments. The best performing model will later be fused into a multimodal recognition system as an additional modality.

4.1.5 Multimodal Fusion

Finally, we will develop a sensor fusion approach to combine the chosen modality or modalities with the physiological signals-based modality to enhance the robustness or accuracy of the model. For this, we make use of the baseline fusion approach, which is developed by fusing behavioral and visual modalities using an early fusion approach mostly inspired by Tzirakis et al.[42] and Arriaga et al.[53]. This design is modular and extensible, which can support possible new networks. Therefore, we can easily analyze the effects on the accuracy by integrating our best performing physiological signals modality to this multimodal approach.

Next, we give details about the methodologies of the individual modalities chosen to create the fusion approach.

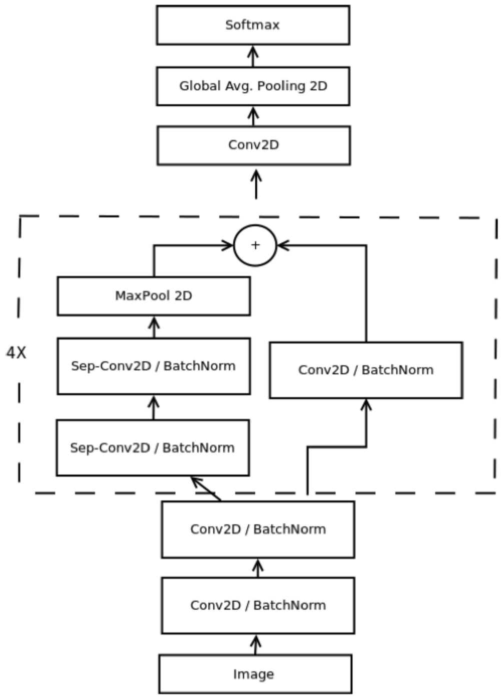

Figure 4.5: Proposed mini-Xception architecture for facial modality

Facial Modality

Analysis of facial expressions is one of the most studied, efficient, and accurate techniques of emotion detection, especially when utilizing CNNs to identify facial features.[54] However, we would like to emphasize that even though they can be used to predict the emotional states accurately on their own, it is not sufficient in the driving context and can be enhanced by the integration of the other modalities, specifically physiological signals modality, in a fusion approach. As discussed previously, facial occlusions or malfunctioning of the camera can lead to failures when only utilizing a facial modality for emotion prediction in the driving context.

To train a facial modality, we use a variation of the Xception model[55] which was modified to be more suitable for real-time affect recognition. This model, namely mini-Xception[53], has 80 times fewer trainable parameters than the original Xception model and is able to extract spatial features by using depth-wise separable convolutions, which is critical for affect recognition from video data and facial expressions. The final architecture uses four residual depth-wise separable convolutions, which are followed by batch normalization and a ReLU activation function. Details of the proposed architecture are presented in Figure 4.5.

Behavioral Modality

Most human communication happens non-verbally, such as body language and facial expressions. Analyzing one's behavioral patterns can have a great impact on detecting emotional states. Studies have shown that behavior modality can be applied to vehicles to classify the emotional states of the drivers by utilizing domain-specific features such as acceleration and steering wheel.[56]. We would like to emphasize that even though they can also be used to predict the emotional states accurately on their own, it is not sufficient in the driving context and can be used to enhance other modalities in a fusion approach.

For the behavioral modality, we utilize a modified version of the mini-Xception model[53] which was previously implemented at the Chair of Robotics, Artificial Intelligence and Real-time Systems at TUM. Since the data from the behavioral sensor is one-dimensional, the model was adapted into a 1D-mini-Xception model.

5 Experiments and Results

In this section, we train and compare three different CNN-based network architectures that utilizes physiological signals. Later, we propose a multimodal emotion recognition approach that does not consist of a physiological signals modality. Lastly, we integrate the best performing physiological signals modality into the previously proposed multimodal emotion recognition approach.

5.1 Data Labeling and Cleaning

In order to begin with the experiments, we needed to clean and restructure our data. The dataset provides the physiological signals, each having a different sampling frequency, and Facial Action Coding System (FACS) signals which were extracted with the CERT software from the facial videos. FACS signals include seven different emotion signals, namely anger, contempt, disgust, fear, joy, sadness, surprise, and neutral. Therefore, we have first downsampled all the physiological signals to 1 fps, thus having 1 data point per second and labeled each data point with the corresponding FACS signal. Resulting data includes synced time series data with BR, HR, pEDA and perinasal EDA signals and the labels.

Table 5.1: An example representation for the restructured data

Time(s)	HR	BR	PEDA	PP	Label
1	..	..	..	..	2
..	..	..	..	..	..

FACS values provide seven different emotions which we have used as the ground truth to label out data. Many researchers have suggested that for various fields, a different set of emotions may be required. For example, in the area of instruction and education students typically experience boredom or delight, however they hardly feel fear or disgust which is an argument for the need for domain-specific classes.[17] According to the survey that has been done on emotion recognition in intelligent vehicles[4], there is a strong preference toward studying emotional states associated

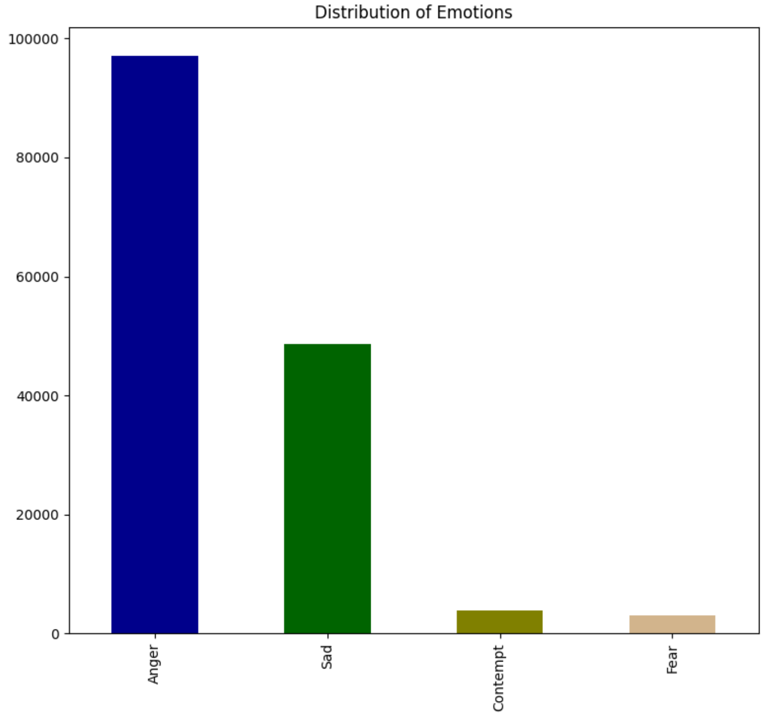

Figure 5.1: Distribution of the negative emotion labels in data

with high arousal and negative valence across the literature. Similarly, in order to classify the abnormal emotions that could be dangerous in cabin environment, we do not need the emotion classes such as disgust. As the experiments never intended to invoke this emotion on the subjects. We only need the set of emotions which could be classified as the abnormal emotions in cabin environment. Therefore, we have dropped all data that was labelled with neutral, disgust, joy and surprise. This reduced our set of emotion classes to anger, sadness, fear, and contempt, thus making our experiments a multi-class classification task. Figure 5.1 shows the distribution of the negative emotion classes in the data.

After inspecting the data, we have observed that there are many missing values. The missing data is in a way that some driving sessions completely miss the data for one signal. Therefore, it is not possible to substitute these missing values with the mean or median values. As we do not have any data from that session and signal to take the mean or median. Therefore, we decided to drop these rows completely.

After cleaning and labeling the data, the distribution of the emotion classes is depicted in Figure 5.1. As it can be seen from the chart, our data is highly skewed. We have the anger emotion as the majority class and almost half as many sadness emotion classes. Finally, we have significantly less observations from contempt and fear emotion classes, which makes them the minority classes. This imbalanced distribution could cause the developed NNs to be biased towards the majority classes.

5.2 Performance Metrics

To measure the performance of the trained machine learning model, it is important to have a completely separated new data set that the model has not seen before. This data set is called the test set. One simple metric to measure the algorithms performance is to look at the accuracy of the predictions that the algorithm made on the test data. Therefore the predicted labels are compared with the ground truth labels from the test set and the accuracy is estimated by dividing the number of correct decisions through the total number of samples inside the test set. It is important to note that the accuracy measure only makes sense if the dataset is balanced. For a balanced dataset the number of samples per class are nearly equal. For imbalanced datasets, which is the case for out dataset, simply measuring accuracy is not enough, especially in multi-class classification. Therefore, different performance metrics are used to measure how good the machine learning algorithm performs in learning the probability distribution of the underlying problem. In the following we introduce the following measures: Precision, Recall, F1, micro, macro and a weighted average.

Precision or positive predicted value is the ratio of true positives to the sum of true and false positives.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

Recall, also referred to as true positive rate, which expresses correctly predicted positive observations with respect to all observations in the class. Often it is useful to weight this metric more in case there is a high cost associated with a false negative.

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

Moreover, F1 score which computes the score by taking both the recall and precision score into consideration as shown in the following equation:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

In addition to recall, precision and F1 score, we also include the micro, macro, and a weighted average for each model. The micro average is calculated based on the global average of true positives, false negatives, and false positives. Micro average takes into account the amount of samples for each individual class and thus is more important in datasets with unbalanced classes. Macro average is simply the average between all classes without considering the number of samples in each. The last metric is the weighted average which is similar to macro average but weighted by label imbalance.

5.3 Physiological signals-based Modality

For the experiment setup, data from each subject were first split into training and test sets. For training set, the train and validation sets were created with a ratio of 4:1. During training, a maximum of 50 epochs were run with early stopping implemented with a patience of 15, i.e., if validation loss had not improved for 15 epochs, the training stopped and the Adam optimizer was utilized with a learning decay. For hyperparameter tuning, Bayesian optimization was used to enable the fair comparison of the DL architectures.[46] However, we have modified the depth of the Resnet and Stresnet architectures to be at most 3 to reduce the training time and the number of trainable parameters considering the in-vehicle environment. The dense output dimension size was also modified to be 128d, mostly to keep the feature maps that will be fused equal sized.

Additionally before the experiments the following preprocessing steps were applied to the data:

- 3–97% winsorization, which removes extreme values form the signal data;
- Butterworth low-pass filter with a 10 Hz cut-off which removes components above the threshold frequency of 10 Hz;
- Min-max normalization;
- Sliding windows of 30 seconds with 1 second slides.

Since our data is highly imbalanced, on a random distribution, several of the validation sets ended up having no observations from the minority class which caused the model to only memorize the data but not generalize on it. Therefore, in order to have the same ratios from each class in each set, we have selected a custom train-validation-test split to use on all the experiments. In addition to custom sets, we also have added class weights to provide bias to the minority classes. We have used the

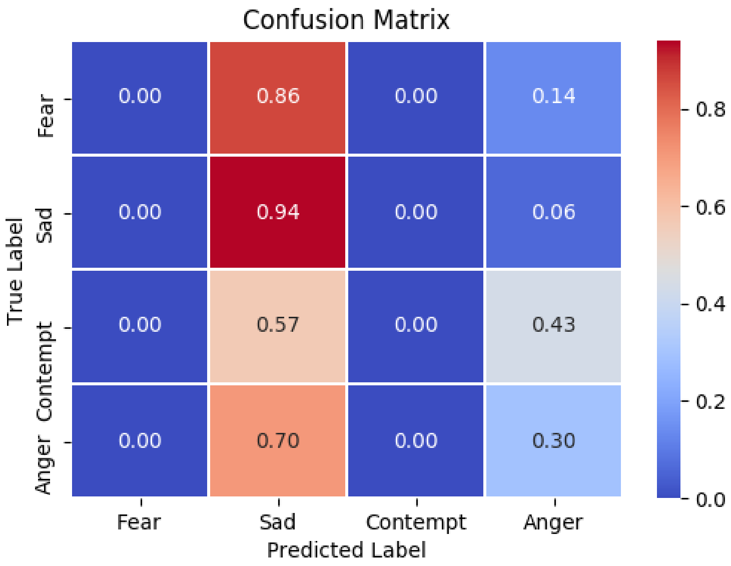

Figure 5.2: Confusion matrix of FCN on negative emotion classes

sklearn.compute_class_weight to calculate the class weights which uses the following formula to calculate the weight associated to the each class.

$$w_j = n/(classes * n_j)$$

where,

- j signifies the class;
- w_j is the weight for each class;
- n is the total number of samples or rows in the dataset;
- classes is the total number of unique classes in the target;
- n_j is the total number of rows of the respective class.

5.3.1 FCN

We utilize the DL architecture FCN[46] for our cleaned and restructured data. For our experiments, we have used the enhanced version of the architecture which is capable of learning from different signal simultaneously. This is crucial for us since we have four different physiological signals data.

Table 5.2: Classification report of FCN on negative emotion classes

	Precision	Recall	F1-score	#Data points
fear	0.0	0.0	0.0	66
sadness	0.35	0.94	0.51	6461
contempt	0.0	0.0	0.0	1328
anger	0.82	0.30	0.43	15078
accuracy			0.46	22933
macro avg	0.29	0.31	0.24	22933
weighted avg	0.64	0.46	0.43	22933

Details of this four class classification experiment is presented in Table 5.2 and Figure 5.2. Although, our model fitted the data well, as it can be seen in the table, it achieves better results in predicting the anger and sadness signals. This result can be interpreted as the consequences of our imbalanced dataset. We have significantly less observations for fear and contempt emotions classes. On the other hand, anger and sadness has high number of observations. This results in biased predictions towards these emotions.

5.3.2 Resnet

We later utilized the DL architecture Resnet[46] for our cleaned and restructured data. For our experiments, we have used the enhanced version of the architecture which is capable of learning from different signals simultaneously. This is crucial for us since we have four different physiological signals data.

Table 5.3: Classification report of Resnet on negative emotion classes

	Precision	Recall	F1-score	#Data points
fear	0.0	0.0	0.0	66
sadness	0.40	0.19	0.26	6461
contempt	0.0	0.0	0.0	1328
anger	0.67	0.89	0.76	15078
accuracy			0.64	22933
macro avg	0.27	0.27	0.26	22933
weighted avg	0.56	0.64	0.58	22933

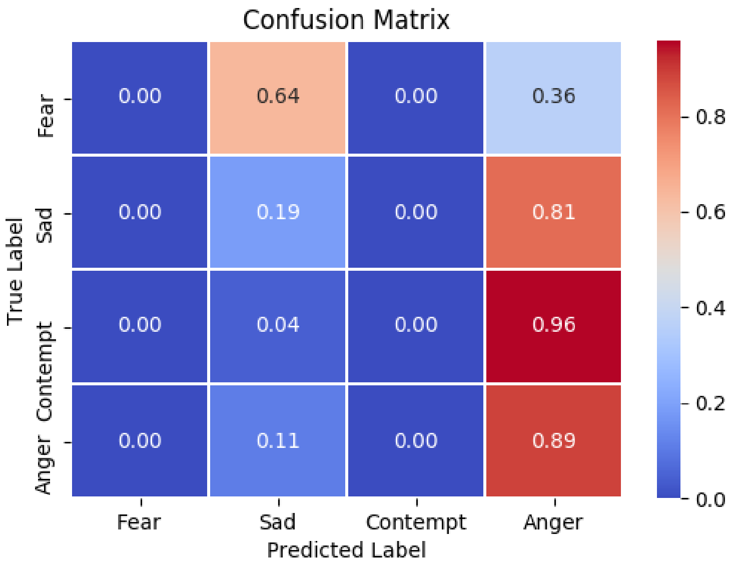

Figure 5.3: Confusion matrix of Resnet on negative emotion classes

Details of this four class classification experiment is presented in Table 5.3 and Figure 5.3. As it can be seen in the table, Resnet achieves a performance close to the FCN architecture. Although it performs better at predicting anger emotion class, percentage of incorrectly classified observations are immensely higher than FCN. Considering that Resnet has deeper neural network structure and more trainable parameters, having comparable results with the lightweight FCN architecture outweighs its advantages.

5.3.3 Stresnet

Lastly, we have utilized the DL architecture Stresnet[46] for our cleaned and restructured data. For our experiments, we have used the enhanced version of the architecture which is capable of learning from different signals simultaneously. This is crucial for us since we have four different physiological signal data.

Details of this four class classification experiment is presented in Table 5.4 and Figure 5.4. As it can be seen in the table, it achieves better results in predicting the anger and sad signals and a overall better performance with the accuracy of 66%. This result can be interpreted as the benefits of analyzing the input data both in spatial and temporal domains.

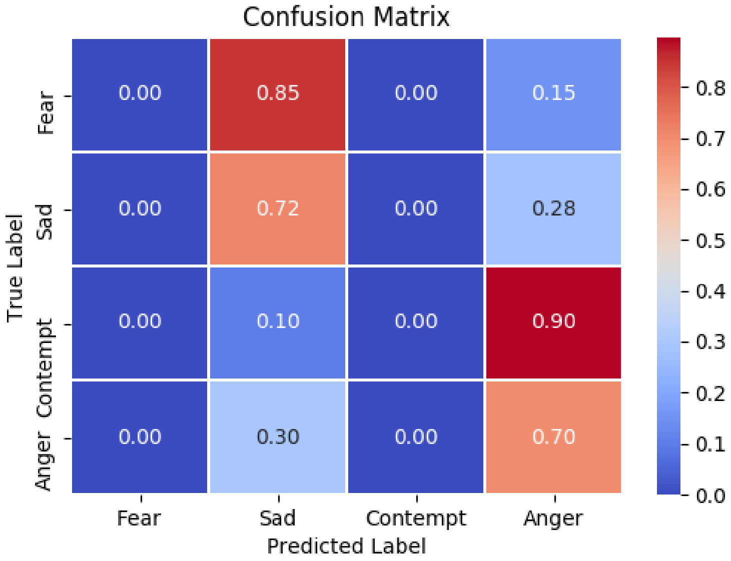

Figure 5.4: Confusion matrix of Stresnet on negative emotion classes

Table 5.4: Classification report of Stresnet on negative emotion classes

	Precision	Recall	F1-score	#Data points
fear	0.0	0.0	0.0	66
sadness	0.49	0.72	0.59	6461
contempt	0.0	0.0	0.0	1328
anger	0.78	0.70	0.74	15078
accuracy			0.66	22933
macro avg	0.32	0.35	0.33	22933
weighted avg	0.65	0.66	0.65	22933

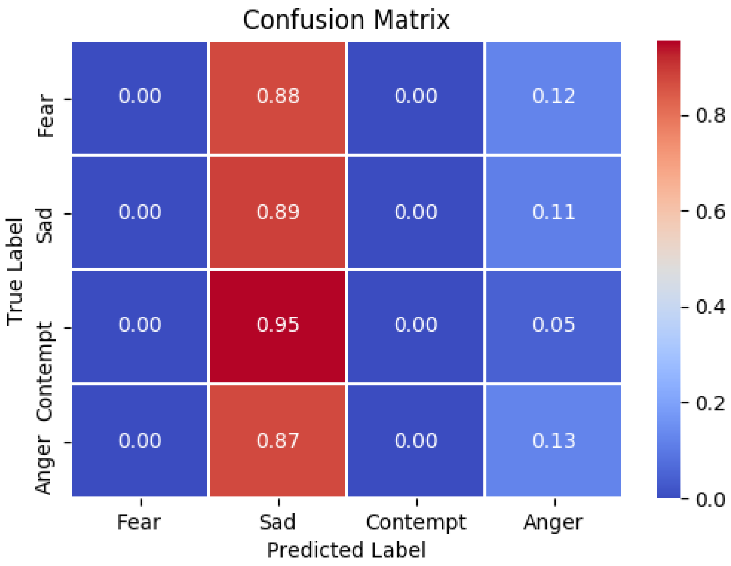

Figure 5.5: Confusion matrix of facial modality on negative emotion classes

5.4 Multimodal Fusion

In order to evaluate and benchmark the developed physiological signals model in a multimodal approach, we built and trained CNNs for behavioral modality and facial modality. In order to fuse the sensors, we followed a feature level fusion approach and extracted features from each modality and concatenate them into one long vector, which is then fed into a CNN-LSTM. To do so, we first trained the chosen modalities separately.

5.4.1 Facial Modality

To train the facial modality, we have used the previously preprocessed video data. Each frame of the video data was extracted and reduced to a 64x64 resolution. This was a good trade-off between performance and speed. As previously discussed, we utilized the mini-Xception[53] to train the facial modality. We trained the model with 50 epochs using a batch size of 16, and with the Adam optimizer on the four previously discussed negative sentiment classes. Additionally, we balanced the classes using the same weighting mechanism used in section 5.3 to adjust the weights based on the number of observations from each class. For training purposes, we kept the last global average pooling layer and the soft-max activation function to produce a prediction.

We modified the network to extract features before the last pooling layer into a 128-dimensional feature vector, which later was fused with equally sized feature vectors from the other CNNs.

Table 5.5: classification report of facial modality on negative emotion classes

	Precision	Recall	F1-score	#Data points
fear	0.0	0.0	0.0	40
sadness	0.33	0.51	0.42	14677
contempt	0.0	0.0	0.0	1100
anger	0.66	0.52	0.58	28563
accuracy			0.61	44380
macro avg	0.25	0.26	0.25	44380
weighted avg	0.54	0.50	0.51	44380

Details of this four class classification experiment is presented in Table 5.5 and Figure 5.5. As it can be seen from the confusion matrix, facial modality leans more towards a biased state of predicting sadness. Studies have shown that most of the aggressive behaviors on the road expressed verbally, and drivers developed a coping mechanism by focusing on safe driving or trying to relax.[57] This can result in anger emotions being less expressive through facial expressions.

5.4.2 Behavioral Modality

To train the behavioral modality, firstly we further processed the data. Similar to the preprocessing that was done for the physiological signals modalities in section 5.1, the most relevant features to behavioral modality were selected and the rest of the features were dropped. The relevant features include acceleration, steering, brake, and speed signals.

As discussed before, high number of *not a number* values existed in the dataset. Therefore, we have dropped all the frames that had these missing values. Similar to the preprocessing steps described in section 5.3, we standardized the entire dataset between -1 and 1, labeled the frames with the four previously discussed negative sentiment classes, and applied a sliding window of 30 data points with a step size of 1 to the data before the training. We used 80% of the total data as training data and the other 20% for the test set. 20% of the training data was used as the validation set. As previously discussed, we utilized the modified version of the mini-Xception[53], the 1D-mini-Xception to train the behavioral modality. Training was performed with 50

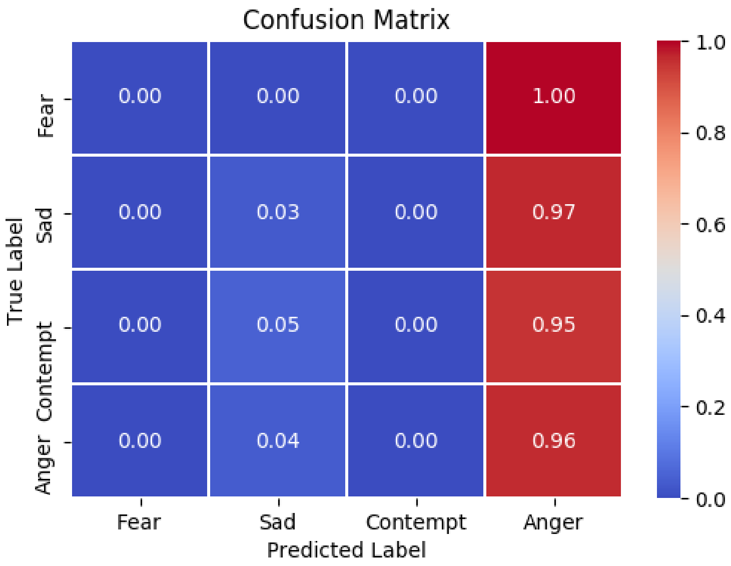

Figure 5.6: Confusion matrix of behavioral modality on negative emotion classes

epochs using a batch size of 16. As optimizer, the Adam optimizer with learning decay was utilized.

Table 5.6: Classification report of behavioral modality on negative emotion classes

	Precision	Recall	F1-score	#Data points
fear	0.0	0.0	0.0	66
sadness	0.25	0.03	0.06	6459
contempt	0.0	0.0	0.0	1328
anger	0.66	0.96	0.74	15074
accuracy			0.64	22927
macro avg	0.23	0.25	0.21	22927
weighted avg	0.54	0.61	0.56	22927

Details of this four class classification experiment is presented in Figure 5.6 and Table 5.6. As it can be seen from the confusion matrix, behavioral modality leans more towards a biased state of predicting anger emotion state. This can be explained by the fact that anger can easily initiate a cycle of violence, which could result in sudden changes in driving behavior. Furthermore, studies have shown that drivers express

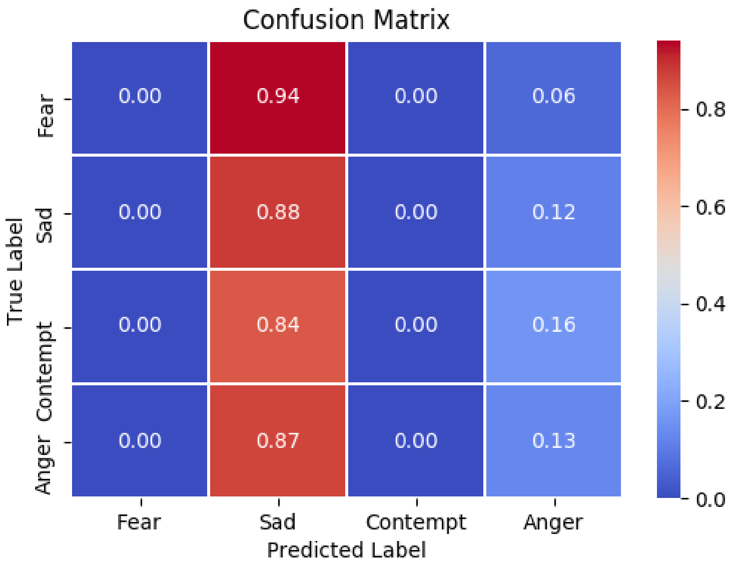

Figure 5.7: Confusion matrix of fused model with behavioral and facial modalities

their anger on the road by their way of using the vehicle.[57]

5.4.3 Baseline Multimodal Fusion

In order to perform the feature level fusion, we fused the equally sized facial and behavioral feature vectors from the CNNs into a combined feature vector. This vector was then used as training input.

Table 5.7: Classification report of fused model with behavioral and facial modalities

	Precision	Recall	F1-score	#Data points
fear	0.0	0.0	0.0	66
sadness	0.29	0.88	0.43	6461
contempt	0.0	0.0	0.0	1328
anger	0.67	0.13	0.22	15078
accuracy			0.33	22933
macro avg	0.24	0.25	0.16	22933
weighted avg	0.52	0.33	0.27	22933

To further process the fused data, we further modified the 1D-mini-Xception and

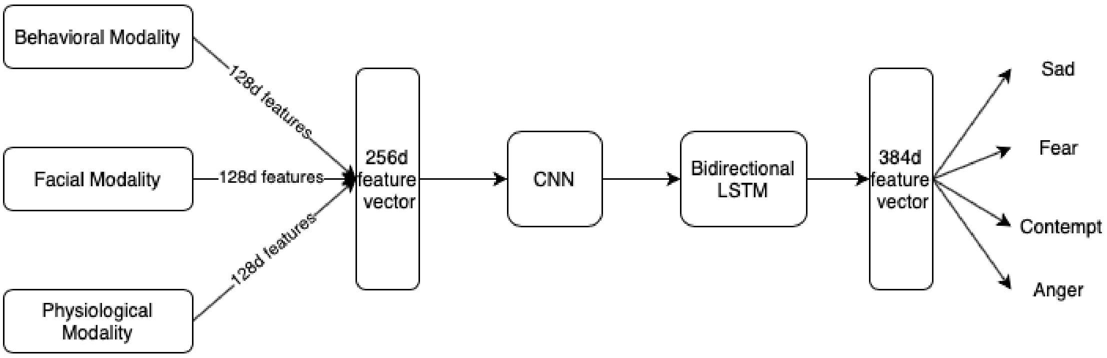

Figure 5.8: Architecture of the final fused model with behavioral, facial, and physiological modalities

added a LSTM layer right before the prediction step in order to incorporate time dependencies. This enabled the network to combine the slower changing behavioral data streams with the fast changing visual data stream and make a prediction of the underlying emotional state.

Details of this four class classification experiment is presented in Table 5.7 and Figure 5.7. As it can be seen from the confusion matrix, the model performs well in predicting the sadness emotion, similar to the facial modality presented in subsection 5.4.1 and reaches an accuracy of 33%.

5.4.4 Final Multimodal Fusion

To evaluate the developed physiological signals modality, we integrated it into the baseline multimodal fusion as the third modality. The overall architecture of this fusion approach is depicted in Figure 5.8. Feature map extracted from the physiological modality is further fused into the input vector of the fusion model. This 384d vector was then fed into the 1D-mini-Xception-LSTM model to make a final prediction of the underlying emotional state.

Details of this four class classification experiment is presented in Table 5.8 and Figure 5.9. As it can be seen from the table, the model has reached an accuracy of 55% by outperforming the baseline fusion approach. Furthermore, the model is able to classify a few contempt emotional state observations correctly, as opposed to the previous models.

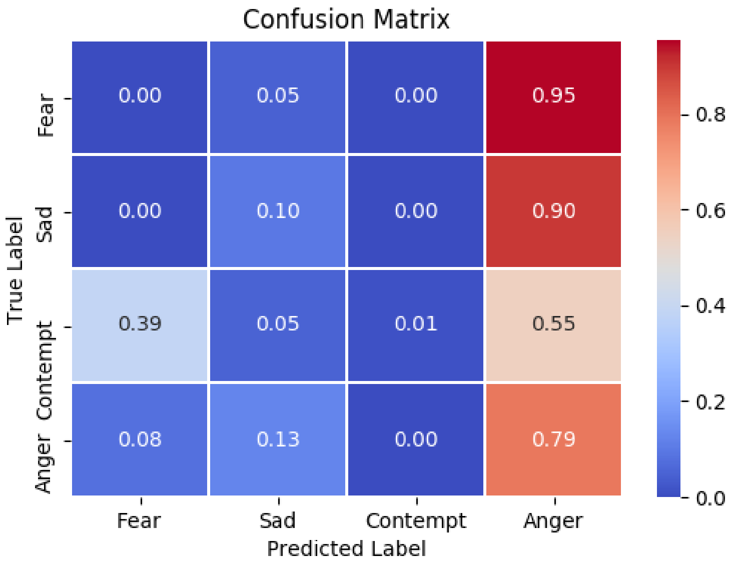

Figure 5.9: Confusion matrix of fused model with behavioral, facial, and physiological modalities

Table 5.8: Classification report of fused model with behavioral, facial, and physiological modalities

	Precision	Recall	F1-score	#Data points
fear	0.0	0.00	0.0	66
sadness	0.24	0.10	0.14	6461
contempt	0.60	0.01	0.02	1328
anger	0.64	0.79	0.71	15078
accuracy			0.55	22933
macro avg	0.37	0.22	0.22	22933
weighted avg	0.57	0.46	0.49	22933

6 Conclusion and Future Work

In this thesis, we have evaluated the importance of physiological signals sensor information in a multimodal negative emotion recognition setup in the context of autonomous driving. We have examined the relevant physiological signals that could be collected in-vehicle environment and have developed three different CNN-based approaches for emotion recognition that can be trained in an end-to-end fashion. These methods include a novel NN approach Stresnet, an FCN, and a Resnet. Furthermore, we have utilized an existing method for multimodal sensor fusion to perform emotion recognition in the context of a vehicle that can be trained in an end-to-end fashion. This model is an early fusion method that consists of fusing the extracted feature vectors from each respective modality and is extendable. To this end, in addition to physiological signals modality, we have developed two more CNN-based modalities which make use of facial expressions and behavioral data.

Among the benchmarked physiological signal modalities, the Stresnet, which analyzes the data in both temporal and spatial domain, outperformed the FCN and Resnet models with 66% accuracy and 33% F1-score and performed well in classifying both sadness and anger emotions. With respect to other individual models, the facial modality predicts emotions at a slightly higher rate than chance level and is better at classifying the sadness emotion class. The behavioral model alone is not able to identify emotions across the participants at an above chance level and is better at classifying anger emotion class. These results strengthen our initial hypothesis of emotion recognition being a multimodal learning problem and that individual modalities are not reliable and robust enough considering the in-vehicle environment. The fusion approach, which uses the fused feature vectors from three modalities, including the best performing Stresnet physiological modality, outperformed the fusion approach, which only uses the features from behavioral and facial expressions modality. The final fused model reached the accuracy of 55%, thus improving the accuracy by 23%. We conclude that these specific physiological signals features are correlated to emotion prediction and that they can be used to increase the robustness of an emotion recognition model in a in-vehicle sensor fusion approach. However, behavioral features alone are not sufficient for accurate emotion estimation in autonomous driving.

We provide a data preprocessing pipeline that extracts, labels, and processes the physiological data information from a multimodal driving dataset across 68 participants that could be fed into three different NN architectures and to any new NN architecture. This pipeline could be useful for future research in evaluating the different preprocessing techniques or benchmarking and comparing more end-to-end neural network architectures that analyze time-series physiological data.

Since ground truth labels used to label the physiological signals data are extracted from the video data, in future work, a more sophisticated ground truth labeling scheme making use of valence and arousal values could be utilized to evaluate the proposed NN architectures and the fusion approach. Additionally, a more balanced dataset could be utilized for the same purposes. As in this dataset, there is a significant imbalance in the observations from each emotion class which results in biased learning. Furthermore, robustness and the reliability of the fusion approach could further be analyzed by adding different modalities, such as audio modality, as our fusion framework is modular and extensible, which can support possible new networks. Lastly, a different fusion approach could be developed and experimented with the same modalities to evaluate the performance of the different fusion approaches.

List of Figures

List of Tables

Bibliography

[1] S. Poria, E. Cambria, R. Bajpai, and A. Hussain. "A review of affective computing: From unimodal analysis to multimodal fusion." In: *Information Fusion* 37 (2017), pp. 98–125.

[2] G. Underwood, P. Chapman, S. Wright, and D. Crundall. "Anger while driving." In: *Transportation Research Part F: Traffic Psychology and Behaviour* 2.1 (1999), pp. 55–68.

[3] S. H. Fairclough, A. J. Tattersall, and K. Houston. "Anxiety and performance in the British driving test." In: *Transportation Research Part F: Traffic Psychology and Behaviour* 9.1 (2006), pp. 43–52.

[4] S. Zepf, J. Hernandez, A. Schmitt, W. Minker, and R. W. Picard. "Driver Emotion Recognition for Intelligent Vehicles: A Survey." In: *ACM Comput. Surv.* 53.3 (2020). ISSN: 0360-0300. DOI: `10.1145/3388790`.

[5] D. Ding, K. Gebel, P. Phongsavan, A. Bauman, and D. Merom. "Driving: A Road to Unhealthy Lifestyles and Poor Health Outcomes." In: *PloS one* 9 (June 2014), e94602. DOI: `10.1371/journal.pone.0094602`.

[6] F. Eyben, M. Wöllmer, T. Poitschke, B. Schuller, C. Blaschke, B. Färber, and N. Nguyen-Thien. "Emotion on the road—necessity, acceptance, and feasibility of affective computing in the car." In: *Advances in human-computer interaction* (2010).

[7] V. Kostov and S. Fukuda. "Emotion in user interface, voice interaction system." In: *in Proceedings of the IEEE International Conference on Systems, Man and Cybernetics* 2 (2000). Nashville, Tenn, USA, pp. 798–803.

[8] S. D'mello and J. Kory. "A Review and Meta-Analysis of Multimodal Affect Detection Systems." In: *ACM Computing Surveys* 47 (Feb. 2015), pp. 1–36. DOI: `10.1145/2682899`.

[9] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang. "A Review of Emotion Recognition Using Physiological Signals." In: *Sensors* 18.7 (2018).

[10] C. M. Jones and I. M. Jonsson. "Automatic recognition of affective cues in the speech of car drivers to allow appropriate responses," in: *in Proceedings of the 17th Australia Conference on Computer-Human Interaction: Citizens Online (OZCHI '05)* 122 (2005). Canberra, Australia, pp. 1–10.

[11] J. A. Groeger. "Understanding Driving: Applying Cognitive Psychology to a Complex Everyday Task." In: *Frontiers of Cognitive Science, Routledge Chapman* (2000).

[12] M. Grimm, K. Kroschel, H. Harris, C. Nass, B. Schuller, G. Rigoll, and T. Moosmayr. "On the necessity and feasibility of detecting a driver's emotional state while driving." In: *in Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction* 4738 (2007). Springer, Lisbon, Portugal, pp. 126–138.

[13] E. Wells-Parker, J. Ceminsky, V. Hallberg, R. W. Snow, G. Dunaway, S. Guiling, M. Williams, and B. Anderson. "An exploratory study of the relationship between road rage and crash experience in a representative sample of US drivers." In: *Accident Analysis and Prevention* 34 (2002), pp. 271–278.

[14] C. Liu, P. Rani, and N. Sarkar. "An empirical study of machine learning techniques for affect recognition in human-robot interaction." In: *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2005, pp. 2662–2667. DOI: 10.1109/IROS.2005.1545344.

[15] S. Basu, A. Bag, M. Aftabuddin, M. Mahadevappa, J. Mukherjee, and R. Guha. "Effects of emotion on physiological signals." In: *2016 IEEE Annual India Conference (INDICON)*. 2016, pp. 1–6. DOI: 10.1109/INDICON.2016.7839091.

[16] P. Ekman. "An Argument For Basic Emotions." In: *Cognition Emotion* 6 (May 1992), pp. 169–200. DOI: 10.1080/02699939208411068.

[17] S. P S and M. G S. "Emotion Models: A Review." In: *International Journal of Control Theory and Applications* 10 (Jan. 2017), pp. 651–657.

[18] D. Rubin and J. Talarico. "A Comparison of Dimensional Models of Emotion: Evidence from Emotions, Prototypical Events, Autobiographical Memories, and Words." In: *Memory (Hove, England)* 17 (Sept. 2009), pp. 802–8. DOI: 10.1080/09658210903130764.

[19] "EMOTION: Theory, Research, and Experience." In: *Theories of Emotion*. Ed. by R. Plutchik and H. Kellerman. Academic Press, 1980, p. ii.

[20] K. Takeda, H. Erdogan, J. H. Hansen, and H. Abut. *In-Vehicle Corpus and Signal Processing for Driver Behavior*. 1st. Springer Publishing Company, Incorporated, 2008. ISBN: 0387795812.

[21] J. A. Healey and R. W. Picard. "Detecting stress during real-world driving tasks using physiological sensors." In: *IEEE Transactions on Intelligent Transportation Systems* 6.2 (2005), pp. 156–166. DOI: 10.1109/TITS.2005.848368.

[22] Z. Ma, M. Mahmoud, P. Robinson, E. Dias, and L. Skrypchuk. "Automatic Detection of a Driver's Complex Mental States." In: *ICCSA*. 2017.

[23] N. Keshan, P. V. Parimi, and I. Bichindaritz. "Machine learning for stress detection from ECG signals in automobile drivers." In: *2015 IEEE International Conference on Big Data (Big Data)*. 2015, pp. 2661–2669. DOI: 10.1109/BigData.2015.7364066.

[24] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006. ISBN: 0387310738.

[25] J. Schmidhuber. "Deep learning in neural networks: An overview." In: *Neural Networks* 61 (Jan. 2015), pp. 85–117. ISSN: 0893-6080. DOI: 10.1016/j.neunet.2014.09.003.

[26] T. Thanapattheerakul, K. Mao, J. Amoranto, and J. H. Chan. "Emotion in a Century: A Review of Emotion Recognition." In: New York, NY, USA: Association for Computing Machinery, 2018. ISBN: 9781450365680. DOI: 10.1145/3291280.3291788.

[27] S. Steidl, A. Batliner, B. Schuller, and D. Seppi. "The hinterland of emotions: Facing the open-microphone challenge." In: *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. Sept. 2009, pp. 1–8. DOI: 10.1109/ACII.2009.5349499.

[28] C. Marechal, D. Mikołajewski, K. Tyburek, P. Prokopowicz, L. Bougueroua, C. Ancourt, and K. Wegrzyn-Wolska. "Survey on AI-Based Multimodal Methods for Emotion Detection." In: *High-Performance Modelling and Simulation for Big Data Applications: Selected Results of the COST Action IC1406 cHiPSet*. Springer International Publishing, 2019, pp. 307–324.

[29] M. Khezri, M. Firoozabadi, and A. R. Sharafat. "Reliable Emotion Recognition System Based on Dynamic Adaptive Fusion of Forehead Biopotentials and Physiological Signals." In: *Computer methods and programs in biomedicine* 122 (July 2015). DOI: 10.1016/j.cmpb.2015.07.006.

[30] Z. Yin, M. Zhao, Y. Wang, J. Yang, and J. Zhang. "Recognition of emotions using multimodal physiological signals and an ensemble deep learning model." In: *Computer Methods and Programs in Biomedicine* 140 (Mar. 2017), pp. 93–110. DOI: 10.1016/j.cmpb.2016.12.005.

[31] R. Subramanian, J. Wache, M. Khomami Abadi, R. Vieriu, S. Winkler, and N. Sebe. "ASCERTAIN: Emotion and Personality Recognition using Commercial Sensors." In: *IEEE Transactions on Affective Computing* PP (Nov. 2016), pp. 1–1. DOI: 10.1109/TAFFC.2016.2625250.

[32] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller. "Deep learning for time series classification: a review." In: *Data Mining and Knowledge Discovery* 33.4 (2019), pp. 917–963.

[33] E. Kanjo, E. M. Younis, and C. S. Ang. "Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection." In: *Information Fusion* 49 (2019), pp. 46–56.

[34] C. Li, Z. Bao, L. Li, and Z. Zhao. "Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition." In: *Information Processing Management* 57.3 (2020), p. 102185.

[35] J.-L. Qiu, X.-Y. Li, and K. Hu. "Correlated Attention Networks for Multimodal Emotion Recognition." In: Dec. 2018, pp. 2656–2660. DOI: 10.1109/BIBM.2018.8621129.

[36] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. "A Multimodal Database for Affect Recognition and Implicit Tagging." In: *IEEE Transactions on Affective Computing* 3.1 (2012), pp. 42–55. DOI: 10.1109/T-AFFC.2011.25.

[37] T. Song, G. Lu, and J. Yan. "Emotion Recognition Based on Physiological Signals Using Convolution Neural Networks." In: Feb. 2020, pp. 161–165. DOI: 10.1145/3383972.3384003.

[38] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. *Multimodal Machine Learning: A Survey and Taxonomy*. 2017. arXiv: 1705.09406 [cs.LG].

[39] V. Vielzeuf, A. Lechervy, S. Pateux, and F. Jurie. *Multi-Level Sensor Fusion with Deep Learning*. Nov. 2018.

[40] Y. R. Pandeya and J. Lee. "Deep learning-based late fusion of multimodal information for emotion classification of music video." In: *Multimedia Tools and Applications* 80.2 (2021), pp. 2887–2905.

[41] G. Caridakis, G. Castellano, L. Kessous, A. Raouzaiou, L. Malatesta, S. Asteriadis, and K. Karpouzis. "Multimodal emotion recognition from expressive faces, body gestures and speech." In: *Artificial Intelligence and Innovations 2007: from Theory to Applications*. Ed. by C. Boukis, A. Pnevmatikakis, and L. Polymenakos. Boston, MA: Springer US, 2007, pp. 375–388.

[42] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou. "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks." In: *IEEE Journal of Selected Topics in Signal Processing* 11.8 (2017), pp. 1301–1309. DOI: 10.1109/JSTSP.2017.2764438.

[43] M. Gjoreski, M. Ž. Gams, M. Luštrek, P. Genc, J. Garbas, and T. Hassan. "Machine Learning and End-to-End Deep Learning for Monitoring Driver Distractions From Physiological and Visual Signals." In: *IEEE Access* 8 (2020), pp. 70590–70603. DOI: 10.1109/ACCESS.2020.2986810.

[44] S. Taamneh, P. Tsiamyrtzis, M. Dcosta, P. Buddharaju, A. Khatri, M. Manser, T. Ferris, R. Wunderlich, and I. Pavlidis. "A multimodal dataset for various forms of distracted driving." In: *Scientific Data* 4.1 (2017), p. 170110.

[45] P. Bota, C. Wang, A. Fred, and H. Plácido da Silva. "A Review, Current Challenges, and Future Possibilities on Emotion Recognition Using Machine Learning and Physiological Signals." In: *IEEE Access* PP (Sept. 2019), pp. 1–1. DOI: 10.1109/ACCESS.2019.2944001.

[46] M. Dziezyc, M. Gjoreski, P. Kazienko, S. Saganowski, and M. Gams. "Can We Ditch Feature Engineering? End-to-End Deep Learning for Affect Recognition from Physiological Sensor Data." In: *Sensors* 20 (Nov. 2020), p. 6535. DOI: 10.3390/s20226535.

[47] Z. Wang, W. Yan, and T. Oates. "Time series classification from scratch with deep neural networks: A strong baseline." In: May 2017, pp. 1578–1585. DOI: 10.1109/IJCNN.2017.7966039.

[48] J. Wang, Z. Wang, J. Li, and J. Wu. "Multilevel Wavelet Decomposition Network for Interpretable Time Series Analysis." In: July 2018, pp. 2437–2446. DOI: 10.1145/3219819.3220060.

[49] J. Wang, Z. Wang, J. Li, and J. Wu. *Multilevel Wavelet Decomposition Network for Interpretable Time Series Analysis*. 2018. arXiv: 1806.08946 [cs.LG].

[50] E. Shelhamer, J. Long, and T. Darrell. *Fully Convolutional Networks for Semantic Segmentation*. 2016. arXiv: 1605.06211 [cs.CV].

[51] K. He, X. Zhang, S. Ren, and J. Sun. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[52] Z. Wang, W. Yan, and T. Oates. *Time Series Classification from Scratch with Deep Neural Networks: A Strong Baseline*. 2016. arXiv: 1611.06455 [cs.LG].

[53] O. Arriaga, M. Valdenegro, and P. Plöger. "Real-time Convolutional Neural Networks for Emotion and Gender Classification." In: (Oct. 2017).

[54] P. Burkert, F. Trier, M. Z. Afzal, A. Dengel, and M. Liwicki. *DeXpression: Deep Convolutional Neural Network for Expression Recognition*. 2016. arXiv: 1509.05371 [cs.CV].

[55] F. Chollet. *Xception: Deep Learning with Depthwise Separable Convolutions*. 2017. arXiv: 1610.02357 [cs.CV].

[56] S. Shafaei, T. Hacizade, and A. Knoll. "Integration of Driver Behavior into Emotion Recognition Systems: A Preliminary Study on Steering Wheel and Vehicle Acceleration." In: June 2019, pp. 386–401. ISBN: 978-3-030-21073-1. DOI: 10.1007/978-3-030-21074-8_32.

[57] J. Deffenbacher, R. Lynch, E. Oetting, and R. Swaim. "The Driving Anger Expression Inventory: A measure of how people express their anger on the road." In: *Behaviour research and therapy* 40 (July 2002), pp. 717–37. DOI: 10.1016/S0005-7967(01)00063-8.