Multimodal Emotion Recognition using facial expression and other physiological condition

Andreas Mathisen and Edvard Halsteinli Unsvåg

Thesis submitted for the degree of
Master in Informatics: Programming and System Architecture
60 credits

Department of Informatics
The Faculty of Mathematics and Natural Sciences

UNIVERSITY OF OSLO

Spring 2022

Abstract

The task of emotion recognition has been attempted solved predominantly by unimodal approaches, where facial emotion recognition is the most frequently used modality. However, regardless of good results in facial emotion recognition, there is still a need for further efforts to improve the quality of emotion recognition methods. Multimodal emotion recognition is an emerging field within emotion recognition set to address this challenging task by combining unimodal approaches, motivated by the fact that humans detect emotions in a multimodal matter. In the field of multimodal emotion recognition, several combinations have been explored; facial expressions and text, facial expressions and speech, and facial expressions and physiological signals.

In this thesis, we explored the field of multimodal emotion recognition by investigating the combination of facial expressions and physiological signals. An overview of different approaches, combinations, and traits was established through a literature study of related work in the field. We experienced a lack of usable datasets from the literature study, and a decision was made to provide true labels to an unlabeled dataset containing facial expressions and physiological signals. Several machine learning models, along with a human survey, were used to label the dataset. Additionally, a preliminary experiment regarding the correlation between facial expressions and blood volume pulse was conducted using spatio-temporal networks, serving as a motivation for the subsequent experiments. Through an experiment trying to classify six basic emotions plus neutral with a 3D-CNN, it was found that a multimodal model using facial expressions and physiological signals resulted in a slight improvement compared to only using facial expressions. These findings suggest that there is potential for improving an emotion recognition model by including physiological signals in a multimodal model, but further research should be conducted to explore how the physiological signals are most effectively utilized.

Acknowledgments

The completion of this thesis would not have been possible without the many people who have helped us along the way. Firstly, we would like to thank our supervisors, Michael Riegler and Pål Halvorsen, for their outstanding guidance and support.

We would also like to thank Elise Fehn Unsvåg for providing us with guidance regarding the thesis structure and helping explain some key concepts.

We would also like to thank Henrik Svoren for providing us with the collection of a multimodal dataset. Henrik was also an essential factor concerning our methodology, with us following some of his suggestions regarding future work. Additionally we would like to thank the providers of CK+ and FER13, who have made significant efforts to collect and annotate data. By sharing these datasets they allowed us to make contributions to the field of multimodal emotion recognition.

Lastly, we want to thank each other for motivating one another when the other was down and for a great teamwork throughout the entire process.

List of Figures

List of Tables

List of Equations

Contents

Chapter 1

Introduction

Multimodal Emotion recognition (MER) is defined as the process of identifying human emotion using several modalities, motivated by the fact that humans detect emotions in a multimodal matter. Understanding emotions is an essential part of being human. Along with the growth in “internet-of-things” (IoT) and wearable technology, we are witnessing growth in the field of MER, which motivates research in the field on how advanced technology can assist in solving the task of emotion recognition. This thesis will focus on how such technology can assist in automatically detecting emotions by exploring the combination of facial expressions and physiological signals.

1.1 Background And motivation

The field of emotion recognition has been dominated by unimodal approaches, meaning approaches that consider one modality. The most popular unimodal emotion recognition approaches uses facial expressions, which tries to detect emotion solely based on facial expressions. However, recent advancements with the combination of several modalities have shown promising results. MER is defined as the process of identifying human emotion by using several modalities. The field of MER is motivated by the fact that humans detect emotions in a multimodal matter by combining the processing of modalities like facial expressions, posture, and prosody to recognize a person’s emotions; therefore, technology should do the same. As a result of the growth in "internet-of-things," wearable technology, and data becoming more available, we are witnessing growth in the field of multimodal emotion recognition. This motivates research in the field on how advanced technology can assist in solving the issue of emotion recognition and how to use such technology in a natural setting.

There is an increasing amount of research covering the multimodal approach to emotion recognition using Machine Learning. While most of the early studies focused on combining facial expressions with speech, considering this was data easily collected from video, the research field has expanded beyond this. Over the last few years, other combinations have been explored using facial expressions, text, speech, and physiological signals. While methods within these combinations perform relatively well, the field, in general, is still relatively fresh and in need of improvements in the quality of emotion recognition. There exists research that investigates the impact of using facial expressions and physiological signals, however, little research combines them in a multimodal approach. This is motivating regarding research of the subject, and the focus of this thesis will be on predicting emotions based on the combination of facial expressions and physiological signals. Emotion recognition is of interest to many actors, and the motivation of this thesis is to motivate the usage of facial expressions, physiological signals, and, in general, contribute to the field of study.

1.2 Goals And Research Questions

Goal *Investigate the effect of multimodality in emotion recognition with the use of facial expressions and physiological signals*

The goal of this thesis is to investigate how a multimodal approach can help recognize emotions with the combination of facial expressions and physiological conditions. A theoretical and practical approach will be adopted to achieve this goal, with a literature study and experiments. This objective is divided into the research questions below. In addition, we will include the objective of creating a labeled multimodal dataset.

Research question 1 *What does the literature suggest as promising approaches to MER using facial expressions and physiological signals?*

A review of related work will be conducted to obtain an overview of well-performing approaches within multimodal emotion recognition. The findings from this review will provide us with an understanding of the field and will be used in experimenting with the combination of facial expressions and physiological signals, which is the focus of the second and third research question.

Objective 1 *Provide ground truths to the multimodal dataset provided by Svoren (2020), using a facial emotion recognition model and human raters.*

To achieve our goal of multimodal emotion recognition, a labeled dataset consisting of both facial expressions and physiological signals was created. This dataset will serve as a contribution to the dataset of Svoren (2020), with the inclusion of true labels and preprocessing of the data.

Research question 2 *How are physiological signals related to facial expressions?*

To serve as motivation for the inclusion of physiological signals, an experiment trying to predict Blood Volume Pulse (BVP) values using facial expressions were conducted.

Research question 3 *What are the effects of including physiological signals in multimodal emotion recognition?*

A unimodal facial emotion recognition (FER) model will be implemented to serve as a baseline for comparison. Then a unimodal physiological emotion recognition classifier will be implemented and combined with the facial emotion recognition model. Lastly, a comparison between the usage of a FER model and a multimodal approach will be conducted.

1.3 Scope And Limitations

Within the area of multimodal emotion recognition, there are many possibilities to explore. However, due to time limitations, we are forced to limit the scope of the thesis to only exploring a tiny portion of the field. Firstly, there are many ways in which we could design our emotion recognition models to perform multimodal emotion recognition. In this thesis, we limit ourselves to exploring a single approach for MER, trying to classify six basic emotions plus neutral. With our approach drawing inspiration from the state-of-the-art literature, we perform a promising exploration of research question 3, exploring the effects of incorporating physiological signals in multimodal emotion recognition. However, with countless combinations of the structure of models, datasets to be used, what emotions to classify, and the number of modalities, several areas is still yet to be explored.

Secondly, several approaches could have been explored regarding the labeling of Toadstool. This thesis uses three classification models, two datasets, and some validation provided by human raters. Moreover, an improved approach to labeling the Toadstool dataset should be further explored due to the fact that high-quality true labels are essential for the performance of an emotion recognition model. As stated, many possibilities and variations of the process of multimodal emotion recognition could potentially yield interesting results but will remain outside the scope of this thesis. Some of these, as well as other potential ideas for future work, will be discussed in chapter 9.

1.4 Research Methods

In order to answer the research questions and accomplish the research goal, several methodologies have been used. Firstly, a literature study regarding the most promising methods in multimodal emotion recognition was conducted to answer Research Question 1. Further, to accomplish our objective of labeling Toadstool, a human survey was conducted to provide labels and combined with the findings from the literature. Secondly, the work conducted in this thesis follows an experimental research strategy, where the value of combining facial expressions with physiological signals is investigated. The work in the experiments is based on the findings from the literature review and Objective 1 to provide answers to Research Questions 2 and 3. The results from the experiment regarding Research Question 2 were further used to answer Research Question 3. The results of the final experiments were quantitatively analyzed by evaluating the impacting factors of combining physiological data with facial expressions and investigating the distribution of correctly and incorrectly predicted emotions. As a result of the experimental strategy, a hypothesis regarding the effects of incorporating physiological signals in multimodal emotion recognition will be tested. However, the contributions of this thesis will extend beyond a proved or disproved hypothesis, providing insight and data to the field of multimodal emotion recognition. (Holz et al. 2006)

1.5 Main Contributions

The work in this thesis will mainly contribute to a deeper insight into the field of multimodal emotion recognition that can be used to build upon for other researchers. Furthermore, with an increased amount of research in this area, we will hopefully contribute to improving the methods for recognizing emotions with the use of machine learning. More specifically, the research conducted in this thesis will contribute with the following:

C1 A literature review on of the field of multimodal emotion recognition

C2 Developing an improved version of Toadstool, with the inclusion of true labels and easier use of the dataset (Mathisen and Unsvåg 2022b)

C3 A user survey regarding the labelling of emotions, contributing to understanding the difficulties of emotion recognition

C4 Experiments with facial expressions and physiological signals in multimodal emotion recognition to investigate the effect of the inclusion of physiological signals (Mathisen and Unsvåg 2022a)

1.6 Thesis Outline

Chapter 2 introduces the relevant theoretical concepts and methods that are used in this thesis, or in related work

Chapter 3 provides an overview of the research conducted in the field of multimodal emotion recognition, as well as methods within facial emotion, recognition, physiological-based emotion recognition (PER), and MER.

Chapter 4 presents the datasets used to train and test the implemented emotion recognition classifiers.

Chapter 5 presents a preliminary experiment where spatio-temporal networks are used to predict BVP based on facial expressions.

Chapter 6 presents Toadstool 2.0, a developed dataset serving as a contribution to Toadstool. The survey regarding human validation will be presented. Further, synchronization and preprocessing of the data, different labeling approaches and a discussing regarding the process is presented.

Chapter 7 includes the experimental setup and describes the architecture of the classifiers developed and the experiments conducted to measure the impact of the inclusion of physiological signals.

Chapter 8 addresses the research questions and objective with an evaluation and discussion of the experimental results

Chapter 9 concludes the thesis by summarizing the research contributions along with suggestions for potential future work

Chapter 2

Background Theory

This chapter covers the theory within the fields of Emotions, Machine Learning, Affective Computing and Multimodal Emotion Recognition that is relevant for this thesis and for related work on multimodal emotion recognition. Additionally, metrics and the tools and libraries used in the thesis are described.

2.1 Theories Of Emotions

Considering that much of the work in this thesis is concerned with emotion recognition in the context of computing and machine learning, touching upon the field of emotions seems sensible. Numerous theorists, philosophers, and computer scientists have tried to answer "what is an emotion?". However, defining a universal definition of emotion is a demanding task, and a widely acknowledged definition is still to be defined. As this thesis focuses on machine learning, defining a definition of emotion is particularly important, considering the target criteria rely on the success of detecting a given emotion. This section will review the most acclaimed takes on emotions, namely that of a discrete categorization of emotions and composing emotion in continuous dimensions.

2.1.1 Discrete Categories

Discrete emotion theory claims that emotions can be divided into a small number of core emotions. Further, that these emotions are biologically determined emotional responses and are the same across all cultures and ethnic backgrounds.

Looking back on the theory of emotions, numerous people have proposed ways of theorizing the subject. Since ancient times, back in the Roman empire, Marcus Cicero thought of emotions and feelings. The roman philosopher organized emotions into four basic categories: Fear, Pain, Lust, and Pleasure (Cicero and Graver (2002)). Further, Darwin (1872) proposed that emotions have an evolutionary history and are shared across cultures and treated emotions as separate discrete entities, such as anger and fear. Paul Ekman (1992), an acknowledged American psychologist, continued the work of Darwin and argued that emotions are shared between cultures, thus, able to be universally recognized. Ekman described emotions as discrete, proposing to categorize six basic emotions: Happy, Sad, Anger, Fear, Surprise, and Disgust. However, the concept of basic emotions is controversial within psychology, as is reflected in the ongoing debate about which emotions should be included. Considering the theory varies within a broad number, regarding which categories to include, there is a basis for skepticism (Moerland, Broekens and Jonker 2018).

One category not included in Ekman's basic emotion is that of neutral. Although neutral is not categorized as a basic emotion, the neutral category is rapidly used in emotion recognition. However, the inclusion of the neutral category comes with a couple of challenges. Firstly, the difference in people's "neutral face" may vary, resulting in difficulties with dividing neutral and other emotions. Secondly, several researchers believe it is not possible to feel neutral because people are constantly feeling something, making the category dismissible (Gasper, Spencer and Hu 2019).

In all the theories of emotion previously mentioned, human emotional experiences are described in words. However, a discrete qualification of emotion can present difficulties since complex mixed emotions can be difficult to interpret precisely. Additionally, different individuals and cultures may describe a similar experience with different words. In order to overcome these difficulties, many authors have adopted the concept of continuous multi-dimensional space models, such as the circumplex model of Russell (1980). In a continuous multi-dimensional space model, emotions are measured along a defined axis, thus, simplifying the process of comparison and emotion discrimination (Bota et al. 2019).

2.1.2 Continuous Categories

Dimensional emotion theory stands in contrast to theories of basic emotions, which post that a discrete and independent neural system subserves every emotion. In contrast, dimensional emotion theory assumes an underlying affective space. Dimensional emotion theory derives from the belief that mixed emotions can be challenging to put into categories, following that emotional states should instead be identified along different dimensions. Proponents of this approach argue that most categorical theorists present emotions as a structured collection of distinct entities and thus fail to capture the intuitions concerning the similarities and differences among emotions.

Some emotions are commonly viewed as opposites, such as joy and sorrow or fear and anger (Plutchik 1982). Once we consider the similarities and differences among emotions, more specific questions inevitably demand our attention. Looking at the mentioned examples of joy vs. sorrow and fear vs. anger, the latter are not opposite of each other in the same way. Joy is a pleasant feeling, whereas fear and anger are both unpleasant. This very opposition in similarity implies that a dimension can be arranged to make more specific assumptions regarding whether emotions are similar to each other.

Several dimensional models have been proposed with a varying number of dimensions. The first dimensional approach dates back to Wilhelm Wundt, who describes momentary emotions as a single point in a three-dimensional space (Reisenzein 1992). This 'emotional space' is spanned by the axes of pleasure-displeasure, excitement-inhibition, and tension-relaxation. At the end of the 1970s, Russell (1980) postulated a two-dimensional model, namely the circumplex model. In this model, active states are represented as discrete points in a two-dimensional space, spanned by the axes valence and arousal. Dimensional models have difficulty separating emotion categories such as anger and disgust, which is a common critique of the theory.

2.2 Machine Learning

Machine learning (ML) is a field within computer science that enables systems and programs to learn through experience and data. Machine learning is commonly split into three approaches; Supervised learning, unsupervised learning, and reinforcement learning. In this thesis, supervised learning will act as the primary approach. Supervised learning is a type of machine learning where algorithms learn from labeled data to be further used in predicting new and unseen data. Methods used for emotion recognition mainly follow the approach of supervised learning. This section introduces the terminology and concepts of the common machine learning classifiers used for Emotion recognition, along with some general terms used in the field.

2.2.1 General ML terms

Epoch

Epoch is a term that indicates a models passage through all the training data, one time.

Early stopping

Early stopping is a method to allow a model to train long enough to learn, without overfitting on the training data. When providing early stopping, the model will stop the training before it has seen all the data, if it has not improved, in compliance with a performance metric, for a certain amount of epochs.

2.2.2 Supervised Learning

Supervised learning is the process of training a learning algorithm on data that has been labeled. In other words, the desired output associated with each sample in the training data is known. When given a sample, the algorithm tries to predict the correct output by looking at the training process. Following this, the learning algorithm adjusts its internal weights based on how close the prediction was to the desired output. This process is repeated with each new sample, letting the learning algorithm gradually improve with each iteration. This iterative optimization algorithm is commonly referred to as gradient descent. When the algorithm stops improving (i.e., reaching a local minimum in terms of gradient descent), the data runs out, or some other specified threshold is reached, the training ends. Common types of supervised learning include classification algorithms, where the output is a discrete category, and regression algorithms, where the output is a value within a range. Supervised learning is the most popular approach to emotion recognition.

Ensemble Learning

Ensemble learning is a subfield of supervised learning using multiple learning algorithms. The intuition behind ensemble learning is that one may achieve better predictive performance by using multiple learning algorithms. A common ensemble learning approach is soft voting. Soft voting collects the sum of the predicted probabilities for all classes, and further uses it to label according to the class with the largest sum probability. Figure 2.1 displays an ensemble learning approach, combining three classifiers.

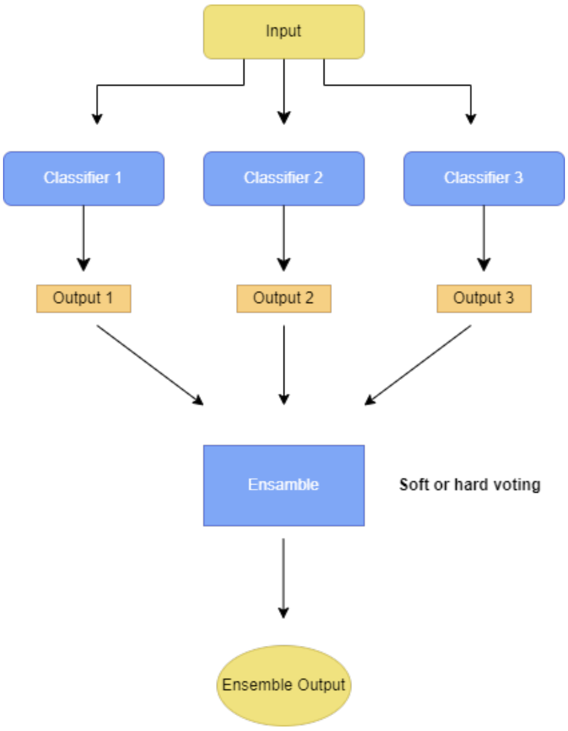


Figure 2.1: Ensemble Learning combining three classifiers

2.2.3 Support Vector Machines

Support Vector Machines (SVM) are supervised learning models often used to classify emotions from face images. In the problem of classifying emotions from face images, the goal is to find a hyperplane that differentiates the classes by maximizing the distance to the nearest training data point of each class. The training data instances are represented as coordinates in an n-dimensional space, where n equals the number of features. Hyperplanes are decision boundaries that help classify the data points, and the dimensions of the hyperplane depend upon the number of features. With the number of input features being 2, then the hyperplane is a line. The data points closest to the hyperplane are called support vectors, and the distance from the support vectors to the hyperplane is called the margin. Subsequently, the goal of SVM is to find the optimal hyperplane with the largest margin possible. Figure 2.2 displays plotted training data that is linearly separable. However, with the assistance of nonlinear kernel functions, it is possible for input data to be transformed into a high-dimensional feature space in which the input data become linearly separable and classifiable. (E. Unsvåg and Gambäck 2018).

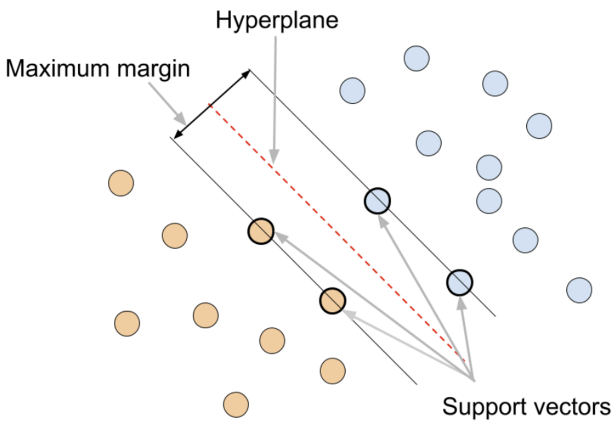


Figure 2.2: SVM hyperplane separation (©E. Unsvåg and Gambäck 2018)

2.2.4 Generalization

Generalization in machine learning refers to how well the concepts learned by a machine learning model applies to a specific example not seen by the model when it was learning. If a model cannot generalize, the model is said to have overfitted on the training data. Overfitting occurs when a model learns the details in the training data to the extent that it negatively impacts the learning of new data. Consequently, a model is learning details in the training data that do not apply to the new data, harming the ability of the model to generalize. Overfitting, along with underfitting, is commonly seen as the two most significant causes of poor performance in machine learning algorithms (Brownlee 2019).

Furthermore, when generalization is the goal, a vast amount of training data is needed, which is challenging, regarding the increase in training time and the collection of the data. Followingly, Poyiadzi et al. (2021) studied the effect of using different age-groups for training FER models because aging affects facial features such as wrinkles. Their study showed that with the inclusion of age-groups in training, an increase in performance when training on unseen age-groups tends to occur. However, most approaches in emotion recognition focus on training a generic model to generalize across age, culture, and gender, along with other categories. Therefore, following the way of most approaches in the field, this thesis will also use the principle of creating generic models.

2.2.5 Deep Learning

Deep learning is a subfield of machine learning that has seen considerable growth in popularity and usefulness in recent years. Due to factors of variation, a major difficulty in real-world artificial intelligence applications is to extract high-level abstract features from raw data. A crucial step is disentangling the factors of variation and discarding the features we do not care about. Deep learning solves this difficulty in feature learning by constructing high-level abstract representations from a combination of simpler representations at different deep layers. Exemplified by its application to image processing, where the lower layers may identify simple representations such as edges. Likewise, higher layers use these simple representations to identify concepts relevant to humans, such as digits, letters, or faces. (Goodfellow, Yoshua Bengio and Courville 2016)

Furthermore, Artificial Neural Networks (ANNs) are networks inspired by a simplification of the biological brain and are a central concept in deep learning. Looking at Neural Networks (NN), the first networks used in the field and the simplest are the feedforward neural networks. We will use these networks to explain some fundamental concepts of neural networks before introducing more advanced networks.

Feedforward networks are directed networks, meaning the information provided to the network only moves in one direction, forward. Subsequently, feedforward networks do not contain any cycles. The most basic network is a single-layer perceptron that can learn linear functions, consisting of an input layer and an output layer. Further, we have a multilayer perceptron (MLP) containing at least one hidden layer, and compared to a single layer perceptron, an MLP can learn linear and nonlinear functions. Figure 2.3 displays a simple multilayer perceptron with one hidden layer and two output classes. The neurons are the basic units of a neural network, which receive input and compute an output. When the NN is provided with inputs, the output is defined by an activation function, being a function that can introduce non-linearity to the output. Further, the way MLPs learn is through the backpropagation step. The backpropagation algorithm adjusts the weights iteratively in the network until the output is "correct" to one's satisfaction. Further, a loss function measures the output with the predicted output, and the error is "propagated" back to the previous layer, where the weights are adjusted according to the error. The goal is to minimize the outcome of the loss function. To achieve this, the gradient descent algorithm is used. Gradient descent is an optimization algorithm that iteratively moves towards finding a local minimum of a function. (E. Unsvåg and Gambäck 2018).

2.2.6 Convolutional Neural Networks

Convolutional neural networks (CNN) are feedforward networks that have proven groundbreaking results in machine learning problems over the last decade, especially the applications that deal with image data. The reason for this is that CNNs are designed to work with grid-structured inputs, that have strong spatial dependencies in the local regions of the grid. Local dependencies have a high resemblance to an image, where adjacent pixels often have similar color values. When we say that CNNs are designed to work with grid-structured inputs, this is thanks to the architecture of the CNN, which generally consists of a convolutional layer,

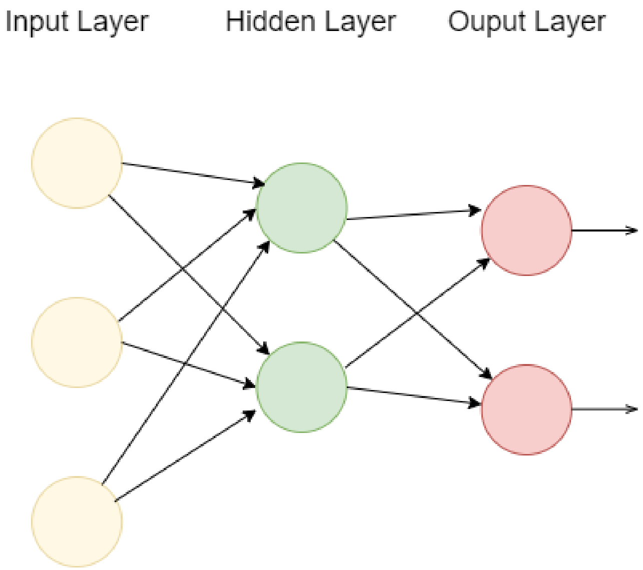


Figure 2.3: Feedforward Neural Network Architecture

a pooling layer, and a fully connected layer. The following sub-subsections will explain the most important aspects of a CNN.

Convolutional Layers

The convolutional layer determines the output of neurons that are connected to local regions of the input through the calculation of the scalar product between their weights and the region connected to the input volume (O'Shea and Nash 2015). The kernel creates these local regions, which slide across the image and divide it into smaller parts. By doing so, neurons in the layer are only connected to one part of the image, in contrast to regular neural networks where all neurons in a layer are connected to all neurons in the next layer. Applying this method reduces the complexity of the computations and ensures that spatial dependencies in local regions of the image are kept. The calculation of the scalar product between the weights and the region-specific input from the image creates a feature map or activation map. This feature map represents the "relevant" features extracted from the image, or in terms of a activation map, which parts of the image are activated.

Pooling Layers

The pooling layer will further downscale the image, similar to the convolutional layer applying a kernel that slides across the feature maps and performs max-pooling or average-pooling on the values. The max-pooling operation will reduce a window equal to the size of the kernel in the feature map to the highest value in that window. Average-pooling will similarly take the average value of the window as a representation of that window. By doing so, the dimensionality is reduced, but the higher-level features are kept. Pooling can be compared to lowering the resolution of an image, where the details of the image are lost, but one can still tell what the image consists of.

Fully Connected Layers

The fully connected layer will act just like in regular neural networks and attempt to produce class scores from the activation's in previous layers, which again are used for the classification of the original image (O'Shea and Nash 2015).

Dropout Layer

Dropout is a regularization method that helps with the problem of overfitting. Hinton et al. (2012) proposed dropout layers as an approach, where it was used on each of the fully connected layers before output. Using it on the fully connected layers, not the convolutional layers, has become the most used configuration. The method consists of "dropping" several outputs, often meaning setting the value to zero with a certain probability. The dropout layer will make the training process noisier and force outputs to take less or more responsibility for the inputs, with a certain probability.

Batch Normalization

Batch normalization is a method used in ANNs to make the network faster and more stable. The method provides normalization of the input to the layers, helping to standardize the inputs to each layer. Batch normalization stabilizes the learning process and reduces the number of training epochs required.

Rectified Linear Activation Function

Rectified linear activation function, commonly known as ReLu, is a linear function that will output the input directly if it is positive; otherwise, it will output zero. ReLu can remove un-activated units (units with zero output) in a randomly initialized network. ReLu is one of the most popular activation functions for deep neural networks.

2.2.7 3-Dimensional Convolutional Neural Networks

3-Dimensional Convolutional Neural Networks are a variation of CNNs that have proven successful in facial emotion recognition in videos. When a 2D-CNN is applied to a video sequence, it looks at each image one at a time, and the output is a 2D tensor. However, when a 3D convolution is applied to a video sequence, the output is a 3D tensor, with the third dimension being time. Therefore, a 3D-CNN has the ability to preserve the temporal aspect of a video sequence. Haddad, Lézoray and Hamel (2020) explains how in video contexts, facial expressions do not manifest themselves instantly, but instead, they build up gradually across time until they reach their peak. Following this, they believe a static approach is uninterpretable, considering the result in the predictions can vary across the frames. A 3D-CNN solves this issue by analyzing and predicting an emotion based on all the frames in a sequence.

2.2.8 Recurrent Neural Networks

As CNNs are neural networks built for data with spatial dependencies, Recurrent neural networks (RNNs) are built for data with temporal dependencies such as sequences of text, speech, and images (video). The RNN does this by inputting data from earlier time steps into the current time step, i.e., in a computer vision problem of classifying events in a video, and the RNN would not treat every frame in the video individually but rather look at the current frame in context with the values generated from previous frames. This results in a better representation of how the video evolves over time. In the context of a facial emotion recognition problem which has been discussed earlier, Kahou et al. (2015) introduce in their paper "Recurrent Neural Networks for Emotion Recognition in Video" an interesting approach; CNN-RNN architecture. The approach contains two steps; (1) an CNN is trained to classify static images containing emotions. (2) train an RNN on the higher layer representation of the CNN inferred from individual frames to predict a single emotion for the entire video. This approach gives us a spatio-temporal evolution of the facial emotions, that is, emotions are learned and detected both in the form of static images (spatial dependent data) and over time in the video (temporal dependent data), giving a more complete picture.

2.2.9 Long Short-Term Memory Neural Networks

Long Short-Term Memory (LSTM) networks are particular types of RNNs that can address the shortcomings of regular RNNs. In the study of Y. Bengio, Simard and Frasconi (1994), they showed how regular RNNs, using gradient descent, performed poorly for tasks involving long-term dependencies. Further, the LSTM architecture, introduced by Hochreiter and Schmidhuber (1997), was explicitly designed to overcome this. LSTM networks consist of three cells; a forget gate, an input gate, and an output gate. The gates has the ability to remove or add information to the memory cell state and store temporal information. The forget gate controls which information to forget or store from the timestep, for later use. Further, the input gate protects the memory state from being disturbed by irrelevant information, and the output gate avoids storing irrelevant information in the memory state. LSTM networks have shown great success in solving tasks where the goal is to capture either long-term or temporal dependencies. A figure of a CNN-LSTM architecture is presented in 5.1 (E. Unsvåg and Gambäck 2018).

2.3 Affective Computing

During social interaction, humans employ rich emotional communication channels by modulating their speech utterances, facial expressions, or body gestures. Humans also rely on emotional cues to resolve the semantics of received messages. Interestingly, humans also communicate emotional information when interacting with machines. However, machines have conventionally been utterly oblivious to emotional information from humans. This reality is changing with the advent of affective computing (Al Osman and Falk (2017)). Computers are beginning to acquire the ability to express and recognize affect and may soon be given the ability to "have emotions" (Picard 1995). Picard calls affective computing "computing that relates to, arises from, or influences emotions ." In other words, the goal of affective computing is to recognize, interpret and process human experiences and emotions.

Detecting or recognizing emotional information usually begins with passive sensors that can capture a person's physical state or behavior. For instance, a microphone may capture speech and tone of voice, and a video camera might capture facial expressions, body posture, and gestures. This section introduces the most common concepts of affective computing.

2.3.1 Affect vs Emotion

Before we proceed, a clarification of a potential source of confusion is needed. The terms affect and emotion can have different meanings in various fields. For instance, according to Shouse (2005), emotion refers to the display of a feeling, whether it is genuine or feigned. However, affect is a "non-conscious experience of intensity." Some psychologists evaluate affect as the experience of emotion. In this thesis, we consider the terms emotion and affect to be synonymous since a sizable amount of work in affective computing use them interchangeably (Bota et al. 2019).

2.3.2 The Circumplex Model Of Emotion

The circumplex model of emotion is a model that distributes emotions in a two-dimensional circular space, containing arousal and valence dimensions, developed by Russell (1980). Russel's valence-arousal model aims to quantify emotions, with arousal and valence forming the horizontal and vertical axes, respectively. Valence refers to the positive and negative span of emotion, whereas arousal refers to the intensity of emotion. Figure 2.4 shows Russels two-dimensional model.

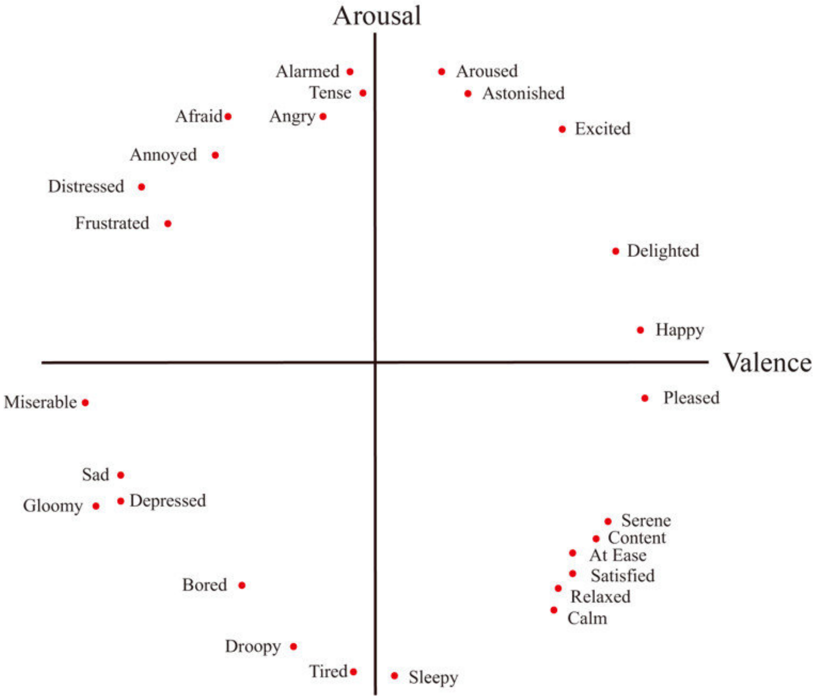


Figure 2.4: The circumplex model of emotion

2.3.3 Emotion Elicitation

Emotion elicitation is the process of triggering emotions to obtain ground truths. One of the biggest challenges when trying to achieve emotion recognition is obtaining ground truth data. How the emotions of the subjects are provoked (elicit) plays a role in the spectrum you will get the emotions and the intensity of each emotion. Due to the high subjectivity and variability in emotion elicitation, it is essential to use a set of pre-validated emotional stimuli to ensure the expression of a wide spectrum, with high intensity. If one does not manage to collect data from the entire spectrum, one faces the risk of classifying a lot of extreme points occurs.

Commonly used methods to elicit emotions are music videos, films, pictures, sound, and virtual reality (VR). Subsequently, the question of "what method to use" rises. The use of images as elicitation material presents the advantages of being user-friendly, low cost, easy and fast to execute in a laboratory. However, images might not be enough to evoke impactful, strong-lasting emotions, or enough to be consciously perceived by the user and physiologically observable. On the other hand, although simple and low cost, music or music videos might be constrained to the evocation of a limited range of positive-negative emotions, highly correlated with the subjects' music taste and the memories it invokes. Thus, films or short-duration audiovisual video clips are the most applied methodology in emotion recognition and have shown to be the most reliable material for emotion elicitation. In this thesis, playing a video game is used as the elicitation method. A method rarely used in the field.

2.4 Facial Emotion Recognition

This chapter will briefly introduce some of the methods and concepts within facial emotion recognition. Firstly a brief overview of the field in general will be presented. Then common approaches to preprocessing and commonly used features for facial emotion recognition will be presented.

2.4.1 Facial Emotion Recognition

Facial emotion recognition is the process of detecting human emotions from facial expressions. Facial emotions are essential factors in human communication that help us understand the intentions of others. In general, people infer other people's emotional states, such as joy, sadness, and anger, commonly using facial expressions. Over the past decades, facial emotion recognition has been gaining increased attention and has established itself as one of the most active fields within affective computing. However, accurate and robust FER by computer models remains challenging due to the heterogeneity of human faces and variations in images, such as different facial poses and lighting. Among the techniques used for FER, deep learning models, especially Convolutional Neural Networks (CNNs), have shown great potential. With respect to the theories of emotions commonly used in computer science, facial emotion recognition aims to map a facial emotion to a category or dimension.

While images have been the most common data source in FER, the use of video is starting to receive more attention. With the emergence of deep learning techniques, predicting dynamic facial emotion expressions has become more interesting. Several deep learning techniques can be considered for sequential data, with the most prominent ones being RNN and LSTM. The work of J. Li, X. Li and D. He (2019) combined a classic 2D CNN with LSTM to cope with the temporal aspect of emotion recognition in videos. In the mentioned approach, the CNN extracts the features over individual frames and passes them to an LSTM, which encodes the temporal dynamics. Lastly, few works have been led on the use of 3D-CNN; however, CNN's 3D kernels may have a superior ability to extract spatio-temporal features within video frames (Haddad, Lézoray and Hamel (2020)).

2.4.2 Preprocessing

Preprocessing is a fundamental step of image processing in emotion recognition. Before extracting features, preprocessing techniques are needed to extract significant features from the images. Ninu Sreedharan (2018) described that preprocessing can improve the image features in order to control the noise and redundant information for adapting the feature extraction step. When it comes to facial images, the images of a face change with variations, such as facial expression, pose, and illumination conditions. For a machine to understand and make sense of such facial images, the preprocessing step becomes essential. A common preprocessing pipeline often consists of the steps of face detection, resizing, data augmentation, and normalization. Face detection refers to detecting a face in an image to remove irrelevant features from the image. Resizing is the process of changing the size of an image in order for it to fit the desired input shape. In data augmentation, slightly modified copies of the image are added to diminish the effect of overfitting. Examples of common data augmentation techniques include flipping, rotation, and zooming. Normalization is a preprocessing method used to reduce variations of the face images, like illumination, to achieve an improved face image. A common normalization approach is grayscaling, converting the image from a three-channel (Red, green and blue) image into a single channel image, providing a more general range of pixel values. Lastly, both resizing and grayscaling contributes to more efficient training of machine learning models.

2.4.3 Features for Facial Emotion Recognition

Feature extraction is the next step after preprocessing in the process of FER. In image processing, feature extraction is a significant stage, whereas it extracts implicit data from graphical data that further can be used as input to a classification model. Following, when the amount of training data is large enough, the difference in performance between models decreases. The main goal is to extract only the most important and descriptive pieces of information. The features chosen to employ in the methods will followingly become the distinguishing impact on performance. This section describes common types of features used in FER, based on a state-of-the-art review of existing research on facial emotion recognition (Canal et al. (2022)).

Histogram of Oriented Gradients

Histogram of Oriented Gradients is a feature descriptor for the purpose of object detection, and is mainly utilized for face and image detection. Simply putting it, the histogram of oriented gradients calculates the gradient in each pixel. And when there is a sharp change in intensity in the image, the magnitude of the gradient increases.

Local Binary Pattern

Local Binary Pattern is a simple and efficient texture operator which labels the pixels of an image by making comparisons between each cell's pixels and their eight neighbors to build a binary number. One of the most important properties of LBP, and a major reason why it is being used, is its robustness to monotonic grayscale changes caused by, for example, illumination.

CNNs

Despite the success of traditional ways of extracting features, recent development in convolutional neural networks has demonstrated significant success in automatically learning features. CNN has yielded impressive performance with its ability to extract undefined features from the training database, compared to traditional approaches where features are defined by hand. When using a CNN, it extracts shift-invariant local features from input images based on the concept of the local receptive field, shared weight, and spatial subsampling (Cheng et al. 2019).

Scale-invariant feature transform

Scale-invariant feature transform (SIFT) is an algorithm used to detect, describe and match local features in an image. SIFT provides less accurate results compared to CNN's, however they require fewer data to generalize with high accuracy. Along with the fact that SIFT requires a small amount of data, the algorithm is invariant to image scale and rotation and performs well with changes in illumination.

Facial Action Coding System

Facial Action Coding System (FACS) is a coding system created by P. Ekman and Friesen (1978), measuring the contraction and relaxation of facial muscles with degrees of intensity and deconstructing it into action units (AU). For instance, a chin raise is categorized as AU17, and a jaw drop is AU26. Further, AU's is combined to decide on the given emotion. Table 2.1 and 2.2 present the different AUs, as well as how it is used to provide emotion labels. In general, regarding the use of FACS, when trying to conduct FER using FACS, one needs to use a specific dataset with facial images cataloged by FACS experts. Resulting in the feature extraction process being provided by humans.

AU	Name	AU	Name
1	Inner Brow Raiser	13	Cheek Puller
12	Lip Corner Puller	20	Lip Stretcher
17	Chin Raiser	21	Neck Tightener
26	Jaw Drop	28	Lip Suck

Table 2.1: Example of 8 Action Units

Emotion	Action Units
Happy	6 + 12
Sad	1 + 4 + 15
Surprise	1 + 2 + 26
Fear	1 + 2 + 4 + 5+ 7 +20 + 26
Anger	4 + 5 + 7 + 23
Disgust	9 + 15 + 17

Table 2.2: Example of coding for the basic emotions

2.5 Physiological Emotion Recognition

In this chapter, we will introduce some of the methods and concepts within physiological emotion recognition and the physiological modalities we will use in this thesis.

2.5.1 Physiological Conditions

Physiological conditions are the condition or state of the body or bodily functions. Physiological signals can be used for affect recognition by detecting biological patterns that are reflective of emotional expressions. These signals are collected through sensors affixed to the subject's body. Many physiological signals could be considered regarding affect detection, and some will be investigated in this thesis.

2.5.2 Motivation

Emotions are reflected in our words, voice, body language, facial expressions, acoustic characteristics, and physiological signals. Even in a negative emotional state, a person may be able to force a smile. For that reason, facial expressions can be an unreliable source of emotion in some cases. While other factors can be faked, it is tough to control the physiological conditions of our bodies. Hence, focusing on emotion recognition using physiological signals controlled by the nervous system seems promising.

Common physiological signals such as electroencephalography (EEG), heart rate variance (HRV), electrodermal activity (EDA), respiration (RSP), and skin temperature (SKT) can be used for recognizing emotions. Nowadays, IoT technology makes physiological data more available along with activity data. In order to track their health, such as heart rate, blood pressure, the number of calories burned, and evaluate their movements, people are interested in buying items that are generally connected to their phones. Moreover, the spike in IoT motivates the use of physiological signals, considering such data is easily accessed.

2.5.3 Preprocessing

Preprocessing is an essential step of physiological emotion recognition (PER). In the task of physiological emotion recognition, it is necessary to eliminate the noise effects at an early stage of emotion recognition by preprocessing, due to the complex and subjective nature of raw physiological signals and the sensitivity to noises from crosstalk, measuring instruments, electromagnetic interferences, and movement artifacts (Shu et al. 2018). The most common preprocessing technique is filtering. Filtering is the act of cutting out noise from specific frequencies in the physiological signal. In chapter 3.3.1, state of the art within preprocessing of physiological signals will be presented.

2.5.4 Modalities For Physiological Emotion Recognition

This subsection describes a few common modalities used for physiological emotion recognition, along with common features extracted from those modalities.

Blood Volume Pulse

Blood volume pulse (BVP) is a measurement of heart rate based on the volume of blood that passes through the tissues in a localized area with each heartbeat. Commonly BVP is measured where a pulse can be easily accessed, such as on a finger or the wrist. The BVP sensor transmits infrared light through the tissue, and the absorption of light is measured by the blood flowing through the vessels. Every time the heart beats, the sensor detects a peak in this absorption, which is shown as the systolic point in Figure 2.5. The interval between the diastolic peaks defines the heart rate, shown as IBI (interbeat interval). The amplitude of the signal, i.e., the difference between the diastolic peak and the systolic peak, defines the subject's vasoconstriction. In other words, the diameter of the blood vessels.

The autonomic nervous system (ANS) regulates bodily functions such as heart rate, body temperature, sweating, blood pressure, digestion, and functions without any conscious voluntary control (McCorry 2007). Following, the sympathetic nervous system (SNS), a substructure of the

ANS, is responsible for the body's involuntary reaction to dangerous or stressful events. When such events occur, the SNS makes sure a flood of hormones is released to raise the alertness of the body, resulting in increased heart rate and extra blood to the muscles. Conclusively, the autonomic nervous system is responsible for dilating or contracting the blood vessel's diameter. Hence, changes in BVP amplitude reflect instantaneous sympathetic activation. For example, vasocontraction is usually decreased when a person relaxes, which is reflected by increased blood flow volume, consequently affecting the BVP amplitude. When anxious or fearful, the opposite is verified. For BVP to be helpful in emotion recognition, we need to extract some features to use in a model. Commonly used features to extract from BVP are mean and Std. Deviation.

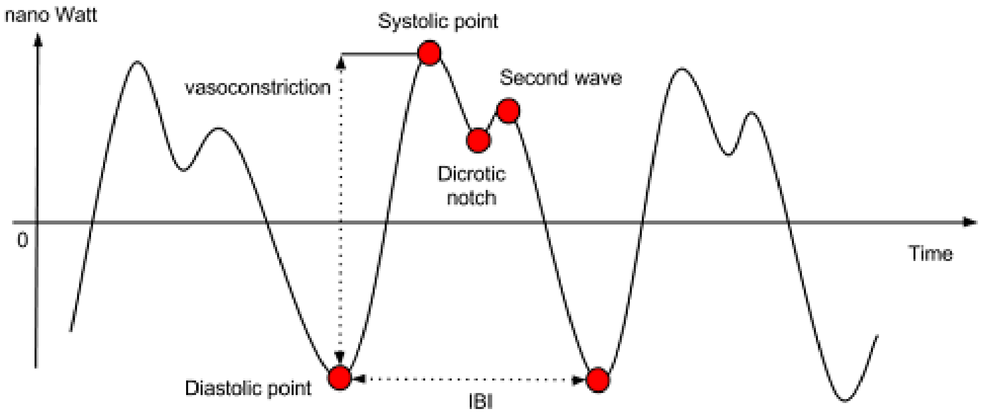


Figure 2.5: Blood volume pulse (© *E4 data - BVP expected signal* n.d.)

Electroencephalogram

Electroencephalography (EEG) is a technology used to measure brain activity. Brain behavior is a sophisticated concept that changes from one person to another and from one emotion to another. Moreover, extracting features from an arbitrarily chosen portion of the EEG signal has found itself important for emotion detection. For instance, identifying epochs where the excitation is at a maximum during the emotion. In multimodal emotion recognition, EEG has shown great success. However, due to the demanding task of collecting EEG signals, EEG as an emotion recognition approach is not a sustainable method. In a recent study by Y. Tan et al. (2021), differential entropy (DA) and power spectral density (PSD) were used as features in a classification model. DA is a measurement of information and represents the amount of information present in the data, while PSD describes the measurement of the signal's power.

Electrodermal Activity

Electrodermal activity (EDA, sometimes known as galvanic skin response) refers to the variation of the skin's electrical conductance in response to sweat secretion. The collection of EDA data is done by applying a low, undetectable, and constant voltage to the skin to measure how the skin conductance varies.

There are two main components to the overall complex, called EDA. The first component is the general tonic-level EDA, which relates to the signal's slower developing components and characteristics. This slower developing component is often referred to as Skin Conductance Level (SCL) and is thought to reflect the general changes in autonomic arousal. On the other hand, Skin Conductance Response (SCR) comes from the phasic component of the EDA signal. It reflects the faster changing characteristics resulting from sympathetic neuronal activity. (Braithwaite et al. 2013)

In a recent study by Gupta et al. (2022), EDA was used as a modality in an emotion recognition model. Some of the features included were the number of peaks per second, Mean, and Std. Deviation.

2.6 Multimodal Emotion Recognition

Multimodality refers to creating meaning using multiple sources to represent the information. Followingly, multimodal emotion recognition is that of trying to recognize an individual's emotional state using multiple features/representations. This chapter will briefly introduce some of the methods and concepts within multimodal emotion recognition and common features, before a further walkthrough of the field is displayed in Chapter 3.

2.6.1 Motivation

At present, the research on emotion recognition is mainly concentrated on unimodal emotion recognition such as text, speech, and facial expressions. Although unimodal emotion recognition has made many breakthrough achievements, they have also exposed some problems over time. For example, it cannot fully describe a particular emotion of the user at the moment, and using multiple modal features to describe a particular emotion together will be more comprehensive and detailed (W. Wei et al. 2019) (Zhang et al. 2020).

Many factors render multimodal affect recognition approaches appealing. Firstly, humans appear in a multimodal context when performing emotion recognition in real life. The voice, body, and face are all perceived as a whole by humans. When trying to learn a computer to reproduce elements of human emotional intelligence, it seems fitting to learn them to utilize the same approach. Secondly, the combination of multiple-affective signals provides a more rich data collection. Combining more than one modality to infer emotion will beneficially complement each other and help alleviate the effect of uncertainty in the raw signals. Lastly, with a more rich data collection, one may experience greater flexibility to classify emotions even when one or more source signals are lacking. In other words, when a particular modality contains less emotional information, the rest of the modality information can provide a supplement for the emotion classification task (Mou, Gunes and Patras 2019) (Zhao et al. 2019).

2.6.2 Multimodal Fusion Techniques

In approaches of multimodal emotion recognition, information extracted from each modality must be reconciled to obtain a single-effect classification result. This is known as multimodal fusion. The literature emphasizes two types of fusion techniques regarding multimodal fusion; fusion-level fusion and decision-level fusion. In the following subsection, we will present the general principles of the two approaches to multimodal fusion and describe their key ideas.

Feature-Level Fusion

Feature-level fusion (early fusion) is a fusion technique that concatenates the features from different modalities to obtain a joint representation before running it through a model. The goal of feature-level fusion is to find the best possible way to concatenate features that can increase emotion recognition performance. Fusion at the feature level using a simple concatenation of the modalities has been successfully used in several applications, with the main advantage being that correlation between modalities is easier utilized. However, because features obtained from different modalities can have different formats, synchronization of the features can be difficult and computationally expensive. Hence, the advantages of combining modalities at the feature level may be limited in some cases (Xie, Sidulova and C. H. Park 2021) (Wu, Lin and W.-L. Wei 2014). Figure 2.6 shows two inputs being concatenated in Feature-Level Fusion, before being handled by a single classifier.

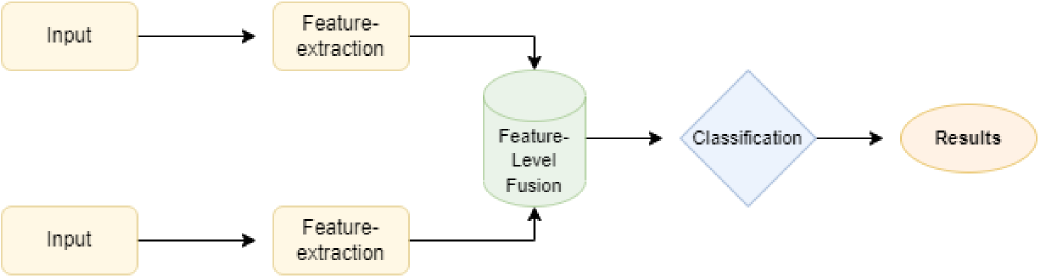


Figure 2.6: Feature-Level Fusion

Decision-Level Fusion

Decision-level fusion (late fusion) is a fusion technique that employs and trains separate classifiers for each modality to combine the outputs from each classifier thereafter. The goal of decision-level fusion is to obtain a final prediction based on two unimodal classifiers, with the main advantage being that decisions have the same format and, hence, can be more easily fused. Following this, the synchronization issues met at early fusion are avoided. Furthermore, using decision-level fusion allows for the application of optimal classifiers suited for each modality, thus providing more flexibility in the classification-step (C. Tan et al. 2020). Figure 2.7 shows two unimodal classifiers utilizing decision level fusion.

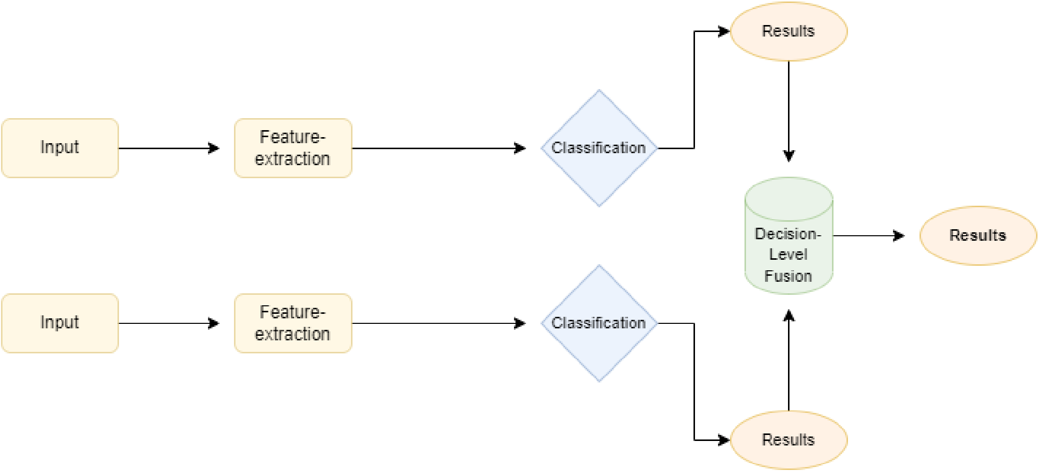


Figure 2.7: Decision-Level Fusion

2.6.3 Mutlimodal Combinations

At present, several combinations have been researched in the field of multimodal emotion recognition. Firstly, a common combination is combining facial expressions with audio signals, with the accessible data collection through video being a main factor for its popularity. Secondly, the combination of facial expressions and textual features has been explored in the field. A motivation for using this combination is increased textual usage, such as chatting. Lastly, we have that of combining facial expressions with physiological signals. Both EEG and EDA have been combined with facial expressions to achieve good results. In chapter 3, state of the art within these combinations will be discussed.

2.7 Evaluation Metrics

This section presents metrics that are often used to evaluate the quality of classification models. These measures use the values of false positives, false negatives, true positives, true negatives, and true positives.

False positives (fp) denote the number of incorrectly classified positive instances, while false negatives (fn) denote incorrectly classified negative instances. True positives (tp) denote the number of correctly classified positive instances, while true negatives (tn) denote the number of correctly classified negative instances.

2.7.1 Accuracy

Accuracy is the proportion of the total number of predictions that were correct. Intuitively, accuracy measures the ability of the classifier to classify correctly across all classes and is useful when all classes are of equal importance. The formula for accuracy is given as:

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{2.1}$$

2.7.2 Precision

Precision is the fraction of positive samples that has been classified correctly among all samples predicted as positive. Intuitively, precision measures the ability the classifier has to correctly label samples. The formula for precision is given as:

$$Precision = \frac{tp}{tp + fp} \tag{2.2}$$

2.7.3 Recall

Recall is the fraction of positive samples that has been correctly classified among all relevant samples. Intuitively, recall measures the ability the classifier has to find all the relevant samples. The formula for recall is given as:

$$Recall = \frac{tp}{tp + fn} \tag{2.3}$$

2.7.4 F1-score

The F1-score combines the values of precision and recall , to further take a harmonic mean between the two. The formula for F1-score is given as:

The F1-score combines the precision and recall of a classifier into a single metric by taking their harmonic mean

$$F_1 score = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{2.4}$$

2.7.5 Matthews Correlation Coefficient

Matthews Correlation Coefficient (MCC) (Reinke et al. 2021) is a statistical rate in the range -1 to 1, which produces a high score only if the prediction obtained good results in all of the four confusion matrix categories (tp,fn,tn,fp), proportionally both to the size of positive elements and the size of negative elements in the dataset (Chicco and Jurman (2020)). The performance metric was originally made for binary classification, but the metric has been extended to the multi-class case with some changes to the equation. The equation in the binary classification case is illustrated in 2.5.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \tag{2.5}$$

In the multi-class case, the MCC can be defined in terms of a confusion matrix C for K classes.

$$MCC = \frac{c \times s - \sum_k^K p_k \times t_k}{\sqrt{(s^2 - \sum_k^K p_k^2)(s^2 - \sum_k^K t_k^2)}} \tag{2.6}$$

To simplify the equation, the intermediate variables are described as:

- $c = \sum_k^K C_{kk}$ the total number of elements correctly predicted
- $s = \sum_i^K \sum_j^K C_{ij}$ the total number of elements
- $p_k = \sum_i^K C_{ki}$ the number of times that class k was predicted

- $t_k = \sum_i^K C_{ik}$ the number of times that class k truly occurred

The equation for the multi-class case is created by Chicco and Jurman (2020).

2.8 Performance Metrics

Performance metrics (also called error measures) are types of metrics that measure the error of a forecasting model. In machine learning, performance metrics are used to compare the predictions of a model with the actual data from a test data set (Botchkarev 2019).

2.8.1 Mean Absolute Error

Mean Absolute Error (MAE) is a model evaluation metric used with regression models. With respect to a test set, the mean absolute error of a model is the mean of the absolute values of the individual prediction errors over all instances in the test set. The difference between the true value and the predicted value is the prediction error for a given instance (Sammut and Webb 2010). The formula for MAE is given as:

$$MAE = \frac{\sum_{i=1}^{n} abs\,(y_i - \lambda(x_i))}{n} \tag{2.7}$$

y_i is the target value for an instance x_i, $\lambda(x_i)$ is the predicted target value for the instance x_i , and n is the number of test instances.

2.8.2 Root Mean Squared Error

Root Mean Squared Error (RMSE) is the square root of the average of squared prediction errors. The effect of each prediciton error on RMSE is proportional to the size of the squared error. This means that larger errors have a disproportionately large effect on RMSE, which makes RMSE sensitive to outliers (Hyndman and Koehler 2006). The formula for RMSE is given as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\lambda(x_i) - y_i)^2}{n}} \tag{2.8}$$

y_i is the target value for an instance x_i, $\lambda(x_i)$ is the predicted target value for the instance x_i , and n is the number of test instances.

2.9 Tools

This section provides a description of the tools and libraries that were used in this thesis.

2.9.1 Pytorch

Pytorch is an open-source machine learning framework based on the Torch library used for applications such as computer vision and natural language processing (Paszke et al. 2019).

2.9.2 Scikit-learn

Scikit-learn is an open-source machine learning library used for efficient precictive data analysis. Scikit-learn is a Python module and is built on top of SciPy. Pedregosa et al. (2011).

2.9.3 Google Forms

Google forms is a tool that lets you easlily create forms for information gathering. Google Forms is developed by Google and supports the inclusion of video in the form, which is the main reason why it was chosen for this thesis. A clear display of statistical features from the answears is also provided by Google.

2.9.4 DeepFace

Deepface is a lightweight face recognition librabry for Python, and has become one of the most popular ones. It includes state of the art models within the field of FER, and handles all procedures for FER in the background. Additionally, the library is open-source, with great documentation, making it easy to access and utilize. (Serengil and Ozpinar 2020)

Chapter 3

Related Work

This chapter first presents a review of the existing research in the field of multimodal emotion recognition, as well as the challenges faced in the field. Then, due to the fact that multimodal emotion recognition is a combination of two unimodal approaches, we will first present the state-of-the-art approaches within FER and PER. Next, the chapter presents the state-of-the-art approaches within the field of multimodal emotion recognition, with the main focus being the use of FER and PER.

3.1 Studies On Mutlimodal Emotion Recognition

Understanding emotions is an essential part of being human. Along with the growth in "internet-of-things" and wearable technology, perhaps this is why we have witnessed remarkable growth in affective science. While the amount of research increases, the field still faces several challenges, both in the actual task of detecting (labeling and eliciting) emotion and the research field in general. Al Osman and Falk (2017) have summarized the following challenges for the task of multimodal emotion recognition.

Firstly, one of the challenges in developing multimodal affect-recognition methods is the need to collect multisensory data from a large number of subjects. Also, it is difficult to compare the obtained results with other studies, given that the experimental setup varies. We divide the databases into three types: posed, induced, and natural-emotional databases. For the posed databases, the subjects are asked to act out a specific emotion while the results are captured. For the induced databases, the subjects are exposed to stimuli (e.g., watching a video) in a controlled setting. For the natural databases, the subjects are exposed to real-life stimuli such as interaction with humans or machines. Regarding the different types of emotion elicitation approaches, comparing results is a challenge. Similarly, it is well established that context affects how humans express emotions (Hess, Banse and Kappas 1995) (Izard 1994). Therefore, the problem of gathering a "ground-truth" arises, as conflicting cases between single modalities may appear. For instance, a user may consciously or unconsciously conceal his/her real emotions through external channels of expression but still reveal them through internal channels of expression, due to the context the person appears in.

Secondly, multimodal affect-recognition methods necessitate the fusion of the modal features extracted from the raw signals, and there is still uncertainty about which fusion technique is the top performer (Lingenfelser, Wagner and Andre 2011). However, decision-level fusion is seemingly the most popular choice. Lastly, it is still unclear what type and number of modalities are needed to achieve the highest level of accuracy. These topics will be further discussed throughout the thesis.

As the amount of work on multimodal emotion recognition increases, some studies aim to create overviews of existing work within the field of study. One of the latest overviews created is the publication of Abdullah et al. (2021), serving as a review of the recent advancements in emotion research using multimodal signals, with feature extraction and classification methodologies using deep learning. In addition, the publication of Seng and Ang (2019) contributes to the research of using multiple modalities in emotion recognition, looking at challenges regarding fusion and classification techniques. They also provide further insight to that of a unimodal vs. a multimodal approach.

There are also several challenges concerning the actual research field. The lack of a benchmark dataset makes it difficult to compare studies and methods, along with the problem of comparing unimodal and multimodal approaches. Admittedly, one may compare unimodality with multimodality within the same dataset. However, comparing accuracy between different datasets is a challenging task since there is no common understanding of the task or terms within the field. Although there is a common and overall goal of detecting emotion, there are variations in how this goal is approached and the subtasks studied. For instance, whether to classify emotions in continuous dimensional or with the use of discrete categories. Additionally, the number of emotional categories to classify and the size of the datasets widely varies, making the comparison hard. The following sections will introduce the state-of-the-art within FER and PER.

3.2 Methods In Facial Emotion Recognition

This section describes methods within preprocessing, feature extraction and classification methods in facial emotion recognition.

3.2.1 Preprocessing Using Facial Emotion Recognition

The primary purpose of facial image preprocessing is to improve the quality of images and enhance the images' features for further preprocessing. Regarding face detection, several algorithms have been used with good performance. In the paper Yang et al. (2018), Haar Cascades method is used to detect whether the image contains a face or not, and Kalsum et al. (2018) used the Viola-Jones algorithm to detect the facial part of the image. Following this, Rahmad et al. (2020) performed a comparison between the combination of Haar Cascade and Viola-Jones algorithm (V-J) and another detection algorithm, the Histogram of Oriented Gradients (HOG). The paper showed that HOG is slightly more accurate than V-J for face detection, mainly regarding images including multiple faces. Figure 3.1 shows face detection of an image from the Toadstool dataset using Haar Cascades.

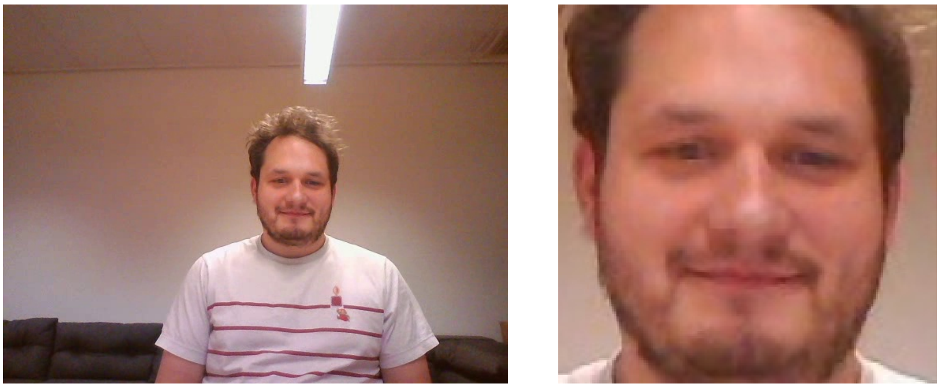


Figure 3.1: Face detection using Haar-Cascades

Khaireddin and Chen 2021 used cropping and data augmentation to achieve state-of-the-art results on the FER13 dataset. In their approach, data augmentation was provided to account for variability in the facial expressions in the dataset. The augmentation included rescaling, shifting, rotating, and cropping, with each of the techniques, applied randomly with a probability of 50 percent. Further, Raghavan and Ahmadi (2021) expressed the challenge facial recognition faces due to illumination conditions. Stating that if the lighting conditions present in the gallery image are different from the probe image, then the face recognition process may ultimately fail. Finally, an experiment on the extended Yale B database (Georghiades, Belhumeur and Kriegman 2001) showed an improvement in face recognition by enhancing the intensity in the regions being inadequately illuminated and decreasing the intensity in the densely illuminated regions while retaining the intensity in the fairly illuminated portions.

3.2.2 Feature Extraction Using Facial Emotion Recognition

After the image has been preprocessed, features that will provide information and reflect its class are extracted. The features must also be represented suitably before being fed into a machine learning classifier. In section 2.4.3 we described the different types of features, with CNNs being one of the most popular approaches. However, comparing the feature extraction performance between different CNNs is a difficult task, considering the feature extraction is done automatically as a result of the CNN layers.

Regarding classical approaches, in other words approaches not included in the field of deep learning, support vector machines are the most applied classical approach to classifying emotions from facial expressions. With respect to the state-of-the-art review performed by Canal et al. (2022), features such as Local binary pattern, histogram of oriented gradients and Scale-invariant feature transform has proved to perform well.

3.2.3 Classification In Facial Emotion Recognition

Supervised machine learning classifiers have been the most frequently used approach for the task of multimodal emotion recognition. As mentioned in Chapter 2, there has been a recent growth in the use of deep learning methods for machine learning tasks. This is also the case for facial emotion recognition. Convolution neural networks (CNN), as presented in chapter 2, are one of the most common methods used in facial emotion recognition and one of the best-performing methods.

As mentioned in the subsection 3.2.1, Khaireddin and Chen 2021 achieved state-of-the-art accuracy on the FER2013 dataset when adopting a VGGNet architecture (Simonyan and Zisserman 2014). Their variant of VGGNet consists of four convolutional stages and three fully connected layers. Each of the convolutional stages contains two convolutional blocks and a max-pooling layer. The convolution block consists of a convolutional layer, a ReLU activation, and a batch normalization layer. Batch normalization is used to speed up the learning process, reduce the internal covariance shift, and prevent gradient vanishing or explosion. A ReLU activation follows the first two fully connected layers. The third fully connected layer is for classification. They achieved an accuracy of 73.28 percent, classifying seven emotions on the FER dataset.

Khattak et al. (2022), tried to address the problem of poor layer selection in CNNs, resulting in performance degradation. They proposed a CNN, which is designed for image classification purposes. Furthermore, different convolutional layers and a varied set of parameters are used for efficient classification. Their result shows that their model outperformed the state-of-the-art CNNs on the CK+ dataset when recognizing seven emotions. They achieved an accuracy of 95.65 percent. However, only a subset of the dataset images was used, which may have increased the accuracy.

3.3 Methods In Physiological Emotion Recognition

This section describes methods within preprocessing, feature extraction and classification methods in physiological emotion recognition.

3.3.1 Preprocessing Using Physiological Signals

In the task of physiological emotion recognition, it is necessary to eliminate the noise effects at an early stage of emotion recognition by preprocessing, due to the complex and subjective nature of raw physiological signals. Selvaraj et al. (2013) used electrocardiogram (ECG) as a modality in a physiological model of emotion. In their approach, a Butterworth low pass filter with a cut-off frequency of 40 Hz was applied to increase the signal quality. An analysis of the effect of such preprocessing was not investigated in the paper. However, ECG signals often contain noise, and a filter is commonly regarded as an acknowledged method.

EEG and BVP signals are typically contaminated by physiological artifacts caused by electrode movement, eye movement or muscle activities, and heartbeat. Nakisa et al. (2018) used a sixth-order (band-pass) Butterworth filter in order to remove artifacts while keeping EEG signals within desired frequency bands. The Butterworth filter will obtain signals in the range 4-64Hz, excluding noise such as non-physiological artifacts that may appear above the 50Hz range. Subsequently, a 3Hz low pass-Butterworth filter was applied to remove noise from the BVP signal. An analysis of the use of different preprocessing was not conducted in the study.

Svoren (2020) transformed an original BVP signal into a sample set of associated amplitudes that could be used for training our CNN models. As the range of values varied widely from person to person, the range of the signal was normalized into a range between -1 and 1. Further, he used the find peaks method from the ScyPy library (Virtanen et al. 2020) to locate the peaks.

3.3.2 Feature Extraction Using Physiological Signals

As in facial emotion recognition, selecting optimal features is a crucial part of the learning process when using physiological signals to achieve a fair and extensive analysis of the problem. Popular features for various physiological signals include frequencies, amplitudes, maxima, and minima (Egger, Ley and Hanke 2019).

Gupta et al. (2022), used all thirty-eight physiological signals present in the K-emotion dataset. For feature extraction, they used an open-source library called PyTeap (M. and Villaro-Dixon 2017), which is a python implementation of "Toolbox for emotion analysis using physiological signals", and extracted all possible features of the signals present. For example, from BVP, features such as heart rate variability, interbeat interval, and mean were extracted. Further, EDA features such as “number of peaks per second”, mean, and the average amplitude of peaks were extracted. Finally, each feature was further divided into smaller windows to adequately capture signal information.

Feature selection is the problem of choosing the most valuable features that best represent the underlying problem. Shukla et al. (2019) conducted a feature selection analysis of EDA features using the AMIGOS dataset (Miranda-Correa et al. 2017). A systematic comparison of 621 features was performed using three feature selection methods (Joint Mutual Information, Conditional Mutual Information Maximization, Double Input Symmetrical Relevance). Their research showed that a high number of features, approximately 95, are required to obtain optimal accuracy. Conclusively, regarding the significance of specific features, the study showed that statistical features related to MFCC (Mel-frequency cepstrum) outperformed all other features.

3.3.3 Classification Using Physiological Signals

In the field of PER, many different classification methods have been suggested by previous studies. As mentioned in Chapter 2, there has been a recent growth in the use of deep learning methods for machine learning tasks. This is also the case for the specific task of PER. In the work of Salari, Ansarian and Atrianfar (2018), EEG, BVP and EDA were used to detect emotions. Several physiological statistical features were extracted and sent to a CNN, where it achieved an accuracy of 85.83% and 75.42% for valence and arousal, respectively. Surpassing those achieved in other papers using common traditional classifiers like SVM, on the DEAP dataset (Koelstra et al. 2011).

A comparative study performed by Gupta et al. (2022) tested and compared the performance of four traditional machine learning models, namely Gaussian Naive Bayes, KNN, DT, and SVM, along with one deep learning model. The classifiers were tested on five different annotation methods on the K-Emocon dataset (C. Y. Park et al. 2020). From the study, it was observed that the deep learning model achieved the best accuracy in all cases except 1, where an SVM beat it by 0.08%. The best accuracy achieved for valence and arousal were 91.12% and 62.19%, respectively. The SVM was the second-best performer in the other four cases.

3.4 State-Of-The-Art In Multimodal Emotion Recognition Using Facial Expressions and Physiological Signals

This section describes state-of-the-art within multimodal emotion recognition using facial emotion recognition and physiological signals. First, the issues concerning datasets will be presented, along with the most commonly used datasets for the task. Then different modality combinations for MER will be presented.

3.4.1 Datasets

The lack of a benchmark dataset for multimodal emotion recognition is an issue, as it becomes difficult to compare models and results based on different data and annotations. Furthermore, considering the challenge mentioned earlier of picking modalities, the datasets created for multimodal emotion recognition have different characteristics based on the task, particularly regarding modalities and the emotions that are possible to predict, which can be both discrete and categorical. In addition, creating datasets for multimodal emotion recognition is time-consuming, as it requires more work compared to the collection of a single modality. Further, several datasets have not been made publicly available. This may be due to privacy issues or simply people not wanting to give datasets out for free. Subsequently, the growth in innovation within multimodal emotion recognition is faced with a problem. However, despite challenges, there has been an increase in datasets created. Table 3.6, displayed at the end of this chapter, provides an overview of the most common datasets used in multimodal emotion recognition.

3.4.2 Modality Combination With Facial Expressions

In this subsection, we will include, to our best knowledge, the best performing modality combinations of facial expressions with another modality.

Facial Expressions And Text

The emotions that prompt individuals to create text with certain words at particular times are what text-based emotion recognition is concerned about. To a certain degree, humans have the ability to understand emotions from text, leading to motivation for computers to do the same. However, without contextual information, the inclusion of sarcasm, and the relationship between the reader and the author, textual interpretation is a complex task for humans and computers.

In a paper from 2020, Lee, Kim and Cheong (2020) investigated the possibility of combining facial emotion recognition with text. The research goal was to classify characters' facial images in a Korean TV series into seven emotions: Angry, Disgust, Fear, Happy, Neutral, Sad, and Surprise. A multimodal deep learning model was implemented using a facial image and text describing the situation as input values. The experiment showed an increase in F1-score when using text descriptions of the characters and facial expressions, compared to a unimodal approach of facial expressions. A summary of the paper is presented in table 3.1.

Facial Expressions And Speech

When people engage in spontaneous conversation exchanges, their speech may reveal their emotional state and personality traits, in addition to the meaning of the words and their conveyance. Schuller and Batliner (2013) has provided an overview of computational paralinguistics, with paralinguistics being the characteristics of the voice that is being used to transmit emotions, addressing the primary techniques for recognition of emotion in human speech.

Luna-Jiménez et al. (2022) investigated the possibility of combining facial expressions with speech. An LSTM was used for FER, using FACS features, achieving an accuracy of 62.13%. Further, a XLSR-Wav2Vec2.0 (Conneau et al. 2020) was used for speech emotion recognition (SER), achieving an accuracy of 81.82%. However, by combining these two modalities with a decision level fusion strategy, they achieve 86.70% accuracy on the RAVDESS dataset, when classifying eight emotions. Results demonstrated that these modalities carried relevant information to detect users' emotional state and their combination allowed them to improve both their final system performance. A summary of the paper is given in table 3.2.

Modalities	Dataset	Feature-Extraction	Classification	Fusion	Classes	Performance
Facial expressions, text	Built own dataset	Not reported	CNN	Decision-Level	Basic Emotions	Increased F1-score for 5 of 7 emotions

Table 3.1: Overview of Lee et al (2020)

Modalities	Dataset	Feature-Extraction	Classification	Fusion	Classes	Accuracy
Facial expressions, speech	RAVDESS	FER: FACS	FER: CNN SER: xlsr-Wav2Vec2.0	Decision-Level	Eight emotions	FER: 62.13% SER: 81.82% MER: 86.70%

Table 3.2: Overview of Luna-Jiménez et al (2021)

3.4.3 Modality Combinations With Physiological Signals

In this section, we will include the best performing modality combinations of physiological conditions with another modality.

Facial Expressions And EEG

Y. Tan et al. (2021) published a paper where a multimodal emotion recognition method was proposed to establish an HRI (human-robot interaction) system with a low sense of disharmony. They performed a multimodal experiment using facial expressions and EEG. The EEG data were collected in a lab environment using an electrode cap, and the facial expressions were collected using a camera, with video being the elicitation tool. The data would further be self-labeled by the subjects and sent through separate classifiers for facial and EEG, trained on FER13 and Seed-IV, respectively.

The facial expressions and EEG results were then combined using the Monte Carlo method for the multimodal experiment. The model classified four emotions with an accuracy of 83.33%, being an improvement compared to both unimodal approaches. A summary of the paper is presented in table 3.3.

Facial Expressions with EDA, Heart Rate and Respiration

Zhong et al. (2017) combined facial expressions with several physiological signals in an attempt to increase the recognition rate of emotion compared to that of a unimodal approach solely using facial expressions. In the paper, they used facial expressions, EDA, heart rate, and respiration data from the multimodal MAHNOB-HCI database (Lichtenauer and Soleymani 2011). For facial expressions, AFFDEX (Bishay et al. 2022), a facial expression analysis toolkit, was used to extract facial features. From the physiological signals, a total of 130 features were extracted. These include time, minima, maxima, frequency, statistical and spectral features, which were further fused with the facial expressions using late fusion. A simple concatenation between the vectors was applied regarding the late fusion technique. The results showed an increase in recognition of both valence and arousal compared to a unimodal approach using facial expressions. A summary of the paper is given in table 3.4.

Combining Physiological Signals

Gupta et al. (2022) researched the possibility of combining physiological signals to recognize emotions. Their research evaluates physiological signals for emotion classification using K-Emocon, which is one of the most recently published datasets in the field of physiological signals. Compared to previous published multimodal datasets using specially selected pictures and videos, K-Emocon proposed a more reliable method of eliciting emotion. Moreover, the participant's emotion is recorded during a debate about Yemeni refugees in a social setting. This paper used all the physiological signals present in the K-Emocon dataset. Regarding the model's performance (Neural network), it achieved the best accuracy for valence at 91.12%, and the best accuracy for arousal was 62.19%. Summary numbers about the paper are found in table 3.5, and more information about K-Emocon is found in table 3.6.

Modalities	Dataset	Feature-Extraction	Classification	Fusion	Classes	Accuracy
Facial Expression, EEG	FER13 and Seed-IV	FER: CNN EEG: DE features	FER: CNN EEG: SVM	Decision-Level	Neutral, Sad Fear, Happy	FER: 69.48% EEG: 79.25% MER: 83.33%

Table 3.3: Overview of Ying Tan et al (2021)

Modalities	Dataset	Feature-Extraction	Classification	Fusion	Classes	Accuracy
Facial Expression, EDA, HR and Respiration	MAHNOB-HCI	FER: TIPF PER: TIPF	FER: SVM PER: SVM	Decision-Level	Valance, Arousal	FER: Valence: 67.3% Arousal: 69.0% MER: Valence: 69.0% Arousal: 71.9%

Table 3.4: Overview of Zhong et al (2017)

Modalities	Dataset	Feature-Extraction	Classification	Fusion	Classes	Accuracy
Physiological Signals	K-Emoticon	PyTeap Library	Neural Network	Decision-Level	Valence, Arousal	Valence: 91.12% Arousal: 62.19%

Table 3.5: Overview of Priyansh et al (2021)

Name	Type	Elicitation	Labelling	Classes	Source
FER13	Unimodal, Facial Expression	Posed and unposed images from Google	Self-assessed, raters	Basic Emotions	*FER13* (2013)
DEAP	Multimodal, Facial Expressions, EEG, photoplethysmography, Skin Temperature	Music Video	Self-assessed	Valcene, Arousal	Koelstra et al. 2011
IEMOCAP	Multimodal, Facial Expressions, Speech	Posed	Self-assessed	Basic Emotions	Busso et al. 2008
CK+	Unimodal, Facial Expressions	Posed	FACS	Basic Emotions	Lucey et al. (2010)
RAVDESS	Multimodal, Facial Expressions, Speech	Posed	Raters	Happy, Sad, Angry, Fearful	Livingstone and Russo 2018
K-emoCon	Multimodal, Facial Expressions, Physiological Signals	Debates on social injustice	Self, Partner and external	Dimensional and categorical	C. Y. Park et al. 2020
MELD	Multimodal Facial Expressions, speech, text	Dataset extracted from dialogues in the Tv-show Friends	Raters	Basic Emotions	Poria et al. 2018
MAHNOB-HCI	Multimodal Facial Expressions, EDA, HR Respiration	Video clips	Self-assessed	Valence, Arousal	Lichtenauer and Soleymani 2011

Table 3.6: Overview of the most used datasets in MER

Chapter 4

Data

Machine learning methods learn from data. While the representation and amount of data needed depend on the complexity of the problem to be solved, it is beneficial that the data is somewhat representative of the related real-world problem. Creating labeled datasets can be a tedious and demanding task. Therefore, the datasets used in this thesis will be already existing datasets. However, we will provide labeling of one dataset with the use of machine learning models. The data from the datasets are used to train and evaluate the classification models for emotion detection, described in chapter 7. Section 4.1 presents the datasets used and their distribution of labeled instances, as well as how the data was collected and labeled. Then, section 4.2 will discuss why these particular dataset were used, as well as why other datasets were discarded.

4.1 Datasets

Several datasets were used to investigate emotion recognition for increased insight and to allow comparisons of the findings. However, the datasets are different in terms of annotations, elicitation, labeling, and the emotions that are being classified. This will be described in the following subsections.

4.1.1 Toadstool (2020)

The most significant dataset for this thesis was provided by Svoren (2020) and has been made publicly available by Simula. Toadstool (Svoren

(2020)) includes both physiological signals and a video of 10 participants. In the process of collecting the data, the participants were placed in a quiet room, only assisted by Svoren, and were told to play a video game for thirty-five minutes. To capture physiological signals from the participants, sensors of E4 Empatica Wristbands were used. This device uses four different sensors to collect various information from the user:

1. A photoplethysmography sensor, measuring blood volume in an area tissue. Further used to calculate BVP.
2. An EDA sensor, measuring the electrical conductivity of the skin.
3. A 3-axis accelerometer, measuring movement and activity.
4. An optical thermometer that measures temperatures.

The video captured the participant's face, posture, and movement as they played, using a python script utilizing a webcam placed on the screen. The video was captured at 30 frames per second. Further, the BVP signal has a sampling rate of 64Hz, the EDA signals has a sampling rate of 4hz, the Accelerometer signal has a sampling rate of 32Hz, and the HR signal has a sampling rate of 1Hz. Table 4.1 shows the data included in the Toadstool dataset.

Label	Video-samples (30 frames)	BVP Signals (64hz)	EDA Signals (4hz)	Accelerometer (32hz)	HR (1hz)
Total Measurements	63 000 frames	134 400 signals	8400 signals	67 200 signals	2100 signals

Table 4.1: Overview of data in Toadstool (2020)

4.1.2 CK+

One of the datasets used in this thesis to train a FER model was The Extended Cohn-Kanade (CK+) dataset, provided by Lucey et al. (2010), where access to the dataset had to be granted. The dataset contains 593 video sequences with facial behavior of 123 different subjects, ranging from 18 to 50 years of age, with various genders and background. For the 593 posed sequences, full FACS coding of peak frames is provided. The FACS coding was further used to validate the posed labels, where only 327 met the criteria for one of seven discrete emotions. Table 4.2 shows the distribution of the annotated sequences in CK+.

Emotion	Amount
Angry	45
Contempt	18
Disgust	59
Fear	25
Happy	69
Sadness	28
Surprise	83
Total	327

Table 4.2: Overview of the original dataset by Patric Lucey et al (2010)

To satisfy our needs, some modifications to the dataset were needed. The developer of the dataset did not include neutral sequences in the dataset. Analyzing the sequences in CK+, all sequences start with a neutral face and gradually evolve into a given emotion. Therefore, we manually provided 52 neutral sequences by extracting the neutral part of sequences from CK+. Additionally, we also had to modify CK+ to be able to use it in a 2D-CNN. In other words, modify the data to include single labeled frames instead of labeled sequences. Accordingly, with the sequences ranging from neutral to peak, the last two frames of each sequence were extracted. In addition, the emotion of "contempt" was removed due to its low amount samples. The distribution of the modified CK+ data with sequences is displayed in table 4.3 and table 4.4 shows the modified version of the CK+ dataset with single frames.

Emotion	Amount
Angry	45
Disgust	59
Fear	25
Happy	69
Sadness	28
Surprise	83
Neutral	52
Total	361

Table 4.3: Overview of the modified CK+ (2010) with sequences

Emotion	Amount
Angry	90
Disgust	118
Fear	50
Happy	138
Sadness	56
Surprise	166
Neutral	104
Total	722

Table 4.4: Overview of the modified CK+ (2010) with single frames

4.1.3 FER 13

Another dataset used for FER was the FER13 dataset, provided by *FER13* (2013). FER13 was used through the DeepFace library (Serengil and Ozpinar 2020), as some of the models included in the library are trained on FER13. The dataset consists of 28,000 labeled images in the training set and 3,500 labeled images in the test set. The dataset was created using images from Google search, searching for images of faces that match the basic emotions plus neutral. Along with synonyms for these emotions. These images were further processed with face detection to obtain images only consisting of a face. Lastly, human labelers were used to reject incorrectly labeled images, correct the face detection's cropping if necessary, and filter out some duplicate images. Images approved by the human labelers would then be resized to 48x48 pixels and converted to grayscale.

Each image in FER13 is labeled with one of the six basic emotions plus neutral. The distribution of labels is provided in table 4.5

Emotion	Training set	Test set
Angry	3995	958
Disgust	436	111
Fear	4097	1024
Happy	7215	1774
Sad	4830	1247
Surprise	3171	831
Neutral	4965	1233
Total	28,000	3,500

Table 4.5: Overview of the FER13 dataset

The FER13 dataset is relatively evenly distributed regarding the labels, except for the emotions of happy and disgust. The happy category is provided with almost double the amount of samples compared to the other categories. In contrast, the disgust category contains around a tenth of the average amount of samples for an emotion.

4.2 Discussion

With respect to the goal of the thesis, namely, investigating the effect of combining facial expressions with physiological signals, we needed a multimodal dataset that included the necessary data. From our literature study, we experienced that two datasets, DEAP (Koelstra et al. 2011) and MAHNOB-HCI (Lichtenauer and Soleymani 2011), included both facial expressions and physiological signals. The MAHNOB-HCI dataset already had some work done regarding the combination of facial expressions and physiological signals, as seen in section 3.4.3. Therefore, we wanted to pursue the DEAP dataset, as it was seemingly less explored in the field of MER, with our desired combination of data. However, as mentioned in 3.4.1, getting access to datasets can be difficult. This was also the case in our approach of obtaining the DEAP dataset (Koelstra et al. 2011), resulting in us not managing to get in contact with the providers of the dataset. Still, in the search for a multimodal dataset, we ended up using a dataset called Toadstool (Svoren (2020)).

As presented in section 4.1.1, Toadstool included both facial expressions and physiological signals. As well as including our desired data, Toadstool seemed appealing due to its natural collection of data. Where DEAP and MAHNOB-HCI was collected with each participant being suited with an EEG helmet, in a seemingly strict environment as displayed in figure 4.1, the data in Toadstool was collected in a more relaxed and natural environment, with an armband being the only instrument the participant had to worry about. This way of collecting data can stimulate more natural and realistic emotions. Additionally, in the context of utilizing emotion recognition as a service in practice, utilizing an armband is a more realistic scenario than using an EEG helmet. Figure 4.1 shows the environment of DEAP and MAHNOB-HCI, while figure 4.2 displays the Toadstool environment.

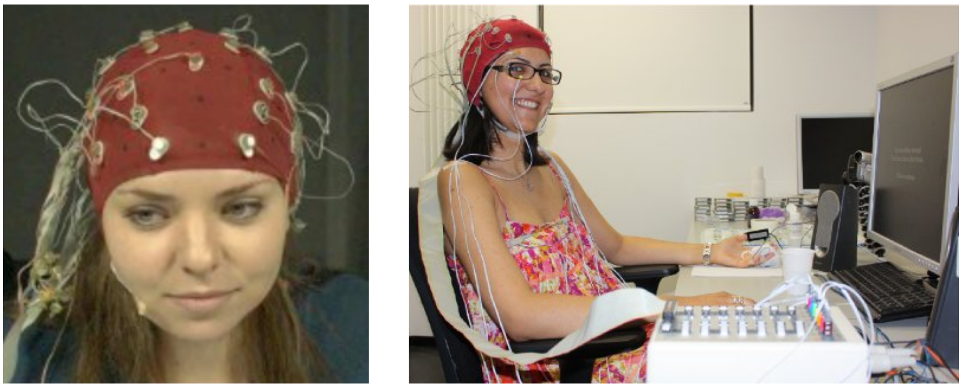


Figure 4.1: Environment of MAHNOB-HCI and DEAP

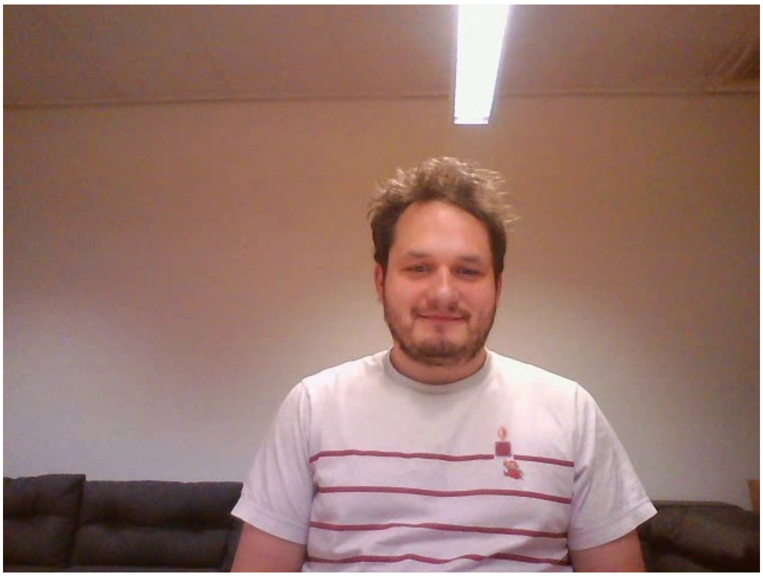


Figure 4.2: Environment of Toadstool

However, the Toadstool dataset did not contain any emotion labels. Therefore, an update of the Toadstool dataset, containing a mapping of the data and labeling, had to be done. The process of labeling Toadstool is described in detail in chapter 6.

Chapter 5

Investigating the correlation between Facial Expressions and Blood Volume Pulse

5.1 Work Of Svoren (2020)

Investigating the correlation between facial expressions and physiological signals can reveal a crucial connection for further work into multimodal emotion recognition using facial expressions and physiological signals. In a preliminary experiment, Svoren (2020) investigated this correlation where a deep convolutional neural network(CNN) was trained to predict BVP amplitude values using a combination of frames of facial expressions from a video and corresponding frames from a video game session. The experiment assumed that by creating an ML model that could predict BVP amplitudes with a low error, one would show that the model confirmed a correlation between the physiological state of the subject and the combination of the facial expression and video game frame.

5.1.1 Plan

The deep CNN, which was based on the TensorFlow implementation of ResNet50 (K. He et al. 2015), was trained on each of the 10 participants from Toadstool individually. In detail, the model's input was two video game frames(the first and last frame of one second) and the facial expression from the last frame of the second corresponding to the video game frames. Since the output of the model and the corresponding true value was a continuous BVP amplitude measurement between 0 and 1, there was little sense in evaluating performance based on overall accuracy as the model would most likely never predict the exact values. More suitable measurements, in this case, was mean average error(MAE) and root squared mean error(RMSE) because these would give a better indication of how far away the model predictions were. With a Zero Rule (ZeroR) classifier as the baseline, the results showed that the baseline experiment performed overall better than the CNN model, addressing the task's difficulty. The ZeroR classifier takes the most frequently occurring true label from the training dataset and uses that as output when predicting on the test dataset. The poor results were addressed by Svoren (2020) and recurrent neural networks (RNN) were suggested as a method to increase performance. The intuition behind the suggestion is that using methods such as RNNs would add temporal information to the prediction of BVP amplitudes. By looking at the evolution of a facial expression over a given time frame rather than the facial expression in a given instance, one could get a more complete picture of the state of a subject, which in turn could improve the prediction of a subject's physiological state in the form of BVP.

5.2 Investigating The Use Of Spatio-Temporal Neural Networks

This chapter aims to present the preliminary experiment built upon the work of Svoren (2020). The preliminary experiment investigates the possible effects of using spatio-temporal neural networks in predicting physiological signal values based on facial expressions. First, section 5.2.1 will present the experiment plan. Next, section 5.2.2 will present a description of the model architectures used. Lastly, in section 5.2.3, the performance of the model used will be presented and compared to a baseline experiment, as well as compared to the results from the experiment of Svoren (2020).

5.2.1 Experiment Plan

Our experimental plan will closely follow the plan of Svoren (2020), with some adjustments in context with the architecture of our models and goals. Similar to what Svoren (2020) did, and based on the nature of the BVP signal explained in 2.5.4, we transformed the original 64Hz BVP signal into a sample set of associated amplitudes. While the input of the CNN from Svoren (2020) was a combination of two video game frames and the corresponding frame of facial expression, our spatio-temporal models were inputted ten facial expression frames extracted from 4 seconds of video. Since the goal was to predict a single BVP amplitude, the mean BVP amplitude from each second was again averaged over the 4 seconds, meaning each datapoint consisted of 10 frames mapped to 1 mean BVP amplitude. The decision of using 4 second long sequences was a result of Ekman (2007); an emotion typically lasts between 0.5 and 4 seconds. The models were trained and tested on the data from each of the 10 participants in the Toadstool dataset separately. The performance of both models was measured by calculating the MAE and RMSE of the output and compared to the MAE and RMSE of a ZeroR classifier.

With the suggestion of using RNNs to capture temporal information from Svoren (2020), our experiment started with implementing a CNN-LSTM model similar to the model implemented by Kahou et al. (2015) for emotion recognition in video. As explained in sections 2.2.8 and 2.2.9, in CNN-LSTM (or CNN-RNNs) approaches, the LSTM takes the features extracted by a CNN over individual frames as inputs and encodes the temporal dynamics.

Research has shown that 3D kernels in 3D CNNs can have a superior ability to extract spatio-temporal features within video frames, as compared to 2D CNNs, even if combined with temporal networks such as RNNs or LSTMs (Haddad, Lézoray and Hamel (2020)). For this reason, we decided to create a 3D CNN to make a comparison of the two spatio-temporal approaches. Our 3D CNN was based on the 3D CNN model architecture created by Haddad, Lézoray and Hamel (2020) for emotion recognition in video. The model took the same input as the previously used CNN-LSTM with ten facial expression frames extracted from 4 seconds of video mapped to the average BVP amplitudes over those 4 seconds.

5.2.2 Model Architectures

Preprocessing

Before inputting the data into the models, both the video and BVP signal had to be preprocessed. A detailed description of the preprocessing steps done on the BVP signals is given in section 7.3.2. How the frames are extracted from the video is explained at the end of section 6.3.2, and how these extracted frames were processed is explained in section 7.2.1.

CNN-LSTM

Resnet-18 was used as the CNN part of the model to extract spatial features from single frames. This model was chosen for its simplicity and good performance in general computer vision tasks (K. He et al. 2015). The layers in the model architecture are shown in table 5.1. The first convolutional layer is followed by a batch normalization layer and a ReLu activation, while the following convolutional layers are only combined with a batch normalization layer.

Layer no:	Type:	Parameters
0	Input	input_shape=(1, 112, 112)
1	2D Convolutional	kernel=(3,3), filters=64, stride=2, padding=3
2	Max Pooling	kernel=(3,3), stride=2
3	[2D Convolutional]x2	kernel=(3,3), filters=64, stride=1
4	[2D Convolutional]x2	kernel=(3,3), filters=128, stride=2
5	[2D Convolutional]x2	kernel=(3,3), filters=256, stride=2
6	[2D Convolutional]x2	kernel=(3,3), filters=512, stride=2
7	AVG Pooling	kernel=(4,4)
8	Flatten	-
9	Linear	input_shape=512, output_shape=300

Table 5.1: The architecture of Resnet-18

The **LSTM** part of the model consists of a stacked LSTM with 2 LSTM cells. At each timestep, i.e., at each of the ten frames, the input to the LSTM layer is the feature vector of length 300 outputted by the CNN.

A stacked LSTM was used because it has been shown to add levels of abstraction of input observations over time (Pascanu et al. 2013). Additionally, it gives the possibility to have a dropout layer between the two LSTM cells, which can prevent overfitting. The output of the

LSTM layer was then sent into two consecutive linear layers with a ReLu activation between, where the last linear layer was the final output layer. Figure 5.1 illustrates how the CNN-LSTM predicts a BVP value based on 10 facial expressions.

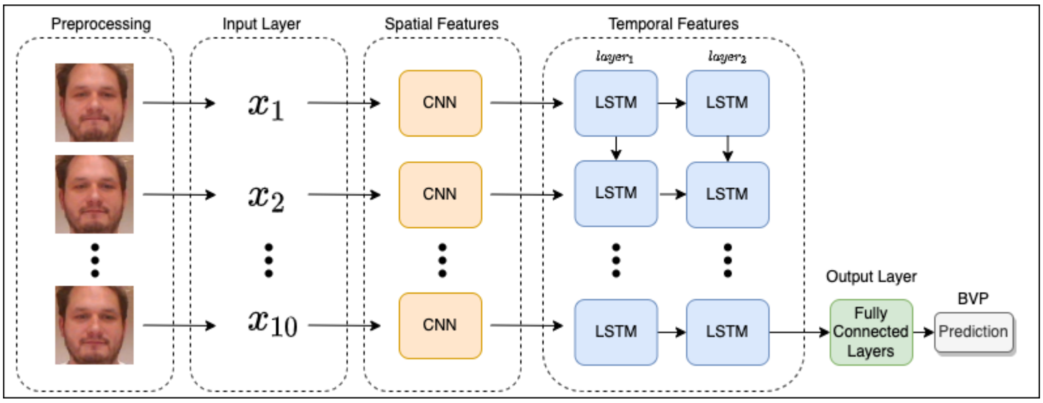


Figure 5.1: CNN-LSTM architecture

3D CNN

As 5.2.3 will show, the 3D-CNN was the best performing model architecture and was therefore used further in our thesis. A more detailed architecture description will therefore be presented in section 7.2.

5.2.3 Results

The experimental plan closely followed the plan of Svoren (2020). However, our results are not directly comparable to the results of Svoren (2020) for two reasons. (1) The input of the 2D-CNN used by Svoren (2020) was a combination of facial expression frames and video game frames, while our models were only inputted facial expression frames. (2) The 2D-CNN predicted a single BVP amplitude, which resulted from averaging 64 measurements over 1 second. In our case, the goal was to predict based on 4-second long sequences, which meant that the averaged BVP values over 1 second were again averaged over 4 seconds. That being said, as a Zero Rule baseline experiment (ZeroR) was done in both contexts, it is highly relevant to compare the performance of the 2D CNN with its respective baseline results with the performances of both spatio-temporal networks with their respective baseline results.

As seen in table 5.2, the MAE was lower or equal in the 2D-CNN compared to the ZeroR MAE in only 3 of 10 participants, while the RMSE of the ZeroR was lower for all participants compared to the RMSE of the 2D-CNN.

As seen in table 5.3, the MAE was lower or equal in the CNN-LSTM compared to the ZeroR MAE in 8 of 10 participants, while the RMSE of the CNN-LSTM was lower for all participants compared to the RMSE of the ZeroR.

As seen in table 5.4 the MAE and RMSE of the 3D-CNN was lower or equal in all participants compared to the MAE and RMSE of the ZeroR.

	CNN		ZeroR	
ID	MAE	RMSE	MAE	RMSE
0	0.076	0.100	0.075	0.099
1	0.104	0.132	0.103	0.131
2	0.071	0.104	0.075	0.100
3	0.050	0.070	0.050	0.070
4	0.078	0.103	0.069	0.094
5	0.091	0.129	0.094	0.121
6	0.091	0.119	0.090	0.116
7	0.110	0.142	0.109	0.139
8	0.061	0.096	0.060	0.090
9	0.126	0.157	0.109	0.134

Table 5.2: Results from predicting BVP amplitudes using CNN Svoren (2020)

	CNN-LSTM		ZeroR	
ID	MAE	RMSE	MAE	RMSE
0	0.022	0.029	0.023	0.033
1	0.049	0.063	0.070	0.087
2	0.070	0.088	0.086	0.113
3	0.032	0.049	0.043	0.065
4	0.067	0.090	0.061	0.095
5	0.080	0.104	0.126	0.162
6	0.060	0.078	0.059	0.078
7	0.082	0.101	0.147	0.177
8	0.065	0.088	0.074	0.108
9	0.098	0.122	0.156	0.194

Table 5.3: Results from predicting BVP amplitudes using CNN-LSTM

	3D CNN		ZeroR	
ID	MAE	RMSE	MAE	RMSE
0	0.022	0.029	0.022	0.032
1	0.050	0.065	0.076	0.097
2	0.068	0.088	0.085	0.112
3	0.031	0.050	0.043	0.065
4	0.059	0.086	0.061	0.095
5	0.081	0.110	0.126	0.162
6	0.059	0.077	0.059	0.078
7	0.077	0.097	0.126	0.153
8	0.063	0.088	0.074	0.108
9	0.099	0.127	0.151	0.186

Table 5.4: Results from predicting BVP amplitudes using 3D CNN

5.3 Discussion

As seen in 5.2.3, the performance of our spatio-temporal networks was significantly better compared to that of the 2D-CNN used by Svoren (2020). Despite being two slightly different contexts, we think that the promising results motivate further investigation and usage of spatio-temporal networks in the context of multimodal emotion recognition.

Chapter 6

Toadstool 2.0 (2022)

Subsection 4.1.1 displayed the content of Toadstool, along with its challenges. In this section we present Toadstool 2.0, a new version of the Toadstool dataset, updated to fit our goal of multimodal emotion recognition. The Toadstool 2.0 dataset includes preprocessed physiological signals that are synchronized with the video of people playing Super Mario, along with labeling of the synchronized data points. This chapter will firstly present a survey regarding human validation of Toadstool 2.0, secondly section 6.2 will present the process of preprocessing the physiological signals, and the synchronization of the data. Thirdly, section 6.3 will display two different approaches to label the dataset, as well as a comparison between the two approaches and the human validation. Lastly, section 6.4 will present the content of the final multimodal dataset.

6.1 Human Validation Of Toadstool

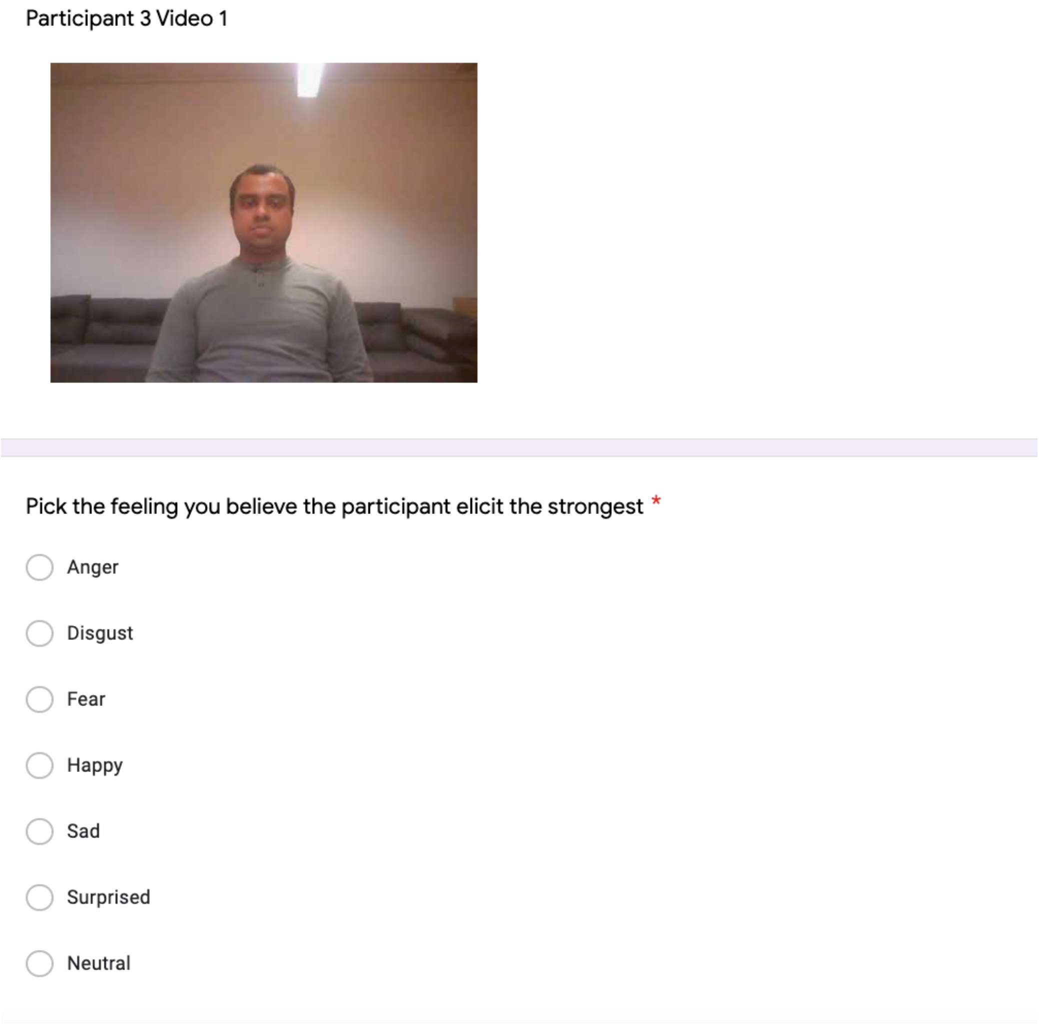
During social interaction, humans employ rich emotional communication channels by modulating their speech utterances, facial expressions, or body gestures. With the underlying assumption that humans are able to detect these emotional channels, along with the previous research of Issa, Fatih Demirci and Yazici (2020) providing 67% human accuracy on the RAVDESS dataset (Livingstone and Russo 2018), a survey of human accuracy was performed to help validate the labeling of Toadstool. This section presents the findings from the survey regarding how humans labeled sequences from the Toadstool dataset. Then, later in the chapter, we will display how the survey contributed to choosing a classification model for labeling Toadstool.

6.1.1 Study Setup

To collect the human validation of the dataset, we asked friends and family to label 30 video sequences from the dataset using google forms. Three sequences, all four seconds long, were randomly extracted from the video of each of the ten participants in the dataset, adding up to 30 sequences. We got 14 people to label the emotion they believed the participants expressed the strongest for each sequence. Regarding whether to label using categorical emotions or a dimensional spectrum, we discovered from our exploration of the fields of FER, PER, and MER that most datasets varied between using dimensional and categorical emotions for classification. With both approaches being valid options, six basic emotions plus neutral were selected due to our previous experience in using categorical emotions. Additionally, to our best knowledge, there were no multi-modal datasets with facial expressions and physiological signals, classified with the six basic emotions plus neutral. Lastly, when we sent out the survey, we made sure to explain the conditions of the video sequences. Namely that the videos consisted of people placed in a room to play Super Mario Bros, with the purpose of collecting facial expressions and physiological data. Figure 6.1 shows the survey structure; a video sequence followed by options to choose from the six basic emotions plus neutral.

6.1.2 Results

While working with Toadstool, we experienced that the majority of the participants evidently expressed a neutral face. This neutral skewness is unsurprisingly displayed in the results from the survey on Toadstool, considering the video sequences were picked randomly. Subsequently, the neutral category was picked as the strongest emotion in 40.2% (169 out of 420) of the votes, additionally, 60% (18 out of 30) of the sequences were labelled neutral.



Participant 3 Video 1

Pick the feeling you believe the participant elicit the strongest *

○ Anger

○ Disgust

○ Fear

○ Happy

○ Sad

○ Surprised

○ Neutral

Figure 6.1: Question from Google Forms

Emotion	Amount	Mean Probability Accuracy
Anger	2	42.9%
Disgust	2	39.3%
Fear	-	-
Happy	8	69.65%
Sadness	-	-
Surprise	-	-
Neutral	18	60.72%
Total	30	53.14%

Table 6.1: Overall stats from the Google Form

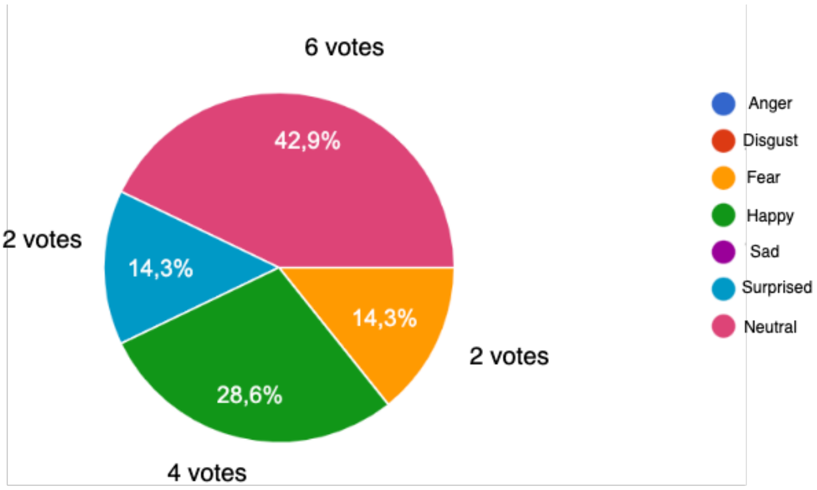


Figure 6.2: Distribution of human labeling of a sequence

Emotion Per Sequence

The 30 sequences were labeled with a mean amount of 3,8 emotions per sequence, showing that the task of labeling videos of people may be a challenging task even for humans. With that said, a person may express several emotions within a time frame of 4 seconds, leaving the raters to choose between several emotions. However, such a scenario is what humans face every day and have to understand. Figure 6.2 presents the distribution of labeling one of the 30 sequences.

6.1.3 Discussion

This survey was provided to help validate the labeling of the Toadstool dataset. Section 6.1 explained the process of how the survey was conducted, as well as its results. However, the survey could have been performed in a better manner. Therefore, we will close up the section by shortly discussing possible missteps and improvements in the survey.

Firstly, in our survey, 14 raters provided their answers, and all raters were aged between 20 and 30. An increased and more diverse group of participants would produce a better basis for validation. Secondly, humans tend to remember things that happened last more clearly than those that came first. This is known as the *recency effect* and may have produced some cognitive bias in the decision-making of the raters. For instance, a rater may label a sequence happy because the last second expressed a happy face, even though the person expressed sadness in the first 3 seconds of the sequence. Thirdly, considering the videos in Toadstool are relatively neutral heavy, a rater may experience labeling neutral three times in a row, resulting in a more extensive search for a non-neutral emotion. Lastly, 3.8 emotions per sequence display the difficulty in labeling a sequence with a single emotion. A possible solution may be to label in a valence-arousal spectrum where the raters can better combine all feelings from the sequence, or let the participants provide self-labels.

6.2 Data preprocessing

To use the physiological data and video provided by Svoren (2020), we had to preprocess the physiological signals and synchronize them with the video. When using the timestamps of the video recording and the sensor recording, we could see that the sensor recording started before the video recording. In addition, the time difference between the start of the sensor data and the start of the video varied across participants. Therefore, to synchronize the video recording with the sensor recording, we needed to use the time difference to find where the video started in the sensor recording. This was done by multiplying the time difference in seconds with the sample rate of the given physiological signal. After synchronizing, the physiological signals were mapped to the video depending on the sample rate; 64 BVP measurements, 4 EDA measurements, 32 Accelerometer measurements, and 1 HR measurement, to one second of video.

The video was recorded at 30 frames per second, but due to the computational cost of iterating through all frames, we decided to extract only the frames planned to train the models. The number of frames to be extracted was three frames per second. This was based on observations of similarities in frames from the same second, and we considered three frames being sufficient to represent the information in one second of the video. How these extracted frames were further used is described separately for the two approaches in sections 6.3.1 and 6.3.2.

6.3 Choosing Labeling Approach

To perform supervised learning, we needed to provide context to the data in the form of labeling. The labeling of the data is a critical process, however creating labeled datasets can be a tedious and demanding task. Furthermore, as seen in 3.6, the majority of the most common datasets are self-assessed, and only a few are solely labeled by external raters. Therefore, for us to operate as external raters did not seem very promising. Along with the fact that it would be quite time-consuming. Additionally, looking at the human accuracy on Toadstool, the raters labeled the sequences with a mean of 3.8 emotions per sequence, highlighting the difficulties of external labeling. Therefore, our approach to labeling the Toadstool dataset was set to rely on machine learning models. Two models were tested and compared, and the following section will introduce the two approaches tested for labeling Toadstool. First, an approach using 2D-CNNs will be presented, considering its popularity and good performance on FER. Then, considering the format of Toadstool being video, we will present an approach using a 3D-CNN.

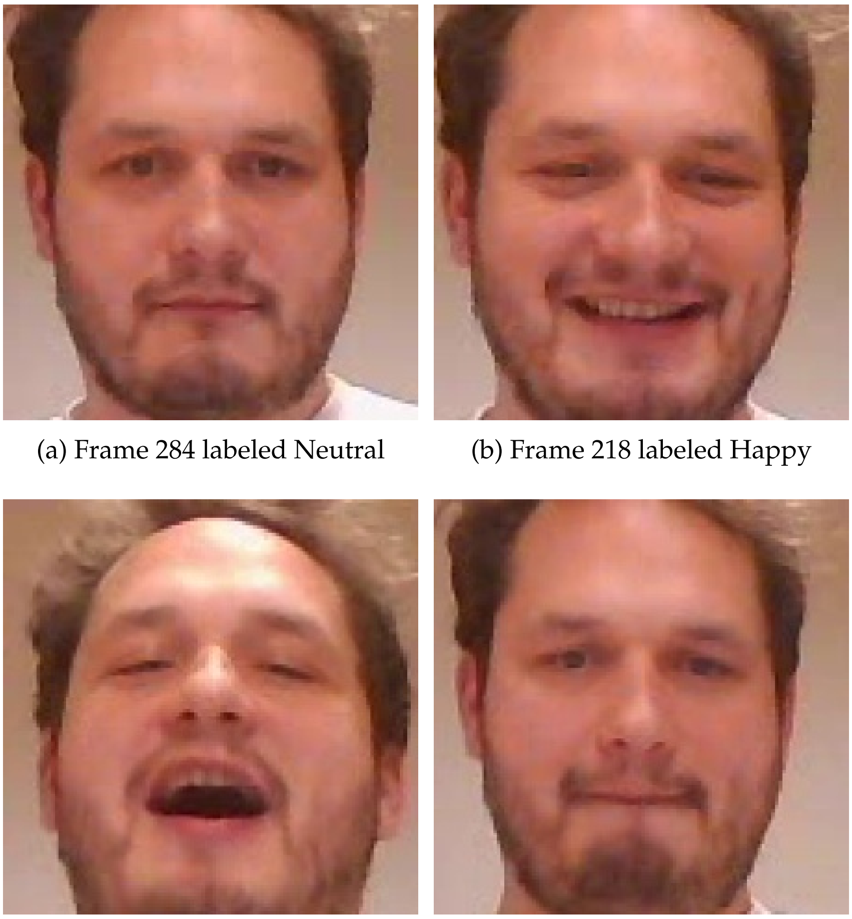
6.3.1 Labelling with 2D-CNN

Through the literature study, we experienced that several 2D-CNN models proved to perform well on FER. For that reason, when choosing a classification model, we decided to label the video using two separate CNNs trained on different dataset, and move on with the most confident label. For the first model, we used the Deepface library (Serengil and Ozpinar 2020). Deepface offers out-of-the-box implementation of state-of-the-art models, including the pipeline of preprocessing and feature extraction. A VGGNet architecture (Simonyan and Zisserman 2014) was selected, due to the fact that it achieved close to state-of-the-art performance when trained on the FER13 dataset (*FER13* (2013)). Regarding the second model, we decided to train it on CK+, with labeled single frames, displayed in table 4.4. Looking at the research of Canal et al. (2022) that compared classical and deep learning approaches on the CK+ dataset, the result showed that CNN outperformed the best classical approach, being an SVM. Therefore, the second model used for labeling Toadstool was also a CNN. ResNet18 was implemented as the 2D-CNN due to its simplicity and good performance in other computer vision tasks (K. He et al. 2015).

From the 3 frames extracted in section 6.2, the first frame was further chosen to label one second of video. Then, we trained both models on their mentioned datasets and sent the frames from all 10 participants into the two models. Table 6.2 displays four random frames and how the model with the highest accuracy decided the label. Further, the correlating emotions are displayed in figure 6.3.

Frame	ResNet Emotion	ResNet Acc.	DeepFace Emotion	DeepFace Acc.	True Label
284	**Neutral**	**91.27%**	Angry	32.52	Neutral
218	**Happy**	**99.98%**	Happy	91.44	Happy
355	Fear	81.43%	**Surprise**	**88.32%**	Surprise
821	Neutral	67.74%	**Angry**	**85.78%**	Angry

Table 6.2: Overview of the labeling of the two models for one participant



(a) Frame 284 labeled Neutral

(b) Frame 218 labeled Happy

(c) Frame 355 labeled Surprised

(d) Frame 821 labeled Angry

Figure 6.3: Labeled emotions of four random frames

The labeled emotions regarding one participant seemed to show a significant disagreement between the two models. Looking at all the participants, we experienced that, on average, the two models had similar labels on approximately 10% of the frames, with the Deepface (Serengil and Ozpinar 2020) model most frequently being the best performer, leading us to question the credibility of the approach.

6.3.2 Labelling With 3D-CNN

Our second approach to labeling the Toadstool dataset was to use a 3D-CNN. A 3D-CNN preserves the temporal aspect of a video sequence and tries to predict an emotion based on the whole sequence. Considering Toadstool consisted of video sequences, such an approach was appealing. We decided to follow the architecture and approach of (Haddad, Lézoray and Hamel (2020)) due to their good performance on the CK+ dataset that we already had available. Looking at the distribution of labels for sequences in CK+, seen in table 4.3, the classes of angry, fear, and sadness contained fewer sequences than average. To overcome the issue of poor distribution of labels and thus avoid overfitting, data augmentation was applied to the mentioned classes. Additionally, three additional steps had to be performed for our model to generalize to the Toadstool dataset. This sub-sub chapter will present how the 3DCNN labeled Toadstool, while the complete architecture of the approach will be presented in chapter 7.

1. Following the approach of Haddad, Lézoray and Hamel (2020) we extracted ten frames from each emotion sequence. However, some emotion sequences from CK+ contained less than ten frames. To overcome this issue, in the approach of Haddad, they added the last frame of the sequence x times until the sequence reached ten frames. However, we experienced that this approach led to a significant number of miss classified neutral labels in our case. When exploring the issue, we discovered that the neutral emotion sequences from CK+ were the most frequent emotion sequence to be missing frames. This resulted in a model learning that static facial expressions at the end of a sequence, should be labeled neutral. Exemplified by a seemingly happy sequence, being labeled neutral due to several similar frames at the end of the sequence. To overcome this issue and make the missing frames less static, we took the last portion of the sequence, corresponding to the number of missing frames, and duplicated them. The approach of Hadadd is displayed in figure 6.4, and our approach to the handling of missing frames is displayed in figure 6.5.

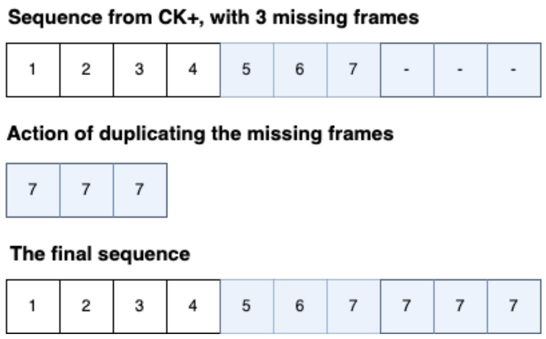

Sequence from CK+, with 3 missing frames
1 2 3 4 5 6 7 - - -
Action of duplicating the missing frames
7 7 7
The final sequence
1 2 3 4 5 6 7 7 7 7

Figure 6.4: Hadadd's approach to handling missing frames from CK+

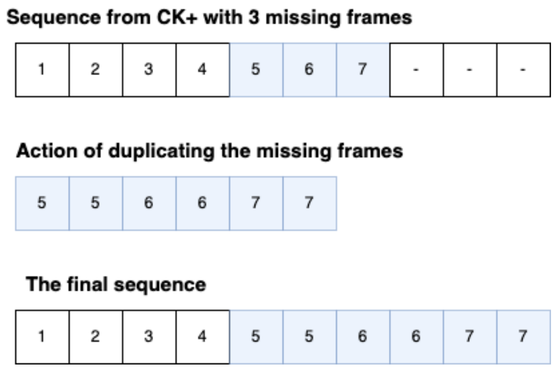

Sequence from CK+ with 3 missing frames
1 2 3 4 5 6 7 - - -
Action of duplicating the missing frames
5 5 6 6 7 7
The final sequence
1 2 3 4 5 5 6 6 7 7

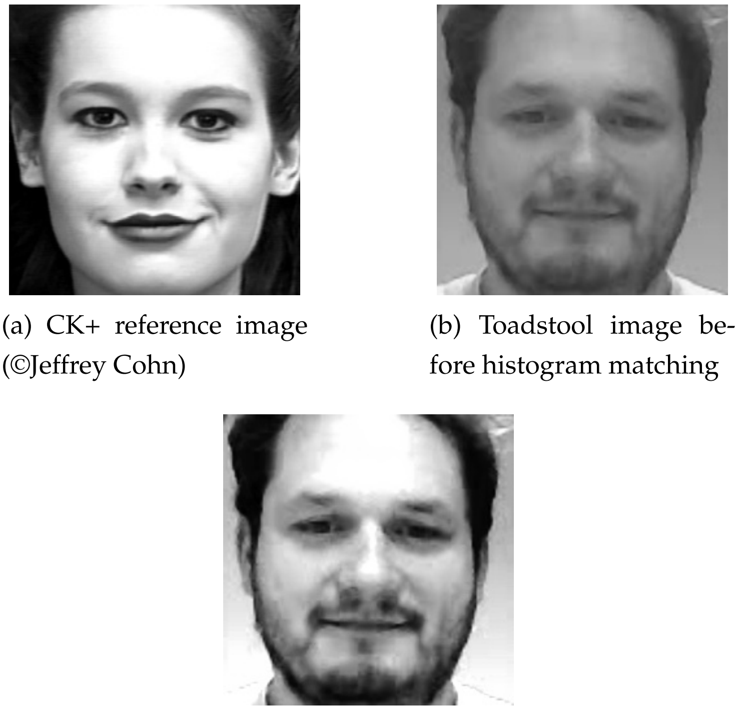
Figure 6.5: Our approach to handling missing frames from CK+

2. In labeling the toadstool dataset, we experienced that the trained 3D-CNN model often was skewed towards one or two labels, resulting in a labeled dataset overrepresented by a few sets of labels. In analyzing a subset of labels and the corresponding facial expression frames, we saw that many sequences were labeled differently compared to the survey answers collected in 6.1, as well as our own judgment on the given sequence. When investigating the issue, we saw that the classification varied considerably depending on the point in the training the state of the model was saved. In our approach, we took the best-performing model, i.e., saving the model state at the epoch with the highest validation accuracy. This meant that despite having a model state which performed well on the CK+ dataset, when we subsequently transferred the model to the context of Toadstool, the classification would be based on a single set of weights which could be skewed toward a few sets of classes. This was due to the nature of model optimization through Stochastic Gradient Descent (SGD). In SGD, the model will update its weights

based on a given learning rate until the model converges to an optimum. One solution to this problem is using *Stochastic Weight Averaging* (Izmailov et al. 2018), where the idea is to save the model state at several points in the optimization, which in our case was the three epochs with the highest validation accuracy in training. The average of the weights from the three best epochs was then used as the model state, which resulted in a more generalized model that was less prone to classifying only a few sets of the classes.

3. Research generally shows that facial expression models are constrained by poor preprocessing steps of facial images. This is because data can vary significantly in chromaticity, image size, face positioning, photography method, lighting, and differences in physical appearance between subjects. These variances in data can occur between samples from the same dataset and, maybe more significantly, between samples across datasets. Normalizing values is often applied to combat differences regarding pixel values across samples, using approximate values for standard deviation and mean of the pixel values of all samples in the dataset. However, as our problem is concerned with normalization across datasets, a normalization of pixel values in the images of CK+ and Toadstool would not work, as the values for standard deviation and mean in pixel values differ between CK+ and Toadstool. To bridge the gap between the domain of CK+ and the domain of Toadstool and, in turn, improve generalization across datasets, we applied an image processing technique called *Histogram Matching* on the images from Toadstool. After the "standard" preprocessing steps; face detection, grayscaling, and resizing, the differences between the images of CK+ and Toadstool were minor; see images 6.6a and 6.6b. Still, these slight differences resulted in substantial differences when classifying, and for that reason, further image processing was needed. The two main differences between the datasets were image contrast levels and image quality. As image quality results from the equipment used in collecting the images, there is little one can do about that. Contrast levels in the images from Toadstool can, on the other hand, be manipulated to match better the contrast levels of the images from CK+ with the mentioned processing technique of histogram matching. Images are character-

ized by a particular pixel intensity distribution, i.e., the intensity histogram. Histogram matching manipulates the intensity histogram of an input image to resemble the distribution of the reference image (Gonzalez and Woods 2018). This is illustrated in 6.6c where histogram matching has been applied on image 6.6b using image 6.6a as reference.



(a) CK+ reference image (©Jeffrey Cohn)

(b) Toadstool image before histogram matching

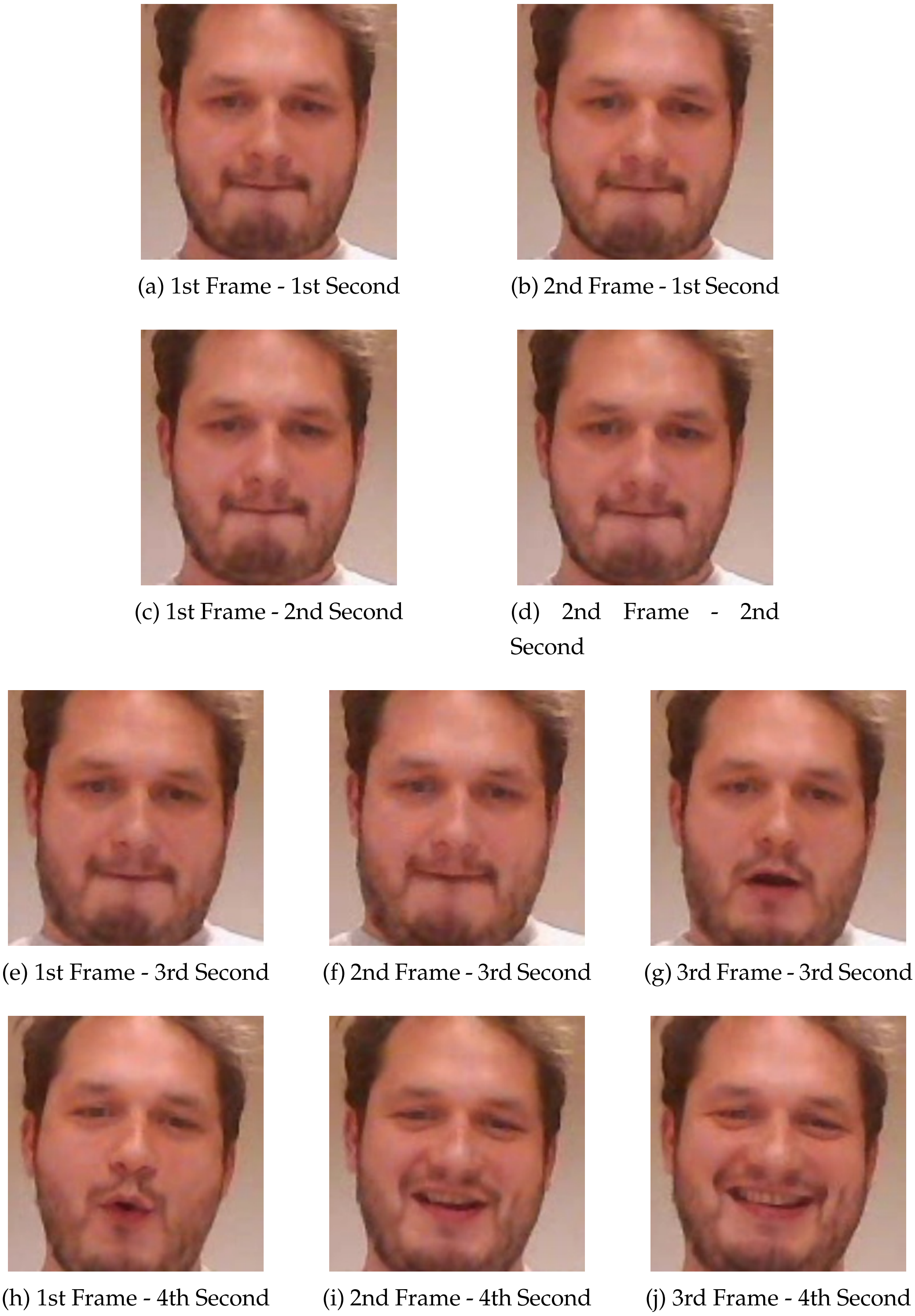
(c) Toadstool image after histogram matching

Figure 6.6: Histogram matching

Our model achieved an accuracy of 95.56% on CK+, and was further used to label 4-second long video-sequences for each participant, as a result of Ekman (2007). In section 6.2 we explained that 3 frames were saved for each second. This meant that 4-second sequences consisted of 12 frames in total. As our model input shape was 10 frames we chose to extract two frames from the first two seconds, and three frames from the last two seconds, where the two frames were the first and last frames in each second, respectively. Lastly, the sequences were extracted using a sliding window approach to ensure we captured as much data as possible. Exemplified by; after extracting a sequence from the first four seconds, the start of the next sequence to be extracted is set to start one second later than the previous. Table 6.3 displays the first four sequences labeled by the 3D-CNN and how the sequences are extracted with one second overlap, while figure 6.7 displays an example of a sequence with ten frames.

Sequences	Time Interval	Label	Accuracy
1505	1505-1509	Fear	86.89%
1506	1506-1510	Fear	62.42
1507	1507-1511	Neutral	56.52%
1508	1508-1512	Happy	97.24%

Table 6.3: Overview of the first four sequences labeled by the 3D-CNN



(a) 1st Frame - 1st Second

(b) 2nd Frame - 1st Second

(c) 1st Frame - 2nd Second

(d) 2nd Frame - 2nd Second

(e) 1st Frame - 3rd Second

(f) 2nd Frame - 3rd Second

(g) 3rd Frame - 3rd Second

(h) 1st Frame - 4th Second

(i) 2nd Frame - 4th Second

(j) 3rd Frame - 4th Second

Figure 6.7: Display of frames from sequence 1508, labeled happy

6.3.3 Choosing Final Labeling Method

After the two models had produced labels for the Toadstool dataset, we used the answers from the human raters to validate the models, deciding which model to use for labeling Toadstool. However, as mentioned, the human validation of Toadstool had shown a distinctive disagreement in labeling, with 3.8 emotions per sequence. Consequently, only the sequences labeled with an accuracy of 50% or more by the raters were used to validate the models, resulting in 23 out of the 30 sequences. Additionally, all sequences labeled neutral by the raters, being 13, were taken out of consideration. The motivation behind this decision was that the video in Toadstool was relatively neutral heavy, meaning the most common expression of the participants was neutral, and we did not want to move on with an approach mainly due to its capability of labeling neutral emotions. Conclusively we ended up with ten labeled sequences from the human validation that we used to compare the two approaches with.

When comparing the two approaches, firstly, the 2D-CNN had to combine its labels to represent a sequence, considering the 3D-CNN and human validation is performed on sequences. To achieve this we combined the labels from 4 frames, adding up to 4 seconds, and labeled it according to the most common label. With each approach representing the same sequences, we could compare their labels, and decide on which approach to use for labeling Toadstool. The comparison revealed a distinctive difference between the human validation and the two classification approaches. Looking at the ten sequences, the 3D-CNN had similar labels on two sequences, whereas the 2D-CNN did not have a single similar label with the human validation. Figure 6.4 shows the comparison of the labels between the 2D-CNN, 3D-CNN, and the human raters, for the ten sequences.

Consequently, both the approaches underachieved based on our expectations. However, with the 3D-CNN relating slightly better to the human validation and the significant relation between physiological signals and 3D-CNNs conducted in an earlier chapter, we ended up labeling the dataset using the 3D-CNN.

Sequence	2D-CNN	3D-CNN	Human Validation
2	Neutral	Surprised	Happy
7	Neutral	Disgust	Happy
9	Neutral	Disgust	Sad
16	Fear	Fear	Happy
18	Neutral	**Anger**	**Anger**
19	Fear	Neutral	Happy
21	Disgust	Surprise	Happy
27	Fear	Happy	Disgust
28	Sad	**Happy**	**Happy**
29	Anger	Neutral	Happy

Table 6.4: Comparison between 2D-CNN, 3D-CNN and Human validation

6.4 Toadstool 2.0

After preprocessing, synchronizing and labeling, the result was a usable multimodal dataset. This section will first present the final content of the labeled dataset.

6.4.1 Content

The final content of Toadstool 2.0 consists of 2097 sequences of 4 seconds, for each of the 10 participants. For each sequence, an emotion label corresponds to 12 frames extracted from video and the 4 physiological signals. Table 6.5 presents 5 examples of the content for one sequence, for one participant, in the Toadstool 2.0 dataset. Further Table 6.6 presents an overview of the Toadstool 2.0 dataset.

Sequence	Frames	BVP	EDA	Accelerometer	HR	Label
1	12	64hz * 4	4hz * 4	32hz * 4	1 * 4	Neutral
2	12	64hz * 4	4hz * 4	32hz * 4	1 * 4	Fear
3	12	64hz * 4	4hz * 4	32hz * 4	1 * 4	Neutral
4	12	64hz * 4	4hz * 4	32hz * 4	1 * 4	Happy
5	12	64hz * 4	4hz * 4	32hz * 4	1 * 4	Anger

Table 6.5: Data for sequences for one participant

Participant	Sequences	Disgust	Fear	Happy	Surprised	Sad	Anger	Neutral
0	2097	101	74	131	4	0	27	1760
1	2097	1190	0	97	0	0	196	614
2	2097	0	122	537	9	42	8	1379
3	2097	0	4	17	1	201	72	1802
4	2097	98	61	248	124	69	32	1465
5	2097	123	35	698	328	80	255	578
6	2097	310	799	315	9	22	41	601
7	2097	232	0	9	100	167	690	899
8	2097	0	1065	6	0	262	0	764
9	2097	66	13	105	129	161	722	901
Total	20970	2120	2173	2163	704	1004	2043	10763
Percent	-	10%	10%	10%	4%	5%	10%	51%

Table 6.6: Overview of the Toadstool 2.0 dataset

6.4.2 Discussion

Toadstool 2.0 was made as a contribution to the Toadstool dataset provided by Svoren (2020). The main contribution of Toadstool 2.0 is its addition of emotion labels, along with synchronizing the video and the physiological signals. This chapter explained the process of making the dataset, along with how it may contribute to the field of multimodal emotion recognition, however the dataset still has potential for improvement.

The labeling of the dataset ended up being handled by a 3D-CNN model, after comparing the model with a 2D-CNN, and validation with the human labeling. However, regardless of the 3D-CNN being the top performer, we believe the labeling process has potential for improvement. Firstly, only two out of ten labels correlating between the 3D-CNN and the human labels, where some of the different labels were close to unanimously chosen by the raters. Secondly, regarding the sequences chosen for the human validation, looking at the label of the sequence appearing a second later than the ones randomly picked, the correlation between the 3D-CNN and human validation was higher. Accordingly, a change in sequence length may be promising. Lastly, the 3D-CNN was trained on the CK+ dataset consisting of 327 posed sequences, where each sequence contains images from a neutral frame (first frame) to peak expression (last frame). However, emotions do not always start as neutral expressions, leading to difficulties classifying a sequence consisting of a single emotion throughout.

By not reducing the sample rate of the physiological signals into one value, we allow for more freedom with the use of the raw signals, whether the user wants to extract the max value, the mean value or other features. Further, with respect to the BVP value, one may explore other peak detection algorithms.

The video from Toadstool was recorded while the participants were playing a video game, however limited research has been done regarding the use of video games as an emotion elicitation method. Looking at the distribution of emotion labels in Toadstool 2.0, the neutral category was significantly large, making us question the emotion elicitation method.

Kessous, Castellano and Caridakis (2009) investigated the imbalance of data in the SAL database (McKeown et al. 2012), and further described two challenges with imbalanced data; the first being that training data with an imbalanced distribution often causes learning algorithms to have poor performance on the minority class, and the second being that the imbalance in the validation/test distribution can affect the performance dramatically. Further, how the emotions of the subjects are provoked (elicit), plays a role in the spectrum you will get the emotions, as well as the intensity of each emotion.

In the work of Pallavicini et al. (2018), they investigated the effectiveness of using virtual reality (VR) survival horror games for emotion elicitation. Results showed that players showed an increased perceived sense of anxiety and happiness when playing a VR video-game, compared to a traditional gaming console. Employing a VR approach for emotion elicitation would be an interesting option to a traditional gaming console. Conclusively, the dataset may have benefited from a method provoking more intense emotions. However, the skewness of facial expression data relates well with the nature of how expressions in the real world appear, making it a relevant problem to face.

Chapter 7

Experiments And Results

This chapter presents the experiments conducted for the purpose of investigating the possible effects the inclusion of physiological signals has on the performance of multimodal emotion recognition classification. The first section describes the experiment plan. The second and third section presents the architecture of the unimodal classifiers implemented to perform multimodal emotion recognition. The fourth section will present the architecture of the multimodal classifier. Lastly, the performance of the classifiers trained on four different participants from the Toadstool 2.0 dataset is presented.

7.1 Experiment Plan

An experiment is designed to prove or disprove a hypothesis. The experiment in this thesis hypothesized that the inclusion of physiological signals will impact the performance of a multimodal emotion recognition classifier. As mentioned in section 2.6, you need to apply a fusion technique in order to combine several modalities. As we experienced from the literature study, decision-level fusion proved to be the most used technique. Therefore, the first part of the experiments involved implementing two unimodal emotion recognition models, a FER classifier and a PER classifier. The two models will also serve as a basis for comparing the multimodal classification. All three classifiers were trained and tested on four participants from the Toadstool 2.0. The choice of not using the remaining six participants was due to their poor distributions of labels. That being said, the data of the four participants was not perfectly distributed either. Therefore, a choice of picking a maximum of 45 random labels for each emotion was taken to minimize the differences between the classes while obtaining a sufficient amount of data. While table 6.6 displays the labels for each participant, table 7.1 displays the final distribution used for training the classifiers.

Further, the data for each participant was split into a training and test set. The reason behind splitting the data into separate training and test sets was to ensure that the model's performance was evaluated on unseen data. Additionally, 3-fold cross-validation was performed for each participant. Figure 7.1 illustrates one out of three iterations, where the data for participant 4 is divided into two training-folds and one test-fold. In the two following iterations, fold one and two will be used as the test-fold, respectively. Lastly, the models ran for 200 epochs for each iteration with an early stopping of 20, saving the epoch with the best accuracy. Conclusively, the accuracy of three epochs was saved for each model and used to take the mean between them.

Lastly, while chapter 6 introduced the 3D-CNN classifier used for FER, this section will fully describe the architecture of our 3D-CNN. The second part of the experiments concerned creating a multimodal model using late fusion to combine the two unimodal models. The multimodal model was also trained and tested on the four participants and later compared to the performance of the FER and PER models.

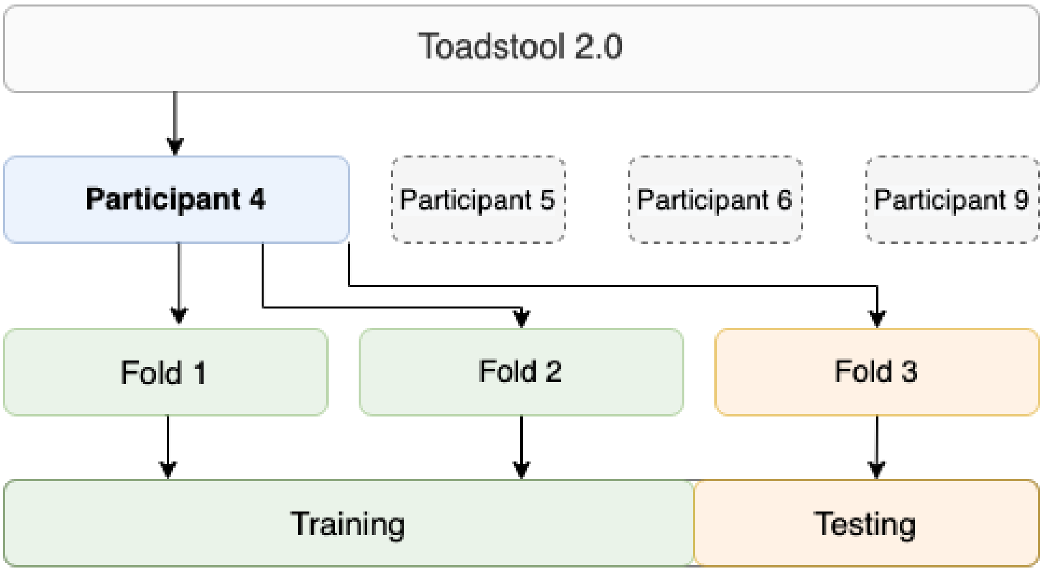


Figure 7.1: Splitting of training set and test set

Participant	Sequences	Disgust	Fear	Happy	Surprised	Sad	Anger	Neutral
4	302	45	45	45	45	45	32	45
5	305	45	35	45	45	45	45	45
6	247	45	45	45	9	22	41	45
9	283	45	13	45	45	45	45	45

Table 7.1: Overview of the Toadstool 2.0 dataset

7.2 FER Model Architecture

In order to measure the effects physiological signals have on emotion recognition, we first had to implement a facial emotion recognition classifier. This section describes the architecture of the FER model, that is, preprocessing techniques, feature extraction methods, and the classification model. The architecture is implemented in compliance with the implementation of Haddad, Lézoray and Hamel (2020). Figure 7.2 illustrates the general concepts of the architecture, from the video sequences to the results from the classification.

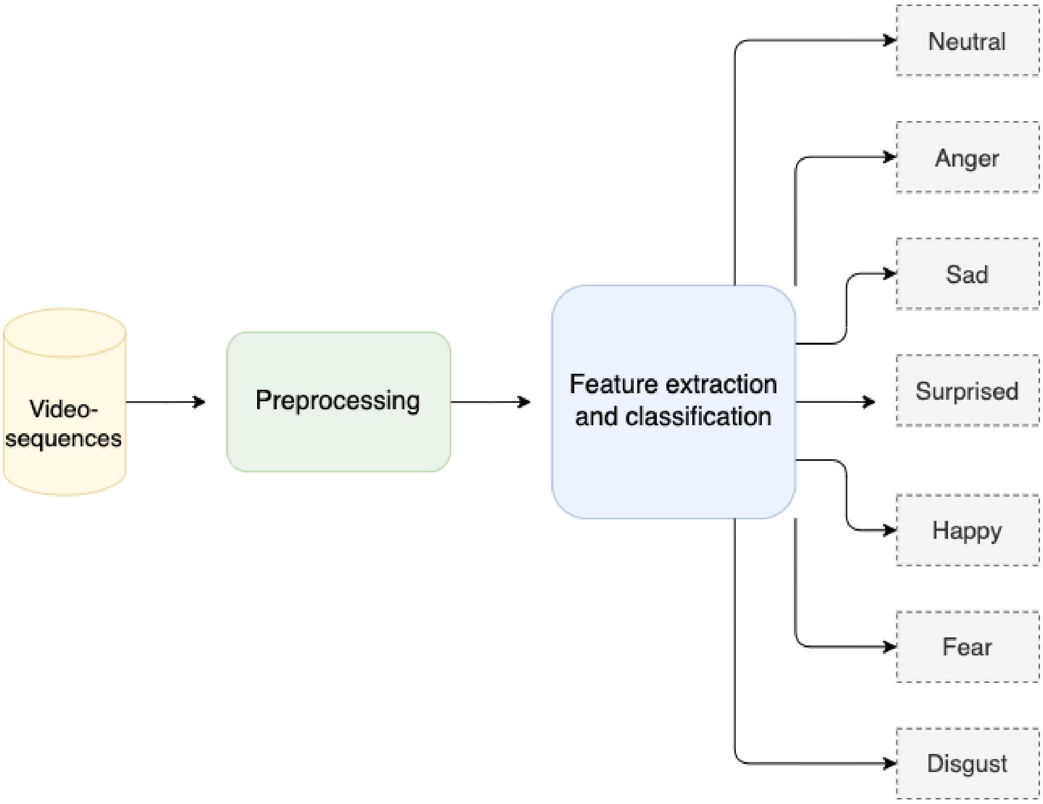


Figure 7.2: FER architecture

7.2.1 Preprocessing

Image processing is an important step in facial emotion recognition and should be performed with care to avoid losing any important features. Fortunately, as experienced throughout this thesis, there exists a common understanding of successful steps within the field of preprocessing. However, datasets are dissimilar to each other and require individual approaches. The PyTorch library (Paszke et al. 2019), described in section 2.9, as well as OpenCV (Bradski 2000), was used for preprocessing of the data. The preprocessing pipeline consisted of the following steps:

1. **Face detection** - As discussed in subsection 3.2.1, the different face detection approaches all perform quite well. Therefore, with the main difference being face detection with multiple faces, and our images consisting of a single face, the Haar Cascade was preferred due to its easy implementation with the use of OpenCV (Bradski 2000).

2. **Resizing** - All images in the dataset were resized to 112x112 pixels to create an equal input size.

3. **Grayscaling** - All images were transformed to grayscale.

4. **Normalization** - Using the Normalize method from PyTorch each value in an image is subtracted by the channel mean and divided by the channel standard deviation.

7.2.2 Feature Extraction and Classification Model

A 3D-CNN was chosen for classification for several reasons. Firstly, as discussed in previous chapters, a 3D-CNN preserves the temporal aspect of a video sequence. Therefore, with Toadstool 2.0 consisting of sequences, using a 3D-CNN seemed promising. Secondly, looking at the experiment conducted in 5, the 3D-CNN proved to correlate with BVP values to a higher degree compared to a 2D-CNN. Consequently, considering our goal was not to implement a top-performing classifier but to investigate the effect of including physiological signals in a multimodal emotion recognition model, and we already had implemented a 3D-CNN proving close to state-of-the-art performance, no other classifications models were tested.

The classification was a multi-class classification task, where the model attempted to classify a sequence of images in the classes of the six basic emotions plus neutral. Feature extraction is performed automatically in a 3D-CNN, however a regularization of the feature extraction part of the network with batch normalization was applied due to its success in reducing internal covariate shift. Selecting the best structure and parameters of a classification model can significantly affect the performance of the model. We follow the approach of Haddad, Lézoray and Hamel (2020), where they end up with well-performing parameters through a thorough exploration of how to optimize the structure and parameters of the network to obtain better performances.

As mentioned, when implementing the 3D-CNN, we followed the specifications laid out in the paper of Haddad, Lézoray and Hamel (2020) as much as reasonably possible. For that reason, the input to the model is a single sequence consisting of 10 frames. The architecture of the 3D-CNN model can be seen in table 7.2.

Layer no:	Type:	Parameters
0	Input	input_shape=(10, 1, 112, 112)
1	3D Convolutional	kernel=(3,3,3), filters=64, padding=(1,1,1)
2	AVG Pooling	kernel=(1, 2, 2), stride=(1, 2, 2)
3	3D Convolutional	kernel=(3,3,3), filters=128, padding=(1,1,1)
4	AVG Pooling	kernel=(2, 2, 2), stride=(2, 2, 2)
5	3D Convolutional	kernel=(3,3,3), filters=256, padding=(1,1,1)
6	3D Convolutional	kernel=(3,3,3), filters=256, padding=(1,1,1)
7	AVG Pooling	kernel=(2, 2, 2), stride=(2, 2, 2)
8	3D Convolutional	kernel=(3,3,3), filters=512, padding=(1,1,1)
9	3D Convolutional	kernel=(3,3,3), filters=512, padding=(1,1,1)
10	AVG Pooling	kernel=(2, 2, 2), stride=(2, 2, 2)
11	3D Convolutional	kernel=(3,3,3), filters=512, padding=(1,1,1)
12	3D Convolutional	kernel=(3,3,3), filters=512, padding=(1,1,1)
13	AVG Pooling	kernel=(1, 2, 2), stride=(2, 2, 2), padding=(0,1,1)
14	Flatten	-
15	Linear	input_shape=(8192), output_shape=(4096)
16	Dropout	Probability=0.2511
17	Linear	input_shape=(4096), output_shape=(2048)
18	Dropout	Probability=0.2511
19	Linear	input_shape=(2048), output_shape=(7)

Table 7.2: The architecture of the 3D-CNN

All convolutional layers have a kernel of size (3, 3, 3), and a padding of (1, 1, 1). All the convolutional layers uses the ReLu, and batch normalization, while the last two linear layers only uses ReLu. The output of the last layer is the number of classes.

7.3 PER Model Architecture

After implementing a FER model, a PER model had to be implemented in order to perform multimodal emotion recognition. This section describes the architecture of the PER model with preprocessing, feature extraction, and classification. Figure 7.3 illustrates the general components of the architecture, from physiological signals in the dataset to the predictions of the classifier. As seen in the figure, a total of 4 physiological signals will be used and combined in order to perform PER. This section will describe each of the architectural components in turn, from preprocessing to classification.

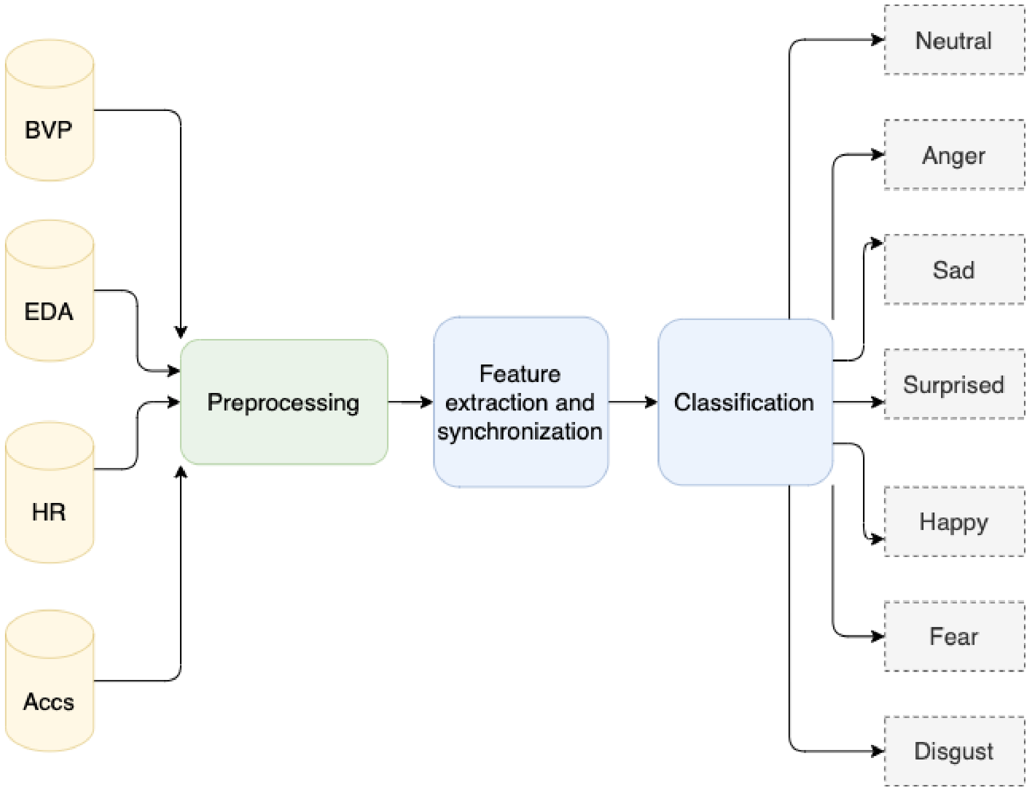


Figure 7.3: PER architecture

7.3.1 Preprocessing

Physiological emotion recognition is a challenging task due to raw physiological signals' complex and subjective nature. In addition to the sensitivity physiological signals have to noise, the data from Toadstool also had to be synchronized with each other to make the data able to combine.

7.3.2 Feature Extraction And Synchronization

The PER model is in itself a multimodal model where each of the four physiological signals acts as its own modality combined through feature level fusion. Because of this, feature extraction and synchronization of the signals were needed before concatenating them into a single input.

The sample rate of BVP was chosen as the standard, resulting in the values for EDA, Accelerometer, and Heart Rate being transformed to 64hz. Using the physiological signal with the highest sampling rate prevents loss of information when downsampling signals. As the sequences were 4 seconds long, the final feature vectors had a length of 256 after concatenating the features from each second. Each of the final features was normalized between 0 and 1. This subsection will present the individual feature extraction and synchronization of the physiological signals

BVP

Specific preprocessing steps were necessary before the BVP data from Toadstool could suit our purpose. As explained in subchapter 2.5.4, we are interested in the amplitude of the BVP signal. Therefore we needed to transform the original BVP signal into a sample set of associated amplitudes. To do this, we applied the approach provided by Svoren (2020). Firstly, a normalization of the signal was performed. As the range of values varies from person to person, the range of the signals was normalized into values between -1 and 1. Secondly, all negative values were replaced with 0, estimating the vasoconstriction to contain only positive values. Thirdly, we found the systolic peaks for each 40hz using a method *find peaks* from the ScyPy library. The distance was set at 40hz because the signal has minor peaks that appear between heartbeats, which are mostly not indicative of the actual vasoconstriction. Lastly, the peak value found for each 40hz distance was set as the current value for all amplitudes appearing in that distance. Conclusively, each 64hz sample ends up consisting of the peak values found. Figure 7.4 shows the stages of the transformation of the BVP signal. In addition to the BVP amplitudes, the mean and max values were extracted from the BVP amplitudes feature, resulting in 3 features for the BVP signal.

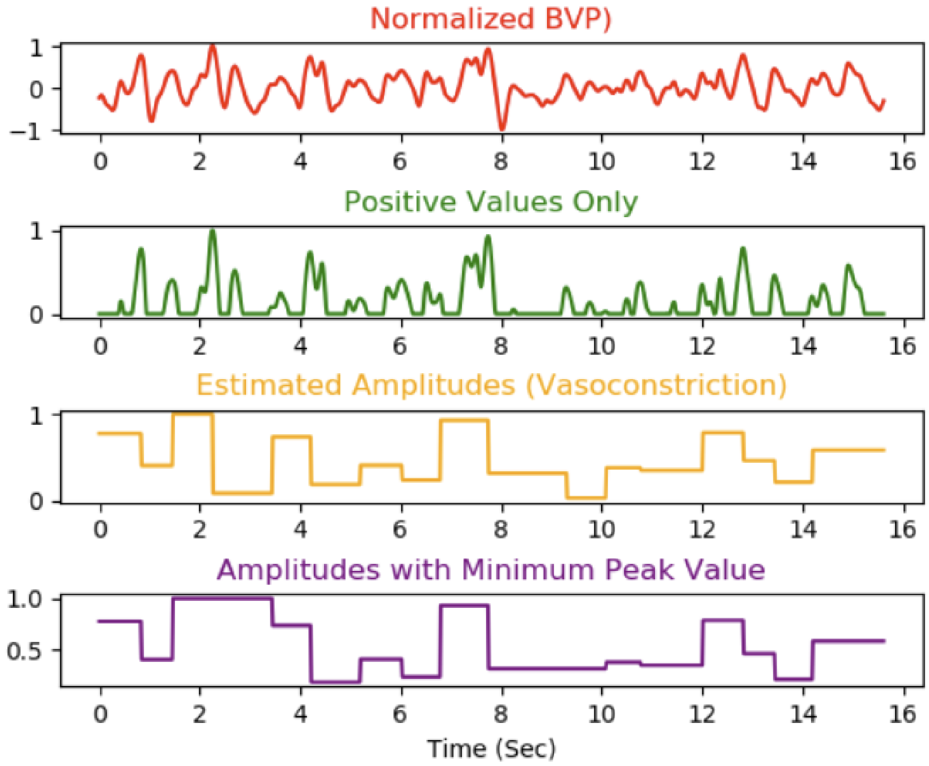


Figure 7.4: The stages of transforming the BVP signal (©Svoren (2020)

EDA

As mentioned in section 2.5.4, EDA is divided into two components, the Tonic component and the Phasic component. As our goal was to learn features related to instantaneous emotional responses, we used the phasic component of the EDA signal because it relates better to the emotional state of a subject in a given instance. The phasic component was collected using the biosignal processing package Neurokit2 (Makowski et al. 2021). Neurokit2 offers a method that extracts Skin Conductance Response (SCR) amplitudes from a raw EDA signal. Figure 7.5 shows how the Skin Conductance Level slowly builds up over time in the tonic component, while the phasic component shows the more immediate SCR amplitudes. The extraction of the SCR amplitudes did not change the dimension of the signal, meaning that the feature vector had to be "stretched" to match the length of 256. This was done by duplicating each value 16 times, as the original sample rate of the EDA signal was 4Hz. The mean and max values of the SCR amplitudes were extracted as additional features, resulting in 3 features for the EDA signal.

Lastly, with respect to the research of Shukla et al. (2019), seen in section 3.3.2, we tried extracting the MFCC feature. However, due to the low sample rate of the EDA-signal, compared to the signal used by Shukla, it was not possible to extract a sufficient amount of data to make the MFCC feature usable.

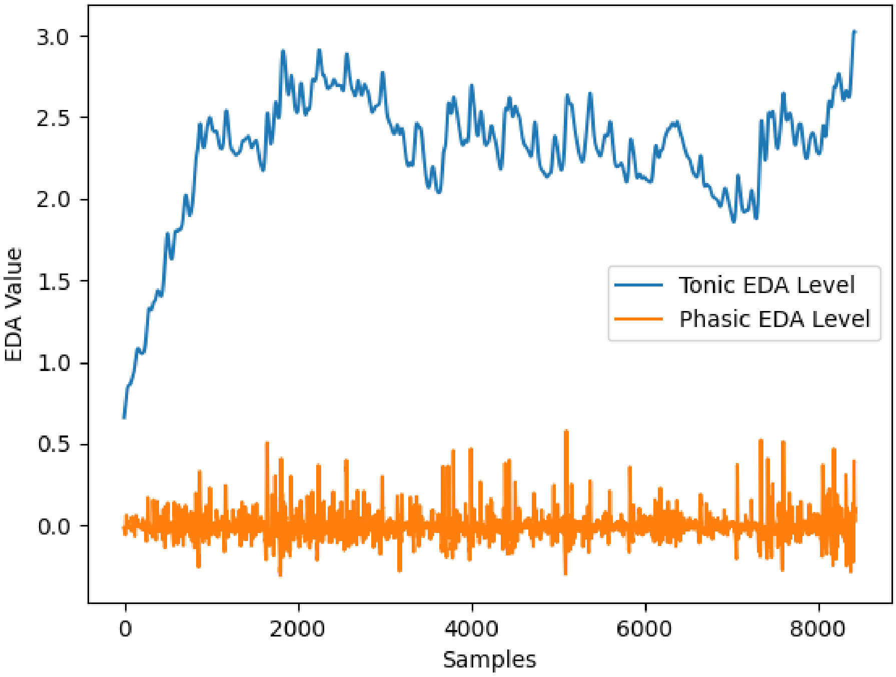


Figure 7.5: Phasic and Tonic component of EDA signal in Toadstool 2.0

Accelerometer

The accelerometer sensor output is three vectors $(\vec{a_x}, \vec{a_y}, \vec{a_z})$ representing the position of the sensor in three-dimensional space. Drawing inspiration from Olsen and Torresen 2016, we decided to look at total movement as a feature, rather than looking at movement along specific axes. Therefore, we transformed the three vectors into one vector describing total movement at the time i, where i=0 is the first sensor recording.

$$a_i = \sqrt{a_{x,i}^2 + a_{y,i}^2 + a_{z,i}^2} \tag{7.1}$$

The sample rate of the raw accelerometer signal was 32Hz, and the transformation into a single vector did not change the length. Therefore, the values were duplicated one time each to match the length of the other signals. Similar to the feature extraction in BVP and EDA, the mean and max values were added as features, resulting in 3 features for the accelerometer signal.

Heart Rate

The sample rate of the heart rate was 1Hz, meaning each value was duplicated 64 times. No further feature extraction or processing was performed besides normalization due to the low sample rate.

7.3.3 Classification Model

Two models were tested and compared to decide on the classification model to use in the multimodal model. As we experienced from the literature study, there has been a recent growth in the use of deep learning methods in PER. However, more classical approaches like SVM have also proven to achieve good results trying to classify physiological signals (Gupta et al. (2022)) (Ayata, Yaslan and Kamasak (2020)). As a result of no clear state-of-the-art approach, we compared the performance of a 1D-CNN and an SVM model and moved on with the best performer. This will be presented in sub-section 7.3.4

Similar to the task of FER in section 7.2, the task of classifying with PER was also a multi-class classification task, where the two models attempted to classify physiological signals in classes of six basic emotions plus neutral. The dataset was initially split into training data and test data, similar to figure 7.1 for each model. When implementing the 1D-CNN and SVM, we followed the specifications laid out by Santamaria-Granados et al. (2019) and Pedregosa et al. (2011), respectively. The input to the models was four physiological signals: BVP, EDA, Accelerometer, and HR, with a total of 10 feature vectors with length 256 for the 1D-CNN. While the SVM was inputted a total of 7 features with length 4. Table 7.3 illustrates the architecture of the 1D-CNN, whereas the SVM is presented solely with text.

1D-CNN

Regarding the input to the 1D-CNN, it used one channel for each feature, adding up to ten channels. All convolutional layers have a kernel of size 3. Further, a stride of 1 was used for all convolutional layers, and a stride of 2 was used for all the pooling layers. ReLu was used in the first convolutional layer, however, batch normalization was performed in each convolutional layer. Further, two linear layers are used, with ReLu and dropout. Lastly, the output of the last layer is the number of classes performed by a linear layer.

Layer no:	Type:	Parameters
0	Input	input_shape=(10, 256)
1	1D Convolutional	kernel=3, filters=64, stride=1
2	Max Pooling	kernel=3, stride=2
3	1D Convolutional	kernel=3, filters=128, stride=1
4	Max Pooling	kernel=3, stride=2
5	1D Convolutional	kernel=3, filters=256, stride=1
6	Max Pooling	kernel=3, stride=2
7	Adaptive Avg Pooling	output_shape=1
8	Flatten	-
9	Linear	input_shape=256, output_shape=64
10	Dropout	Probability=0.25
11	Linear	input_shape=64, output_shape=7

Table 7.3: The architecture of the 1D-CNN

SVM

The SVM model was implemented with the use of the Scikit-learn library (Pedregosa et al. (2011)). The model input was downsampled to the mean and max from each second of each physiological signal (besides HR) measurement. Compared to the input shape of the 1D-CNN which was (10, 256), the input of the SVM was (7, 4) which again was flattened to a single feature vector of length 28. The reduction of dimensionality was done as we observed poor results when using the original input on the SVM, indicating that the SVM may not be able to extract relevant information from such high dimensional features. The SVM used a non-linear kernel function, namely an RBF kernel, as the classification was not a linear classification problem due to the high dimensionality of the input.

7.3.4 Choosing classifier

Two models were tested for PER, a 1D-CNN and an SVM. When deciding on which model to use in the MER model, we compared the performance of both classifiers. Both models were trained and tested using BVP, EDA, HR, and Accelerometer. Table 7.4 presents the accuracy of the two models.

Participant	1D-CNN	SVM
4	**0.407%**	0.400%
5	0.357%	**0.358%**
6	**0.353%**	0.298%
9	**0.367%**	0.352%

Table 7.4: Comparison of 1D-CNN and SVM

The comparison of the two classifiers showed that the 1D-CNN performed better on 3 out of the 4 participants. Therefore, we decided to move on with the 1D-CNN to perform multimodal emotion recognition in combination with the 3D-CNN.

7.4 MER Model Architecture

After the two unimodal classifiers had been implemented, we could combine them to perform multimodal emotion recognition. This section will describe the architecture of the MER model, with the fusion of the two unimodal classifiers. Figure 7.6 illustrates the general components of the architecture, with the combination of the two models.

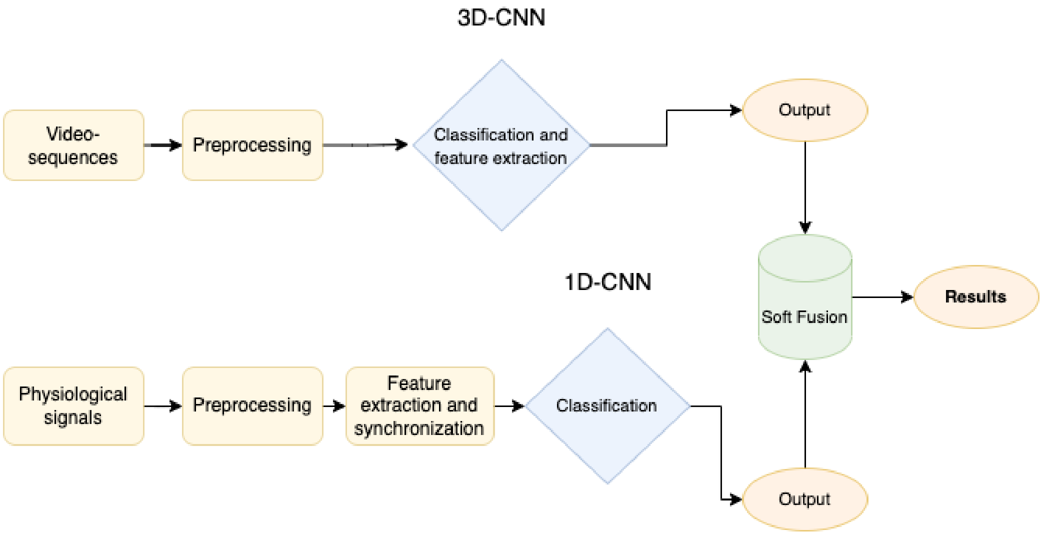


Figure 7.6: MER architecture

7.4.1 Fusion technique

As previously stated, a decision-level fusion technique was applied in our MER model. As the name suggests, the goal of decision-level fusion is to combine the decisions taken by different classifiers to achieve better performance than the individual decisions of the classifiers. The 3D-CNN and the 1D-CNN were the classifiers combined in our approach. Soft fusion, a decision-level fusion method, was used to combine the classifiers. Soft fusion allows one to decide the weight each classifier should have in the final classification. This weight is further multiplied by the softmax of the output for each sequence. The combination of the two modalities is shown in the following equation:

$$Labelmatrix = W_{fer} \cdot Softmax(O_{fer}) + W_{per} \cdot Softmax(O_{per}) \tag{7.2}$$

where O_{fer} and O_{per} is the output from a sequence for the FER model and PER model, respectively. Further W_{fer}, the weight for FER, was set to 0.7, and W_{per}, the weight for PER, was set to 0.3. Following this, the FER model was set to have a greater impact on the classification. This was simply a result of intuitive trial and error, with several combinations of weights.

7.5 Results

All three classification models were trained and tested on four participants from the Toadstool 2.0 dataset. The tables in this section present the performance of the classifiers in terms of accuracy, precision, recall, F1-score and MCC. The section first presents the results from the classification of the FER model, secondly, the results from the classification of the PER model, then the results from the MER model will be presented. Lastly, a comparison of the classifiers will be presented.

The average score of precision, recall, F1-score, and MCC will be displayed for each model, along with accuracy. Additionally, the F1-score for each emotion will be presented for each classifier. The decision to investigate F1-score is based on the fact that precision measures the extent of error caused by false positives. In contrast, recall measures the extent of error caused by false negatives, and in our case, these measures are equally undesirable. Therefore, using F1-score seemed promising, being a harmonic mean of precision and recall. On top of that, observing the distribution of classes in Toadstool 2.0, the dataset appeared to include some class imbalance. When a dataset is unbalanced, i.e. the number of samples in one class is significantly larger than in another class, the accuracy reliability decreases because it would provide an over-optimistic estimation of the classifier's ability to predict the majority class. MCC would produce a high score only if the model obtained good results for all seven emotions (Reinke et al. 2021), acting as a positive addition to the measure of accuracy and F1-score. The MCC score ranges from -1 to 1, where 1 is a complete agreement between labels and prediction, -1 is a complete disagreement, and a score of 0 would indicate that the model is producing random predictions.

7.5.1 FER Classifier

The first part of the experiment consisted of implementing a 3D-CNN to recognize emotion from image sequences. This section will present the results from training and testing the 3D-CNN on the Toadstool 2.0 dataset.

For the FER classifier, participant 4 achieved the best accuracy with

86.4%. Participant 4 also achieved the best F1-score, with 85.8%. This is probably due to participant 4 having the best distribution of emotions. The class that achieved the highest F1-score was the happy class in participant 4, with 98.7%. The MCC-score correlated with the accuracy for all participants.

Participant	Precision	Recall	F1-score	MCC	Accuracy
4	86.4%	85.9%	**85.8%**	**0.84**	**86.4%**
5	82.8%	82.3%	81.4%	0.79	82.2%
6	57.0%	61.8%	58.0%	0.68	73.0%
9	74.1%	77.0%	75.0%	0.82	85.5%

Table 7.5: Results of the 3D-CNN on Toadstool 2.0

Participant	Anger	Disgust	Fear	Happy	Sad	Surprised	Neutral
4	83.7%	93.1%	88.4%	**98.7%**	83.2%	86.3%	67.2%
5	92.2%	81.7%	67.2%	97.6%	80.8%	86.5%	63.5%
6	79.7%	68.9%	82.6%	85.5%	22.2%	0.0%	66.9%
9	91.4%	89.8%	0.0%	87.9%	85.4%	83.9%	86.5%

Table 7.6: Results of F1-score for FER

7.5.2 PER Classifier

The second part of the experiment consisted of implementing a 1D-CNN to recognize emotion from physiological signals. This section will present the results from training and testing the 1D-CNN on the Toadstool 2.0 dataset.

For the PER classifier, participant 4 achieved the best accuracy with 40.0%. Looking at the F1-score for PER, we identified that each participant had at least one class below 10.0%. This is most likely due to a lack of training data for certain classes. The class that achieved the highest F1-score was the happy class in participant 4, with 54.5%. The MCC-score correlated with the accuracy for all participants.

Participant	Precision	Recall	F1-score	MCC	Accuracy
4	36.0%	41.6%	**36.4%**	**0.30**	**40.0%**
5	34.1%	34.3%	30.1%	0.24	34.7%
6	26.3%	27.4%	22.5%	0.19	32.9%
9	28.2%	30.9%	27.9%	0.21	33.9%

Table 7.7: Results of the 1D-CNN on Toadstool 2.0

Participant	Anger	Disgust	Fear	Happy	Sad	Surprised	Neutral
4	54.3%	6.6%	48.6%	54.5%	38.8%	19.8%	31.8%
5	51.3%	26.4%	10.0%	39.7%	19.8%	21.9%	41.5%
6	17.6%	41.1%	42.2%	37.3%	0.0%	0.0%	19.2%
9	18.8%	21.4%	0.0%	46.6%	36.8%	40.0%	31.8%

Table 7.8: Results of F1-score for PER

7.5.3 MER Classifier

The last part of the experiment consisted of implementing a multimodal emotion recognition model by fusing the two classifiers. This section will present the results from training and testing on the Toadstool 2.0 dataset.

For the MER classifier, participant 4 achieved the best accuracy with 87.0%. In addition, participant 4 achieved the best F1-score, with 86.4%. The class that achieved the highest F1-score was the happy class in participant 4, with 98.7%. The MCC-score correlated with the accuracy for all participants. Conclusively, 4 out of 4 participants achieved better accuracy on the MER model. This will be further investigated in the following subsection.

Participant	Precision	Recall	F1-score	MCC	Accuracy
4	88.2%	86.6%	**86.4%**	**0.85**	**87.0%**
5	83.1%	82.5%	81.7%	0.79	82.6%
6	58.8%	61.9%	59.0%	0.68	73.4%
9	74.2%	77.3%	75.3%	0.83	85.8%

Table 7.9: Results of the MER classifier on Toadstool 2.0

Participant	Anger	Disgust	Fear	Happy	Sad	Surprised	Neutral
4	88.8%	96.2%	93.2%	98.7%	82.5%	80.1%	64.9%
5	92.2%	79.9%	68.0%	97.6%	83.4%	87.3%	63.9%
6	81.4%	73.3%	78.4%	89.0%	22.2%	0.0%	68.7%
9	93.9%	91.0%	0.0%	88.4%	83.2%	84.0%	85.8%

Table 7.10: Results of F1-score for MER

7.5.4 Comparing the classifiers

After looking at the results of each of the classifiers, to investigate the effect the PER model had on the MER model, a comparison between the FER model and the MER model was performed. This subsection will compare the accuracy of the classifiers and the F1-score for the different classes for each participant.

Participant 4

When training and testing on participant 4, the FER classifier achieved an accuracy of 86.4%. Further, the MER classifier achieved an accuracy of 87.0%, resulting in an increase of 0.6%. When comparing the F1-score of the two classifiers, anger, disgust, fear, got an improved F1-score in the MER classifier. Happy got a similar F1-score between the two models, while the F1-score of sad, surprised, and neutral decreased. The most improved class was anger, with an increase of 5.1%. The PER classifier performed best on participant 4, with an accuracy of 39.0%

Participant 5

When training and testing on participant 5, the FER classifier achieved an accuracy of 82.2%. Further, the MER classifier achieved an accuracy of 82.6%, resulting in an increase of 0.4%. Additionally, when comparing the F1-score of the two classifiers, fear, sad, surprised and neutral got an improved F1-score in the MER classifier. Anger and happy got a similar F1-score, while the F1-score of disgust decreased. The most improved class was sad with an increase of 2.6%.

Participant 6

When training and testing on participant 5, the FER classifier achieved an accuracy of 73.0%. Further, the MER classifier achieved an accuracy of 73.4%, resulting in an increase of 0.4%. When comparing the F1-score of the two classifiers, anger, disgust, happy and neutral got an improved F1-score in the MER classifier. Surprised and sad got a similar F1-score between the two models, while fear decreased. The most improved class was disgust with an increase of 4.4%.

Participant 9

When training and testing on participant 5, the FER classifier achieved an accuracy of 85.5%. Further, the MER classifier achieved an accuracy of 85.8%, resulting in an increase of 0.3%. Additionally, when comparing the F1-score of the two classifiers, anger, disgust, happy and surprised got an improved F1-score in the MER classifier, with anger being the most improved, with an increase of 2.5%. The F1-score of fear was similar between the two models, while the F1-score of sad and neutral decreased.

Chapter 8

Evaluation And Discussion

8.1 Evaluation

This section provides an evaluation of the experimental results. This section will cover the different aspects of the experiments, from labeling Toadstool to the results of the classifiers.

8.1.1 Labeling Toadstool

Several datasets have been used throughout this thesis. This subsection will contain an evaluation of the mentioned datasets.

CK+ and FER13

CK+ and FER13 were used to train two separate FER classifiers to help with the process of labeling sequences from Toadstool. Both CK+ and FER13 are well-acknowledged datasets commonly used in the field. While both datasets are used for the task of FER and annotated similarly with the basic emotions, they have a significant difference in the amount of training data, with FER13 being the largest dataset. This is probably why the Deepface model achieved the highest accuracy for most frames in Toadstool. Additionally, when comparing the models trained on CK+ and FER13 tested on Toadstool, approximately 10% of the labels correlated. This significant difference may be a result of FER13 being collected from google images, thus including a variety of images, while CK+ is a dataset collected with posed emotion-sequences. This shows that although datasets are annotated similarly, the actual features learned from the dataset can vary significantly.

Toadstool 2.0

Toadstool 2.0 contains data from 10 participants, but only 4 participants were considered in our experiment due to poor distribution of classes among the other 6. The amount and type of data are essential for machine learning models and will heavily impact a model's performance. In Toadstool 2.0, the data size for each participant is identical, however, the distribution of labels varies. Accordingly, the main difference between the participants is the distribution of labels, which is probably the main impact of the different results from the classifiers.

The labels in Toadstool 2.0 result from the best performer between two 2D-CNNs compared to a 3D-CNN, where the 3D-CNN ended up being the model to label the data, as it related slightly better to the human validation. The results from training and testing a FER model on the four best distributed participants in Toadstool 2.0 were quite good. With an average accuracy of 81.7% on the test set, a FER model is indeed able to learn and label a significant amount of the distinctive classes from Toadstool 2.0. Subsequently, our approach to labeling Toadstool proved to provide decent true labels. However, although a model can distinguish between the classes in Toadstool 2.0, it does not prove the labels are correct. For instance, some of the labels that were unanimously picked by the humans, the 3D-CNN labeled differently. Therefore, with the human survey ultimately being the deciding factor in labeling Toadstool, the credibility of the approach is questionable. Furthermore, the labels from the two 2D-CNNs, as well as the labels from the best performing 2D-CNN and 3D-CNN showed a significant disagreement in labels. Conclusively, further investigation of labeling Toadstool is necessary and will be discussed in section 9.2.

8.1.2 Classification Results

Regarding the classification models, only one solution to FER and MER was implemented, while a comparison between two PER models was performed. The final solutions were based on findings and experience from related work and project convenience.

Regarding the FER model, the work of Haddad, Lézoray and Hamel (2020) played an important role. The preprocessing steps chosen were regularly used methods in the field, such as gray-scaling, resizing and normalization. 3D-CNN was chosen as the model due to the experiments provided in chapter 5 and 6.

When choosing a PER model, we tested and compared an SVM and a 1D-CNN and moved on with the best performer, being the 1D-CNN implemented with inspiration from Ayata, Yaslan and Kamasak (2020). As seen in table 7.4 both the 1D-CNN and SVM got 40% as the highest accuracy. In most contexts, such performance can be regarded as a low accuracy score, indicating that predicting seven different emotions based on four different physiological signals is challenging. The preprocessing steps consisted of synchronizing the physiological signals to fit each other and form the signal to contain the essential parts of the data.

The MER model was consequently implemented using the mentioned FER and PER models. A decision-level method was used due to its common usage in the field, explored in chapter 3. The accuracy and general performance increased using a MER model, compared to the unimodal cases. Additionally, with the intention of investigating the effects of including physiological signals in MER, the primary focus of the experiments was not to optimize the classifiers. However, Haddad, Lézoray and Hamel (2020) had already performed an exploration regarding finding well-performing parameters for their model, resulting in us using the same parameters for our model.

Some measures have been taken to avoid model overfitting, such as 3-fold cross-validation, splitting the data into training and test sets, and providing early stopping in training the models. However, with that being said, our model may still be prone to overfitting. The results are also affected by the uneven distribution of labels in the target classes. This is shown by low F1-scores for surprised in participant 6 and fear in participant 9, along with a gap of 27.4% between participant 4 and participant 6 in average F1-score.

Furthermore, the calculation of MCC proved to follow the measurement of accuracy, which, based on the theoretical definition, should only happen if the dataset is balanced. This may be because MCC is originally used in binary classification.

Figure 8.1 shows the confusion matrix for participant 4, classified with the MER model, which yielded the best classifier performance among the participants. Further, figure 8.2 shows the confusion matrix for participant 4, classified with the FER model. Firstly, comparing the true labels between the matrices, six labels were changed between the classifications, where four labels were changed from false to true, and two labels were changed from true to false. Conclusively, the MER model managed to correctly classify two more labels, compared to FER. Anger, disgust, fear and sadness were changed to be correctly classified in the MER classifier, and surprise and neutral were changed from correct to miss-classified. Figure 8.3 shows the confusion matrix for participant 4, classified with the MER model when increasing the dataset size from max 45 samples to max 90. The increase in dataset size resulted in a more significant class imbalance, which again is resembled by the decrease of 16% in accuracy compared to the original class distribution.

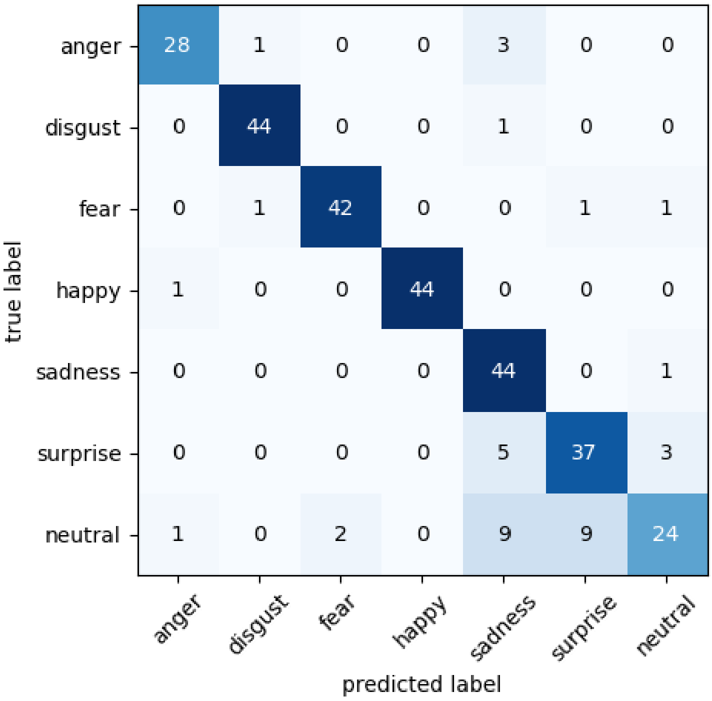


Figure 8.1: Confusion matrix of participant 4 classified with MER

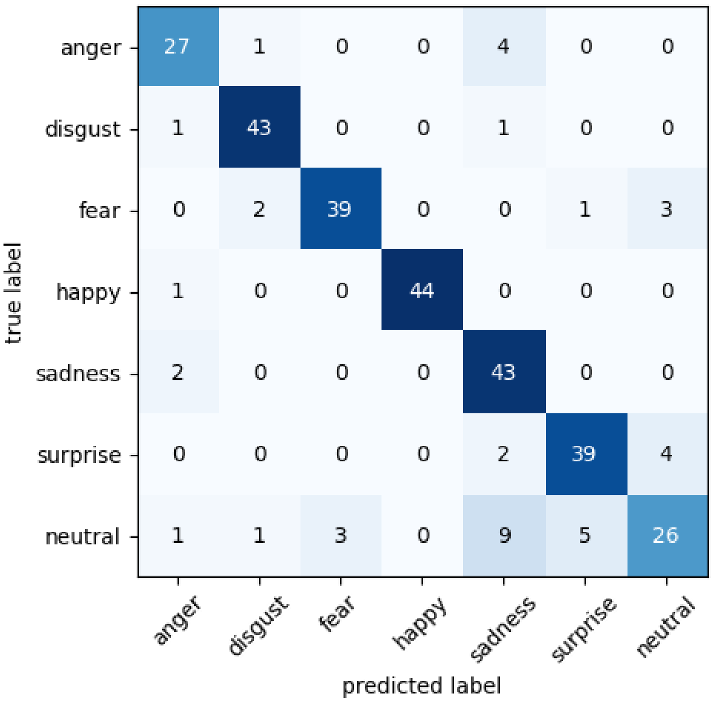


Figure 8.2: Confusion matrix of participant 4 classified with FER

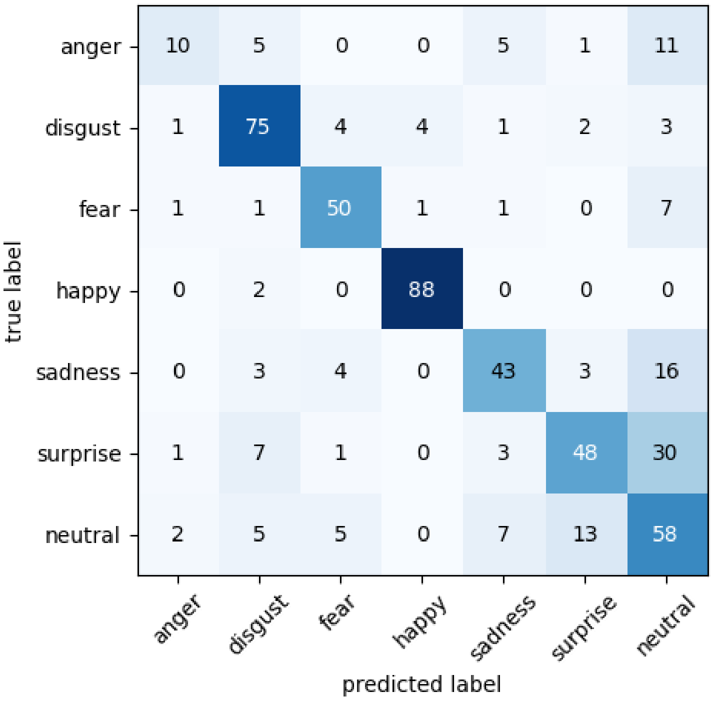


Figure 8.3: Confusion matrix of participant 4 classified with MER with increased dataset

8.2 Discussion

The goal of this thesis was to investigate the effect of multimodality in emotion recognition with the use of facial expressions and physiological signals. Three research questions and one objective were formulated to reach the goal. This will be addressed in this section.

Research Question 1 *What does the literature suggest as promising approaches to MER using facial expressions and physiological signals?*

With respect to Research Question 1, existing research on FER, PER, and MER was investigated. When exploring the field of MER, FER and PER also had to be investigated individually.

Facial Emotion Recognition

When exploring the field of existing research, the field of FER proved to be well documented, with several acknowledged datasets and a common pipeline regarding the model architecture; preprocessing, feature extraction, and classification. Firstly, different datasets require different preprocessing. However, the preprocessing step proved to follow a particular standard pipeline throughout the studies, with face detection, data augmentation, resizing and normalization. Secondly, regarding feature-extraction and classification, exploring the field of study showed a significant usage of deep learning methods, where several of the best performing models used a variation of a CNN.

Physiological Emotion Recognition

The field of PER was not as explored compared to that of FER. However, it was undoubtedly explored to the state of having a common pipeline of action within the field and some commonly used datasets. Nevertheless, all the common datasets in the field used valence and arousal for classification. Subsequently, we believe our introduction of a new dataset with categorical true labels will positively contribute to the field. Further, similar to FER, a model architecture consisting of preprocessing, feature extraction, and classification was used in the field of PER. Firstly, regarding preprocessing, the fact that raw physiological signals are complex, eliminating noise at an early stage proved its importance throughout the field. This was often done with the use of different types of filters. Secondly, the task of choosing and extracting the most valuable features is essential in PER. Popular features for physiological signals include amplitudes, maxima, and minima. Further, some feature selection analyses have been conducted in the search for the best features, however this should be further investigated. Additionally, it is unclear what type and number of modalities to include in a PER model. Lastly, the classification of PER showed similarities to the field of FER, with CNN proving to perform well.

Multimodal Emotion Recognition

Our goal was to investigate what the literature suggested as promising approaches to MER using facial expressions and physiological signals. Both models of FER and PER appeared commonly in the field and proved to perform well in combination with other emotion recognition models such as text and speech, serving as motivation for the use of FER and PER in an MER model. Moreover, the combination of FER and PER also showed promising results. Regarding datasets, there exist some commonly used datasets, however these datasets are collected in a strict scenario due to extensive measuring tools, seen in figure 4.1. Additionally, a dataset with our desired data was not easily accessed.

Late fusion proved to be the most common approach in the field. With the possibility of combining two models, we experienced that the field of MER offered a substantial amount of variations. Exemplified by; The architecture of a FER model has several options regarding preprocessing, feature extraction, and classification, where the classification has a variety of parameters you can tweak. Additionally, the FER model can be trained on various datasets, which can be built in different ways regarding what emotions it includes, how the emotions are elicited, whether the dataset is self-assessed, labeled with raters, or both. All these options regarding FER are also the case for PER. Additionally, little research regarding the comparison of fusion techniques has been conducted and should be explored further.

Lastly, with the growth in easily wearable and accessible technology, the field of MER has experienced a more effortless approach in the creation of datasets. Without undermining the work of Svoren (2020), who unquestionably put in a significant amount of effort in the collection of Toadstool (Svoren et al. 2020), a computer and an E4 Empatica wristband were all that was needed in order to collect a sufficient amount of data.

Objective *Provide ground truths to the multimodal dataset provided by Svoren (2020), using a facial emotion recognition model and human raters.*

Toadstool 2.0 was created for several reasons. Firstly, as discussed in 4.2, we were not able to get access to the DEAP dataset (Koelstra et al. 2011). Furthermore, the IEMOCAP dataset (Busso et al. 2008) had already been explored to a certain degree. Secondly, the natural collection of data in Toadstool (Svoren et al. 2020), along with the fact that we could label Toadstool with basic emotions plus neutral, thereby contributing to the field with a classification not available to our knowledge, seemed meaningful and exciting.

The contributions of Toadstool 2.0 are true labels for 4-second long sequences, with corresponding synchronized physiological data of BVP, EDA, HR, and Accelerometer. The decision to use 4-second long sequences was based on Ekman (Ekman (2007)) and his argumentation regarding how long an emotion lasts. More research should be conducted to understand which length of sequences is most effective for gathering the best ground truths. Additionally, all physiological signals collected by Svoren (2020) were included and synchronized to the video. Keeping the sample rate of the physiological signals was a purposive choice to allow for freedom for further use, regarding what features to extract from the samples.

Lastly, as stated by Svoren (2020), one area that has potential for improvement is the data collection process. Svoren addresses the fact that it might have been a good idea to inform the participants to keep as still as possible, due to the fact that BVP measurements are somewhat sensitive to distortion from movement. Lastly, the evaluation of the distribution in participant 4 addressed the imbalance in the data. As experienced through the literature study, emotion elicitation is vital to acquiring a balanced dataset. From the distribution of Toadstool 2.0 displayed in table 6.6, the dataset was labeled as 51% neutral. Along with the decrease in accuracy when training on 90 samples compared to 45, investigating other forms of emotion elicitation for the data collection seems promising.

Research Question 2 *How are physiological signals related to facial expressions?*

The reasoning behind the experiment done in chapter 5 was that if one could create an ML model that could predict BVP values based on facial expressions with low error, one would show a correlation between the two. Drawing experience from the preliminary experiment of Svoren (2020) we saw that the correlation between BVP and facial expressions in a given instance was small. With this in mind and from exploring emotion theory, we found it natural to instead look at the potential correlation over a given time frame, i.e., with spatiotemporal features. The results of our experiment were promising in the context of comparing it to a ZeroR classification method, and they served as motivation for further investigation of the spatiotemporality of emotion.

The experimental results of chapter 5 serve as a partial answer to how physiological signals are related to facial expressions, as the experiment only looked at BVP as the physiological signal. One could attempt to predict additional physiological signals based on facial expressions, such as EDA, HR, and Accelerometer data. Doing so would give a broader answer to the question, and it could give more insight regarding; to what degree different physiological signals correlate with facial expressions. Additionally, one could attempt several other baseline experiments, questioning to what degree an improvement from a ZeroR classification is indicative of a good model.

Research Question 3 *What are the effects of including physiological signals in multimodal emotion recognition?*

With respect to Research Question 3, the experimental results indicate that physiological signals can be used in addition to facial expressions to slightly increase the predictive power and improve the performance, compared to only using a FER model. While the FER model performed relatively well, correctly labeling a significant amount of test data, the inclusion of physiological signals resulted in a slightly improved model performance. In our experiment we included all physiological signals available, and extracted common features from each. From the SCR and BVP amplitudes, we extracted the mean and max features. Further, from the accelerometer, the mean and max values were extracted from the total movement vector. Lastly, the normalized HR value was used as a feature. However, little optimization of the use of these modalities were explored. For instance, a standard peak algorithm was used to detect peaks in the BVP values. As addressed by Svoren (2020), this is not ideal due to the fact that you might miss out on the most informative BVP amplitude samples. Applying a custom peak detection algorithm should be investigated further.

For each participant, a maximum of 45 samples were used in each class to reduce the imbalance between the classes while maintaining a sufficient amount of training data, in accordance with the amount of data in CK+ (Lucey et al. (2010)). The data was distributed into a training set and a test set consisting of 66.6% and 33.3%, respectively. Additionally, data augmentation could have been applied to the classes including less than 45 samples to balance the split of 45 even further. The data split was a result of the 3-fold cross-validation. However, 4-fold cross-validation, resulting in a distribution of 75% and 25% for the training set and test set, respectively, may be promising, as it allows the model to learn more from the training.

Concerning the MER matrix and table for participant 4, presented in section 8.1 and subsection 7.5.3, neutral had the most miss classified labels, along with a significant lower F1-score, compared to the other classes with maximum amount of training samples (45). With neutral not suffering from lack of training data, the poor performance for classifying the neutral class may be due to variations in the neutral faces. Humans may express their neutral face differently, addressed by the fact that several sequences in the human survey were labeled with a range of four emotions, where neutral was one of the four emotions, displayed in figure 6.2. In other words, dividing neutral from other feelings may be difficult.

When manually investigating the sequences that were changed from being miss-classified by FER to being classified correctly by MER, it was impossible to define whether the change was correct. Meaning, that if a sequence was miss classified as angry by FER and further got correctly classified as fear by the MER model, when manually looking at that sequence, it was impossible to determine whether the change was correct. From this observation, we drew two key points. Firstly, a key motivation behind multimodal emotion recognition is that it allows several models to compliment each other when one lacks information. Which seemingly is what was experienced with our manual check, with not being able to tell whether the label was correct from a facial expression. Secondly, we had to trust our true labels regarding whether the correction was correct or not. However, an addition of self-labeling in the data collection process would provide comprehensive information regarding the credibility of the true labels.

Lastly, in our approach, all physiological modalities available were used, resulting in a slight accuracy increase. However, the effect of different modalities is still unclear, along with the number of modalities to include. In the MER classifier, we experienced sequences being shifted from correctly labeled to miss-labeled, leading us to believe some physiological data provide misleading information to the classifier. More research should be conducted regarding the number of modalities to include in a PER model. In addition, in compliance with the investigation of the effect of BVP in chapter 5, a similar investigation should be conducted regarding the effect of the other modalities.

Ethical considerations

A model is only as good as the data it is trained on. Models in the field of FER, which we have seen operate as a central part in emotion recognition, will therefore naturally be biased towards what it has learned. Simplified by: If every time a person experiences criminal figures, and the criminal figure displays the same facial expression, the person would be biased into classifying similar facial expressions as criminal figures in the future. Like humans using their brains to divide different faces, facial recognition technologies attach numerical values to the human face to divide different faces. Subsequently, FER divides humans into sets of legible signs, which historically has proved to be an injudicious task. In the light of that, Barocas and Selbst (2016) states that if data miners are not careful, the process of data extraction can result in disproportionately adverse outcomes concentrated within historically disadvantaged groups in ways that look a lot like discrimination. For commercial emotion recognition technology to be able to operate without bias, a perfect dataset is needed. However, even if FER could be equally accurate for all people, it may still not be used fairly with just. For example, we see disparate effects when FER is used in policing and judicial systems, operating in a discrimination manner against people of color. Barocas and Selbst (2016) argues that these effects often are unconscious, implicit biases and inertia within society's institutions rather than intentional choices. However, if the world is still suffering from unintentional racial bias, a solution to that may be sensible to target first. Conclusively, a more comprehensive conversation around the deployment of emotion recognition systems is vital, given the state of the world being imperfect.

Chapter 9

Conclusion And Future Work

9.1 Conclusion

Multimodal emotion recognition is a growing field with great potential. Several studies aim to assist in developing effective tools for MER, and the research field is growing. However, there are several challenges linked to the detection of emotion using a multimodal approach, such as the collection and labeling of accurate ground truths. In addition, there is a need for further investigation of the inclusion of physiological signals in MER. The work in this thesis aimed to address this gap by contributing to the Toadstool dataset and investigating the potential of including physiological signals to detect six basic emotions plus neutral.

An overview of information potentially applicable to emotion recognition methods was established by reviewing existing literature related to FER, PER, and MER. Existing literature presents a variety of characteristic approaches to FER, PER, and MER that can be of importance. Future researchers can use this overview as a starting point for investigating how MER can be utilized in emotion recognition. This thesis also contributed to the Toadstool dataset, namely Toadstool 2.0. This dataset aimed to synchronize the video and the physiological data and provide true labels for each sequence. The work from the objective resulted in a labeled dataset that offers extensive freedom of use. Creating Toadstool 2.0 shed light on the difficulties regarding emotion recognition with the disagreements between human raters in the survey, as well as between the FER models, illustrating the ambiguity of subtle emotions.

An investigation of the correlation between facial expressions and BVP was conducted in a preliminary experiment. A correlation was confirmed in the context of our experiment, which motivated further investigation of the combination of physiological signals and facial expressions in emotion recognition. Additionally, the preliminary experiment introduced the usage of spatiotemporal networks, which became central in the following experiments. Based on our literature study and preliminary experiment findings, we created an MER model to investigate the inclusion of physiological signals in emotion recognition. The experiments were conducted by training and testing FER, PER, and MER classification models on the Toadstool 2.0 dataset, showing the performance of the MER model and confirming the usage of the Toadstool 2.0 dataset in parallel.

The FER model provided a good foundation for comparison, achieving high accuracy given a sufficient data. The experimental results showed that the inclusion of physiological signals in the MER model caused a slight improvement compared to FER. Of all tested participants, the participant with the best distribution of true labels achieved the most significant increase in accuracy, corresponding to the findings in the literature study and evaluation regarding the importance of good label distribution. Correspondingly, the participant with the poorest distribution achieved the lowest accuracy. When investigating the significance of the PER model, no common pattern regarding what type of emotion is most prone to the inclusion of PER was discovered. However, in the cases where the PER model changed a miss-classified label to correctly classified, the label suggested by PER was often the top-performing label. Suggesting that the significance of the inclusion of physiological signals depends on the model's performance, and that more research is necessary to improve the PER model. The results from this thesis combine to suggest a potential for including physiological signals with facial expressions to improve performance of emotion recognition. However, more research should be conducted to understand the number of physiological signals to include, what features to extract from the signals, and how these features are most effectively utilized.

9.2 Future Work

The field of emotion recognition is growing, and in recent years, several studies have focused on recognizing emotions with the use of machine learning. However, there is still need for more work in the area. Especially in the field of multimodal emotion recognition, there seems to be a lot of unexplored territory. This section will present suggestions for how to further extend the research conducted in this thesis, or improve upon. As well as other work in the field, independent from this thesis, that the field of multimodal emotion recognition could benefit from.

Approaches to labeling and collecting Toadstool

In this thesis, we used a 3D-CNN, trained on CK+, to label the video sequences from Toadstool. However, there are several ways the labeling could be performed. Considering accurate true labels are a cornerstone in a well-performing machine learning model, future researchers should further investigate other variations of labeling Toadstool. Examples include investigating the sequence length in the model, a shift in the order of images in the sequences in CK+, other classification models, and a more extensive survey regarding human validation. Concerning the data collection of Toadstool, it would also be interesting to perform the collection again with another form of emotion elicitation because the dataset is highly neutral skewed. An example would be to try out other video games. In addition, a process where the participant got the possibility to label their own sequences would be interesting and perhaps result in more accurate labels.

Fusion techniques

Due to the simplicity of decision level fusion one could easily test the effect of other techniques such as product of confidence measures, decision template fusion, Dempster-Shafer and Bayesian belief integration (Ruta and Gabrys 2000).

Feature extraction from PER modalities

In the training of the PER model, simple and common features were chosen from the different modalities. Although the PER model proved to slightly increase the performance of the MER model, investigating the extraction of other types of features may improve the performance. Moreover, in our approach, all physiological signals were used. Following this, investigating the optimal number of modalities to include in the PER model seems promising. Additionally, it is possible to improve the feature selection and selection of modalities through a more extensive investigation of the correlation between, for instance, EDA and facial expressions.

Classification approaches to PER

This thesis put more effort into classifying with FER, compared to PER. For that reason, further effort should be put into improving the PER model, to increase the performance of MER. We ended up using a 1D-CNN for classification after comparing it to a SVM model. However, changes to the entire pipeline should be explored, from preprocessing to classification. As mentioned in section 7.3.2, the PER model is in itself a multimodal model. Looking at other fusion techniques could therefore be interesting. For example, using decision level fusion, one could train a model for each physiological signal and combine the classifications. This approach would remove the need to synchronize the signals prior to classification, which would preserve the original signal.

Bibliography

Abdullah, Sharmeen M.Saleem Abdullah et al. (Apr. (2021)). 'Multimodal Emotion Recognition using Deep Learning'. In: *Journal of Applied Science and Technology Trends* 2.02, pp. 52–58. DOI: 10.38094/jastt20291. URL: https://jastt.org/index.php/jasttpath/article/view/91.

Al Osman, Hussein and Tiago Falk (Feb. (2017)). 'Multimodal Affect Recognition: Current Approaches and Challenges'. In: ISBN: 978-953-51-2915-8. DOI: 10.5772/65683.

Ayata, Deger, Yusuf Yaslan and Mustafa Kamasak (Jan. (2020)). 'Emotion Recognition from Multimodal Physiological Signals for Emotion Aware Healthcare Systems'. In: *Journal of Medical and Biological Engineering* 40. DOI: 10.1007/s40846-019-00505-7.

Barocas, Solon and Andrew D. Selbst ((2016)). 'Big Data's Disparate Impact'. In: *California Law Review* 104.3, pp. 671–732. ISSN: 00081221. URL: http://www.jstor.org/stable/24758720 (visited on 30/04/2022).

Bengio, Y., P. Simard and P. Frasconi ((1994)). 'Learning long-term dependencies with gradient descent is difficult'. In: *IEEE Transactions on Neural Networks* 5.2, pp. 157–166. DOI: 10.1109/72.279181.

Bishay, Mina et al. (2022). *AFFDEX 2.0: A Real-Time Facial Expression Analysis Toolkit*. DOI: 10.48550/ARXIV.2202.12059. URL: https://arxiv.org/abs/2202.12059.

Bota, Patrícia J. et al. (2019). 'A Review, Current Challenges, and Future Possibilities on Emotion Recognition Using Machine Learning and

Physiological Signals'. In: *IEEE Access* 7, pp. 140990–141020. DOI: 10.1109/ACCESS.2019.2944001.

Botchkarev, Alexei (2019). 'A New Typology Design of Performance Metrics to Measure Errors in Machine Learning Regression Algorithms'. In: *Interdisciplinary Journal of Information, Knowledge, and Management* 14, pp. 045–076. DOI: 10.28945/4184. URL: https://doi.org/10.28945%2F4184.

Bradski, G. (2000). 'The OpenCV Library'. In: *Dr. Dobb's Journal of Software Tools*.

Braithwaite, Jason J et al. (2013). 'Guide for Analysing Electrodermal Activity & Skin Conductance Responses for Psychological Experiments'. In: *CTIT technical reports series*.

Brownlee, Jason (Aug. 2019). *Overfitting and Underfitting With Machine Learning Algorithms*. https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/.

Busso, Carlos et al. (Dec. 2008). 'IEMOCAP: Interactive emotional dyadic motion capture database'. In: *Language Resources and Evaluation* 42, pp. 335–359. DOI: 10.1007/s10579-008-9076-6.

Canal, Felipe Zago et al. ((2022)). 'A survey on facial emotion recognition techniques: A state-of-the-art literature review'. In: *Information Sciences* 582, pp. 593–617. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2021.10.005. URL: https://www.sciencedirect.com/science/article/pii/S0020025521010136.

Cheng, Eric-Juwei et al. (2019). 'Deep Sparse Representation Classifier for facial recognition and detection system'. In: *Pattern Recognition Letters* 125, pp. 71–77. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2019.03.006. URL: https://www.sciencedirect.com/science/article/pii/S0167865519300868.

Chicco, Davide and Giuseppe Jurman (Jan. (2020)). 'The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation'. In: *BMC Genomics* 21.1, p. 6. ISSN: 1471-2164. DOI: 10.1186/s12864-019-6413-7. URL: https://doi.org/10.1186/s12864-019-6413-7.

Cicero, M. T. and M. R. Graver ((2002)). *Cicero on the emotions.*

Conneau, Alexis et al. (2020). 'Unsupervised Cross-lingual Representation Learning for Speech Recognition'. In: *CoRR* abs/2006.13979. arXiv: 2006.13979. URL: https://arxiv.org/abs/2006.13979.

Darwin, Charles ((1872)). *The Expression of the Emotions in Man and Animals.*

E4 data - BVP expected signal (n.d.). URL: https://support.empatica.com/hc/en-us/articles/360029719792-E4-data-BVP-expected-signal.

Egger, Maria, Matthias Ley and Sten Hanke (May 2019). 'Emotion Recognition from Physiological Signal Analysis: A Review'. In: *Electronic Notes in Theoretical Computer Science* 343, pp. 35–55. DOI: 10.1016/j.entcs.2019.04.009.

Ekman ((2007)). *Emotions revealed: recognizing faces and feelings to improve communication and emotional life.*

Ekman, P. and W. V. Friesen ((1978)). 'Facial Action Coding System'. In: *M/C Journal* 8.6.

Ekman, Paul ((1992)). 'An argument for basic emotions'. In: *Cognition and Emotion* 6.3-4, pp. 169–200. DOI: 10.1080/02699939208411068. eprint: https://doi.org/10.1080/02699939208411068. URL: https://doi.org/10.1080/02699939208411068.

FER13 ((2013)). https://www.kaggle.com/datasets/msambare/fer2013.

Gasper, Karen, Lauren A. Spencer and Danfei Hu (2019). 'Does Neutral Affect Exist? How Challenging Three Beliefs About Neutral Affect Can Advance Affective Research'. In: *Frontiers in Psychology* 10. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2019.02476. URL: https://www.frontiersin.org/article/10.3389/fpsyg.2019.02476.

Georghiades, A.S., P.N. Belhumeur and D.J. Kriegman (2001). 'From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose'. In: *IEEE Trans. Pattern Anal. Mach. Intelligence* 23.6, pp. 643–660.

Gonzalez, R.C. and R.E. Woods (2018). *Digital Image Processing*. Pearson. Chap. 3, "140–148". ISBN: 978-0-13-335672-4. URL: https://books.google.no/books?id=0F05vgAACA.

Goodfellow, Ian J., Yoshua Bengio and Aaron Courville (2016). *Deep Learning*. http://www.deeplearningbook.org. Cambridge, MA, USA: MIT Press.

Gupta, Priyansh et al. (Jan. (2022)). 'Emotion Recognition During Social Interactions Using Peripheral Physiological Signals'. In: pp. 99–112. ISBN: 978-981-16-3727-8. DOI: 10.1007/978-981-16-3728-5_8.

Haddad, Jad El, Olivier Lézoray and Philippe Hamel ((2020)). '3D-CNN for Facial Emotion Recognition in Videos'. In: *ISVC*.

He, Kaiming et al. (2015). 'Deep Residual Learning for Image Recognition'. In: *CoRR* abs/1512.03385. arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

Hess, Ursula, Rainer Banse and Arvid Kappas (Aug. 1995). 'The Intensity of Facial Expression Is Determined by Underlying Affective State and Social Situation'. In: *Journal of Personality and Social Psychology* 69, pp. 280–288. DOI: 10.1037/0022-3514.69.2.280.

Hinton, Geoffrey E. et al. ((2012)). 'Improving neural networks by preventing co-adaptation of feature detectors'. In: *CoRR* abs/1207.0580. arXiv: 1207.0580. URL: http://arxiv.org/abs/1207.0580.

Hochreiter, Sepp and Jürgen Schmidhuber (Dec. (1997)). 'Long Short-term Memory'. In: *Neural computation* 9, pp. 1735–80. DOI: 10.1162/neco.1997.9.8.1735.

Holz, Hilary et al. (Dec. 2006). 'Research methods in computing: what are they, and how should we teach them?' In: *ACM SIGCSE Bulletin* 38, pp. 96–114. DOI: 10.1145/1189136.1189180.

Hyndman, Rob J. and Anne B. Koehler (2006). 'Another look at measures of forecast accuracy'. In: *International Journal of Forecasting* 22.4, pp. 679–688. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2006.03.001. URL: https://www.sciencedirect.com/science/article/pii/S0169207006000239.

Issa, Dias, M. Fatih Demirci and Adnan Yazici ((2020)). 'Speech emotion recognition with deep convolutional neural networks'. In: *Biomedical Signal Processing and Control* 59, p. 101894. ISSN: 1746-8094.

Izard, Carroll E. (1994). 'Innate and universal facial expressions: evidence from developmental and cross-cultural research.' In: *Psychological bulletin* 115 2, pp. 288–99.

Izmailov, Pavel et al. (2018). 'Averaging Weights Leads to Wider Optima and Better Generalization'. In: *CoRR* abs/1803.05407. arXiv: 1803.05407. URL: http://arxiv.org/abs/1803.05407.

Kahou, Samira Ebrahimi et al. (Nov. (2015)). 'Recurrent Neural Networks for Emotion Recognition in Video'. In: DOI: 10.1145/2818346.2830596.

Kalsum, Tehmina et al. (Jan. (2018)). 'Emotion Recognition from Facial Expressions using Hybrid Feature Descriptors'. In: *IET Image Processing* 12. DOI: 10.1049/iet-ipr.2017.0499.

Kessous, Loic, Ginevra Castellano and George Caridakis (Mar. (2009)). 'Multimodal Emotion Recognition in Speech-based Interaction Using Facial Expression, Body Gesture and Acoustic Analysis'. In: *Journal on Multimodal User Interfaces* 3, pp. 33–48. DOI: 10.1007/s12193-009-0025-5.

Khaireddin, Yousif and Zhuofa Chen (2021). 'Facial Emotion Recognition: State of the Art Performance on FER2013'. In: *CoRR* abs/2105.03588. arXiv: 2105.03588. URL: https://arxiv.org/abs/2105.03588.

Khattak, Asad et al. (Jan. (2022)). 'An Efficient Deep Learning Technique for Facial Emotion Recognition'. In: 81.2, pp. 1649–1683. ISSN: 1380-7501. DOI: 10.1007/s11042-021-11298-w. URL: https://doi.org/10.1007/s11042-021-11298-w.

Koelstra, Sander et al. (Dec. 2011). 'DEAP: A Database for Emotion Analysis Using Physiological Signals'. In: *IEEE Transactions on Affective Computing* 3, pp. 18–31. DOI: 10.1109/T-AFFC.2011.15.

Lee, Jung-Hoon, Hyun-Ju Kim and Yun-Gyung Cheong (Feb. (2020)). 'A Multi-modal Approach for Emotion Recognition of TV Drama Characters Using Image and Text'. In: pp. 420–424. DOI: 10.1109/BigComp48618.2020.00-37.

Li, Jialin, Xueyi Li and David He ((2019)). 'A Directed Acyclic Graph Network Combined With CNN and LSTM for Remaining Useful Life Prediction'. In: *IEEE Access* 7, pp. 75464–75475. DOI: 10.1109/ACCESS.2019.2919566.

Lichtenauer, JEROEN and MOHAMMAD Soleymani (2011). *Mahnob-hci-tagging database.*

Lingenfelser, Florian, Johannes Wagner and Elisabeth Andre (Nov. 2011). 'A systematic discussion of fusion techniques for multi-modal affect recognition tasks'. In: pp. 19–26. DOI: 10.1145/2070481.2070487.

Livingstone, Steven R. and Frank A. Russo (May 2018). 'The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English'. In: *PLOS ONE* 13.5, pp. 1–35. DOI: 10.1371/journal.pone.0196391. URL: https://doi.org/10.1371/journal.pone.0196391.

Lucey, Patrick et al. (July (2010)). 'The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression'. In: pp. 94–101. DOI: 10.1109/CVPRW.2010.5543262.

Luna-Jiménez, Cristina et al. ((2022)). 'A Proposal for Multimodal Emotion Recognition Using Aural Transformers and Action Units on RAVDESS Dataset'. In: *Applied Sciences* 12.1. ISSN: 2076-3417. URL: https://www.mdpi.com/2076-3417/12/1/327.

M., Soleymani and Villaro-Dixon (2017). *Toolbox for Emotional feAture extraction from Physiological signals (TEAP)*.

Makowski, Dominique et al. (Feb. 2021). 'NeuroKit2: A Python toolbox for neurophysiological signal processing'. In: *Behavior Research Methods* 53.4, pp. 1689–1696. DOI: 10.3758/s13428-020-01516-y. URL: https://doi.org/10.3758%2Fs13428-020-01516-y.

Mathisen and Unsvåg (2022a). *Multimodal Emotion Recognition using facial expression and other physiological condition*. https://github.com/andmathisen/multimodal_emotion_recognition.

— (2022b). *Toadstool 2.0*. https://datasets.simula.no/toadstool2.0/.

McCorry, Laurie (Sept. 2007). 'Physiology of the Autonomic Nervous System'. In: *American journal of pharmaceutical education* 71, p. 78. DOI: 10.5688/aj710478.

McKeown, Gary et al. (2012). 'The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent'. In: *IEEE Transactions on Affective Computing* 3.1, pp. 5–17. DOI: 10.1109/T-AFFC.2011.20.

Miranda-Correa, Juan Abdon et al. (2017). *AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups*. DOI: 10.48550/ARXIV.1702.02510. URL: https://arxiv.org/abs/1702.02510.

Moerland, Thomas, Joost Broekens and Catholijn Jonker (Feb. 2018). 'Emotion in Reinforcement Learning Agents and Robots: A Survey'. In: *Machine Learning* 107. DOI: 10.1007/s10994-017-5666-0.

Mou, Wenxuan, Hatice Gunes and Ioannis Patras (June 2019). 'Alone versus In-a-group: A Multi-modal Framework for Automatic Affect Recognition'. In: *ACM Transactions on Multimedia Computing, Communications, and Applications* 15, pp. 1–23. DOI: 10.1145/3321509.

Nakisa, Bahareh et al. (Sept. (2018)). 'Long Short Term Memory Hyperparameter Optimization for a Neural Network Based Emotion Recognition Framework'. In: *IEEE Access* PP, pp. 1–1. DOI: 10.1109/ACCESS.2018.2868361.

Ninu Sreedharan, Rajakumar Boothalingam (Sept. (2018)). 'Grey Wolf optimisation-based feature selection and classification for facial emotion recognition'. In: *IET Biometrics* 7 (5), 490–499(9). URL: https://digital-library.theiet.org/content/journals/10.1049/iet-bmt.2017.0160.

O'Shea, Keiron and Ryan Nash (2015). 'An Introduction to Convolutional Neural Networks'. In: *CoRR* abs/1511.08458. arXiv: 1511.08458. URL: http://arxiv.org/abs/1511.08458.

Olsen, Andreas Fsrøvig and Jim Torresen (2016). 'Smartphone accelerometer data used for detecting human emotions'. In: *2016 3rd International Conference on Systems and Informatics (ICSAI)*, pp. 410–415. DOI: 10.1109/ICSAI.2016.7810990.

Pallavicini, Federica et al. ((2018)). 'Effectiveness of Virtual Reality Survival Horror Games for the Emotional Elicitation: Preliminary Insights Using Resident Evil 7: Biohazard'. In: *Universal Access in Human-Computer Interaction. Virtual, Augmented, and Intelligent Environments*. Ed. by Margherita Antona and Constantine Stephanidis. Cham: Springer International Publishing, pp. 87–101.

Park, Cheul Young et al. (May 2020). *K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations*. Version 0.2.0. Zenodo. DOI: 10.5281/zenodo.3814370. URL: https://doi.org/10.5281/zenodo.3814370.

Pascanu, Razvan et al. (2013). *How to Construct Deep Recurrent Neural Networks*. DOI: 10.48550/ARXIV.1312.6026. URL: https://arxiv.org/abs/1312.6026.

Paszke, Adam et al. (2019). 'PyTorch: An Imperative Style, High-Performance Deep Learning Library'. In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Pedregosa, F. et al. ((2011)). 'Scikit-learn: Machine Learning in Python'. In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Picard, R. W. (1995). *Affective Computing*.

Plutchik, Robert (1982). 'A psychoevolutionary theory of emotions'. In: *Social Science Information* 21.4-5, pp. 529–553. DOI: 10.1177/053901882021004003.

Poria, Soujanya et al. (2018). 'MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations'. In: *CoRR* abs/1810.02508. arXiv: 1810.02508. URL: http://arxiv.org/abs/1810.02508.

Poyiadzi, Rafael et al. ((2021)). 'Domain Generalisation for Apparent Emotional Facial Expression Recognition across Age-Groups'. In: *CoRR* abs/2110.09168. arXiv: 2110.09168. URL: https://arxiv.org/abs/2110.09168.

Raghavan, Jayanthi and Majid Ahmadi (Jan. (2021)). 'Preprocessing Techniques to Improve CNN based Face Recognition System'. In: pp. 1–20. DOI: 10.5121/csit.2021.110101.

Rahmad, Cahya et al. (Jan. (2020)). 'Comparison of Viola-Jones Haar Cascade Classifier and Histogram of Oriented Gradients (HOG) for face detection'. In: *IOP Conference Series: Materials Science and Engineering* 732, p. 012038. DOI: 10.1088/1757-899X/732/1/012038.

Reinke, Annika et al. (2021). *Common Limitations of Image Processing Metrics: A Picture Story*. DOI: 10.48550/ARXIV.2104.05642. URL: https://arxiv.org/abs/2104.05642.

Reisenzein, Rainer (Jan. 1992). *A Structuralist Reconstruction of Wundt's Three-Dimensional Theory of Emotion1*.

Russell, J. A. (Jan. (1980)). 'A circumplex model of affect. Journal of Personality and Social Psychology'. In: pp. 1161–1178.

Ruta, Dymitr and Bogdan Gabrys (Jan. 2000). 'An Overview of Classifier Fusion Methods'. In: *Computing and Information Systems* 7, pp. 1–10.

Salari, Soorena, Amin Ansarian and Hajar Atrianfar (Feb. (2018)). 'Robust emotion classification using neural network models'. In: pp. 190–194. DOI: 10.1109/CFIS.2018.8336626.

'Mean Absolute Error' (2010). In: *Encyclopedia of Machine Learning*. Ed. by Claude Sammut and Geoffrey I. Webb. Boston, MA: Springer US, pp. 652–652. ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_525. URL: https://doi.org/10.1007/978-0-387-30164-8_525.

Santamaria-Granados, Luz et al. ((2019)). 'Using Deep Convolutional Neural Network for Emotion Detection on a Physiological Signals Dataset (AMIGOS)'. In: *IEEE Access* 7, pp. 57–67. DOI: 10.1109/ACCESS.2018.2883213.

Schuller, Björn and Anton Batliner (Oct. (2013)). *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*, pp. 1–321. ISBN: 9781119971368. DOI: 10.1002/9781118706664.

Selvaraj, Jerritta et al. (July (2013)). 'Emotion recognition from Facial EMG signals using Higher Order Statistics and Principal Component Analysis'. In: *Journal- Chinese Institute of Engineers* 37. DOI: 10.1080/02533839.2013.799946.

Seng, Kah and Li-Minn Ang (July (2019)). 'Multimodal Emotion and Sentiment Modeling From Unstructured Big Data: Challenges, Architecture, Techniques'. In: *IEEE Access* PP, pp. 1–1. DOI: 10.1109/ACCESS.2019.2926751.

Serengil, Sefik Ilkin and Alper Ozpinar (2020). 'LightFace: A Hybrid Deep Face Recognition Framework'. In: *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, pp. 23–27. DOI: 10.1109/ASYU50717.2020.9259802. URL: https://doi.org/10.1109/ASYU50717.2020.9259802.

Shouse, Eric (Dec. (2005)). 'Feeling, Emotion, Affect'. In: *M/C Journal* 8.6.

Shu, Lin et al. (2018). 'A Review of Emotion Recognition Using Physiological Signals'. In: *Sensors* 18.7. ISSN: 1424-8220. DOI: 10.3390/s18072074. URL: https://www.mdpi.com/1424-8220/18/7/2074.

Shukla, Jainendra et al. (Feb. (2019)). 'Feature Extraction and Selection for Emotion Recognition from Electrodermal Activity'. In: *IEEE Transactions on Affective Computing* PP, pp. 1–1. DOI: 10.1109/TAFFC.2019.2901673.

Simonyan, Karen and Andrew Zisserman (2014). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv: 1409.1556 [cs.CV].

Svoren, Henrik ((2020)). 'Emotional Mario - Using Super Mario Bros. to Train Emotional Intelligent Machines'. MA thesis.

Svoren, Henrik et al. (2020). 'Toadstool: A Dataset for Training Emotional Intelligent Machines Playing Super Mario Bros'. In: *Proceedings of the 11th ACM Multimedia Systems Conference*. MMSys '20. Istanbul, Turkey: Association for Computing Machinery, pp. 309–314. DOI: 10.1145/3339825.3394939.

Tan, Clarence et al. (Sept. 2020). 'FusionSense: Emotion Classification Using Feature Fusion of Multimodal Data and Deep Learning in a

Brain-Inspired Spiking Neural Network'. In: *Sensors* 20, p. 5328. DOI: 10.3390/s20185328.

Tan, Ying et al. ((2021)). 'A multimodal emotion recognition method based on facial expressions and electroencephalography'. In: *Biomedical Signal Processing and Control* 70, p. 103029. ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2021.103029. URL: https://www.sciencedirect.com/science/article/pii/S1746809421006261.

Unsvåg, Elise and Björn Gambäck (2018). 'The Effects of User Features on Twitter Hate Speech Detection'. MA thesis. DOI: 10.18653/v1/W18-5110.

Virtanen, Pauli et al. (2020). 'SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python'. In: *Nature Methods* 17, pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

Wei, Wei et al. (July 2019). 'Multi-modal facial expression feature based on deep-neural networks'. In: *Journal on Multimodal User Interfaces* 14. DOI: 10.1007/s12193-019-00308-9.

Wu, Chung-Hsien, Jen-Chun Lin and Wen-Li Wei (2014). 'Survey on audiovisual emotion recognition: databases, features, and data fusion strategies'. In: *APSIPA Transactions on Signal and Information Processing* 3, e12. DOI: 10.1017/ATSIP.2014.11.

Xie, Baijun, Mariia Sidulova and Chung Hyuk Park (2021). 'Robust Multimodal Emotion Recognition from Conversation with Transformer-Based Crossmodality Fusion'. In: *Sensors* 21.14. ISSN: 1424-8220. URL: https://www.mdpi.com/1424-8220/21/14/4913.

Yang, D. et al. (Jan. (2018)). 'An Emotion Recognition Model Based on Facial Recognition in Virtual Learning Environment'. In: *Procedia Computer Science* 125, pp. 2–10. DOI: 10.1016/j.procs.2017.12.003.

Zhang, Jianhua et al. (2020). 'Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review'. In: *Information Fusion* 59, pp. 103–126. ISSN: 1566-2535. DOI: https://doi.

org/10.1016/j.inffus.2020.01.011. URL: https://www.sciencedirect.com/science/article/pii/S1566253519302532.

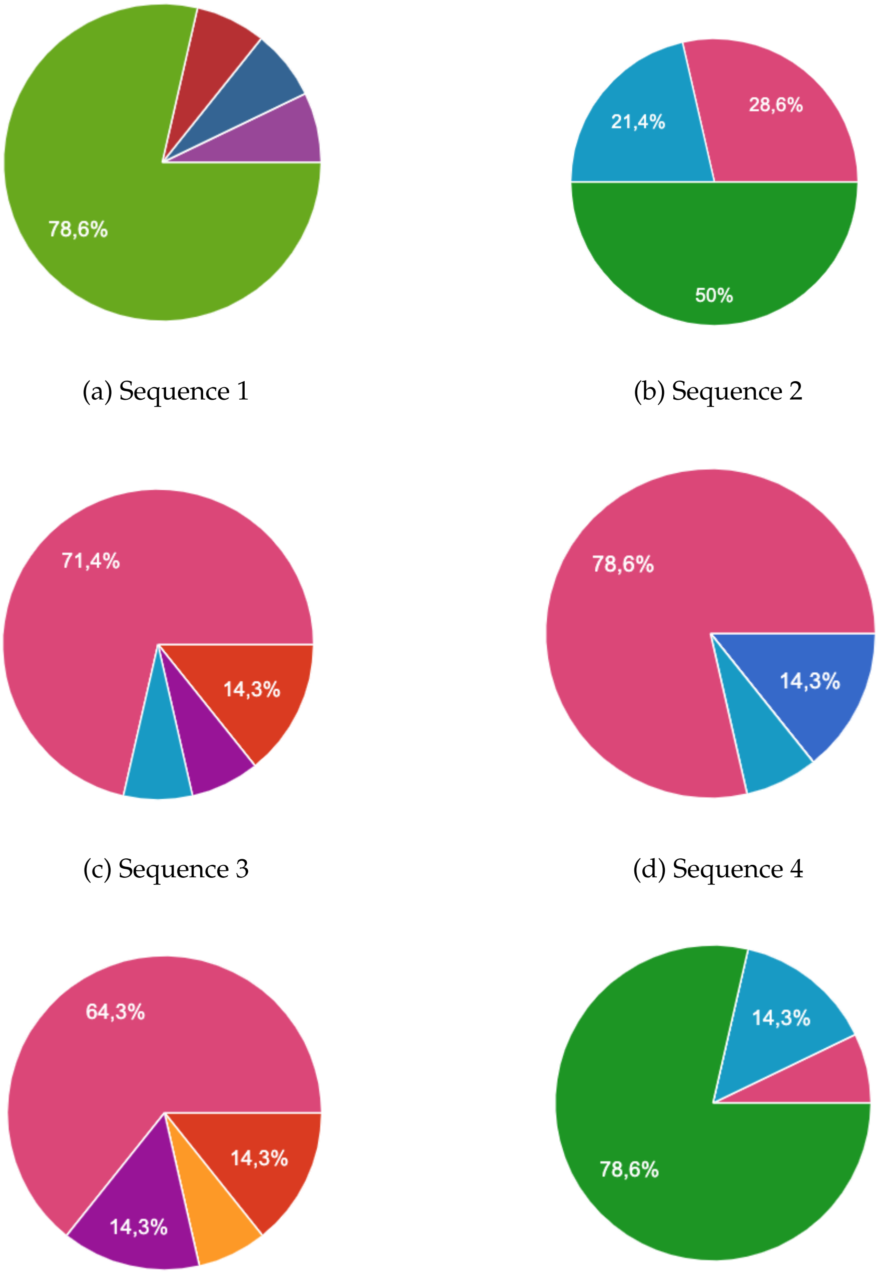
Zhao, S. et al. (Feb. 2019). 'Personalized emotion recognition by personality-aware high-order learning of physiological signals'. English. In: *ACM Transactions on Multimedia Computing, Communications, and Applications* 15.1S. ISSN: 1551-6857. DOI: 10.1145/3233184.

Zhong, Boxuan et al. ((2017)). 'Emotion recognition with facial expressions and physiological signals'. In: *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–8. DOI: 10.1109/SSCI.2017.8285365.

Appendices

Appendix A

Sequences From Human Validation



(a) Sequence 1

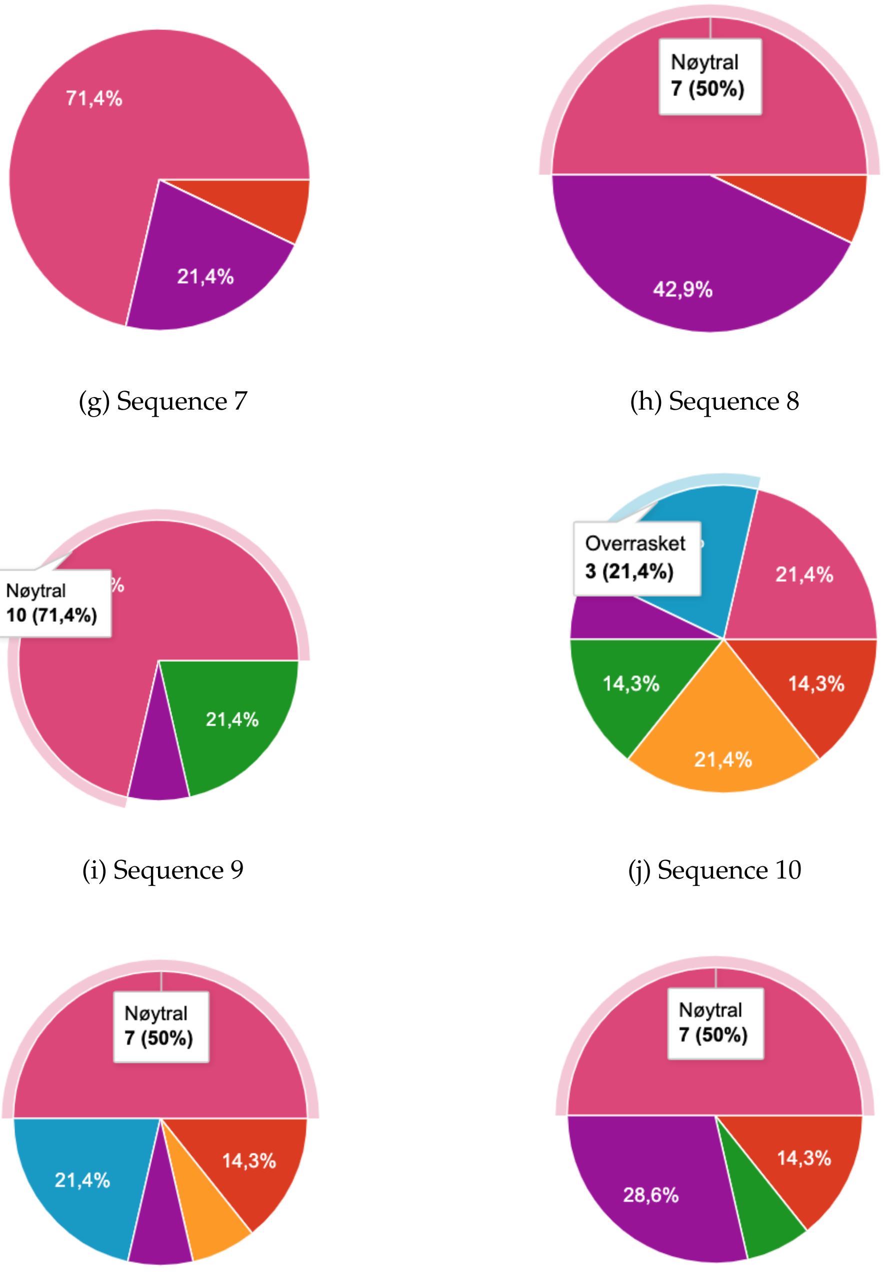
(b) Sequence 2

(c) Sequence 3

(d) Sequence 4

(e) Sequence 5

(f) Sequence 6



(g) Sequence 7

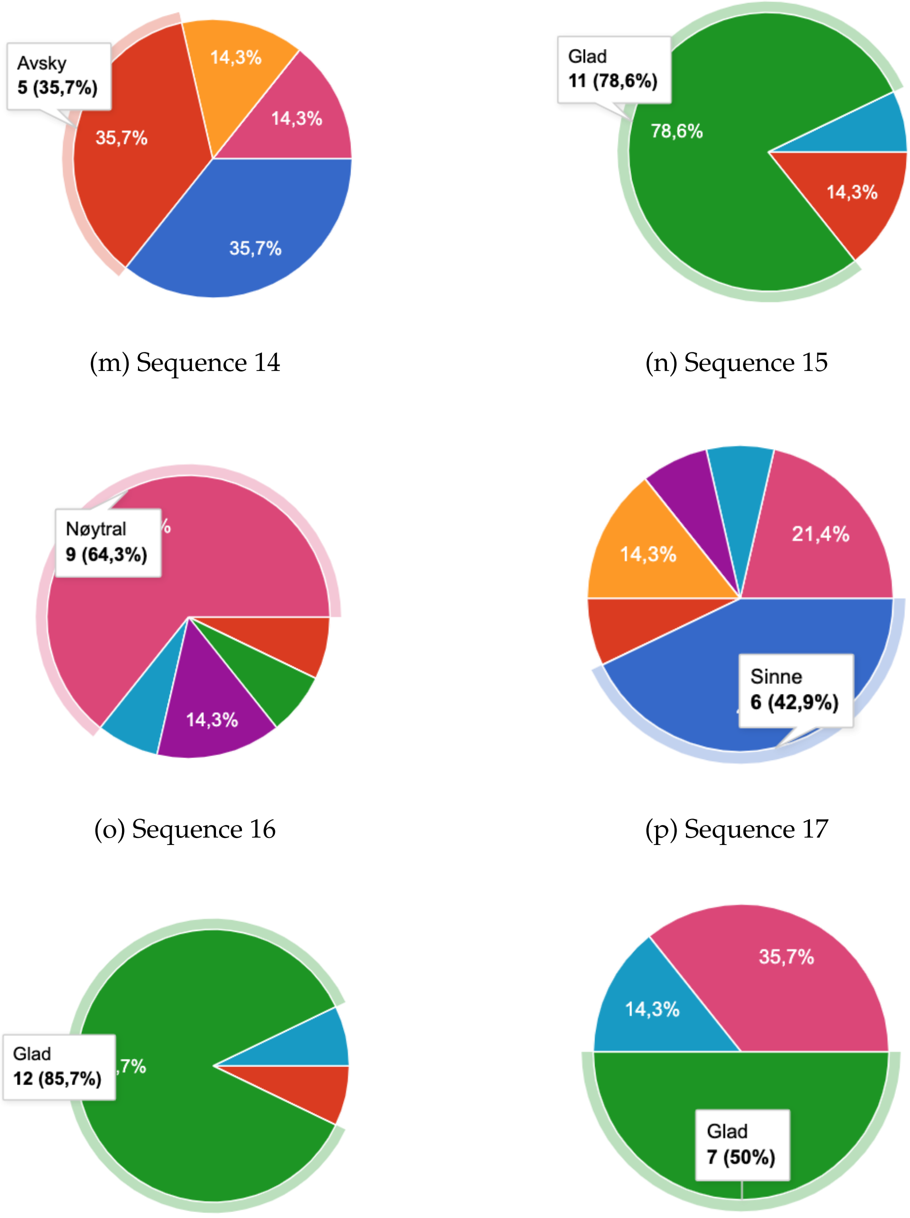
(h) Sequence 8

(i) Sequence 9

(j) Sequence 10

(k) Sequence 11

(l) Sequence 13



(m) Sequence 14

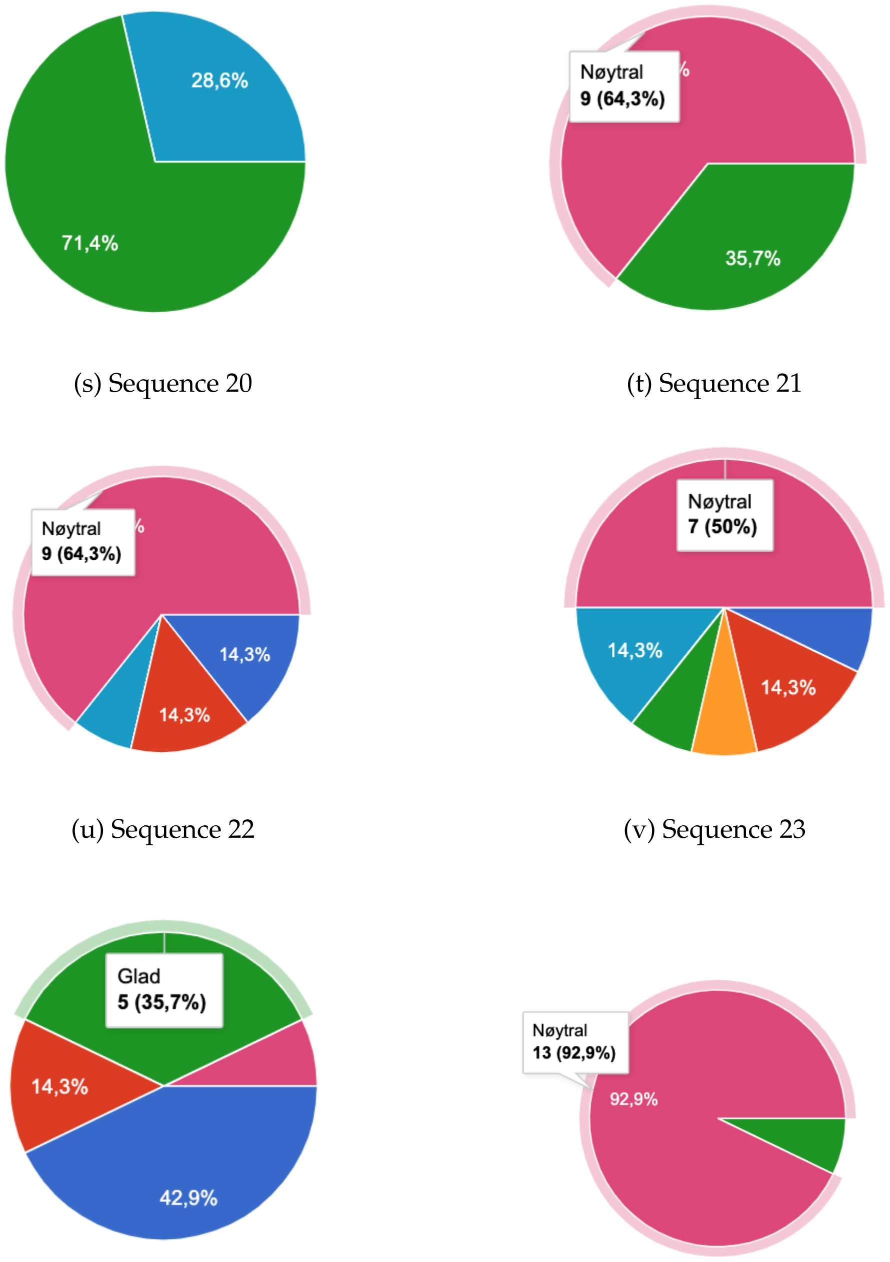
(n) Sequence 15

(o) Sequence 16

(p) Sequence 17

(q) Sequence 18

(r) Sequence 19



(s) Sequence 20

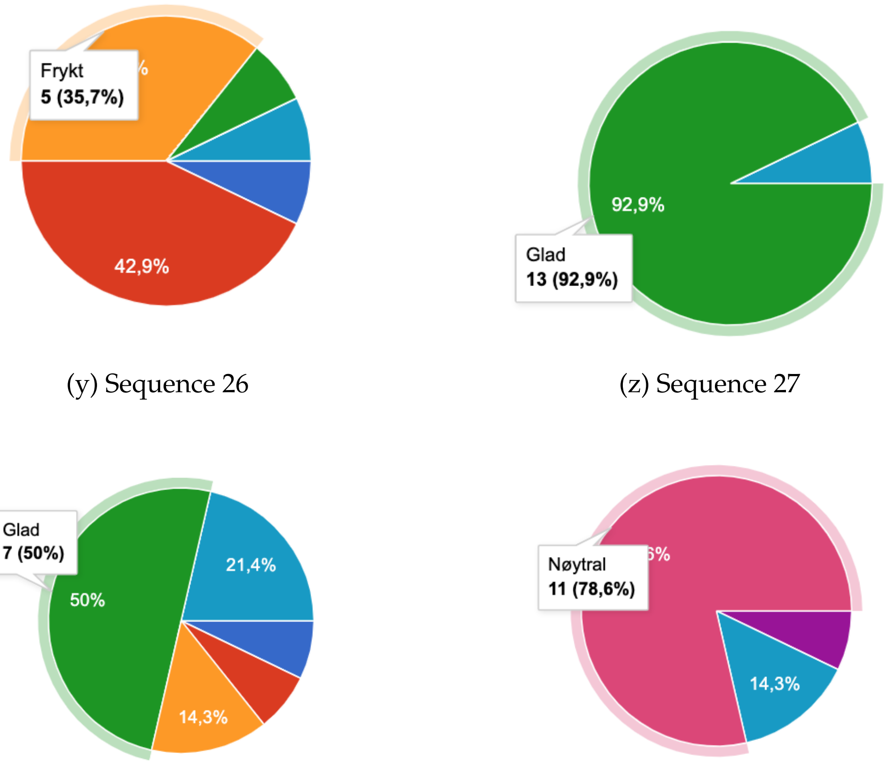
(t) Sequence 21

(u) Sequence 22

(v) Sequence 23

(w) Sequence 24

(x) Sequence 25



(y) Sequence 26

(z) Sequence 27

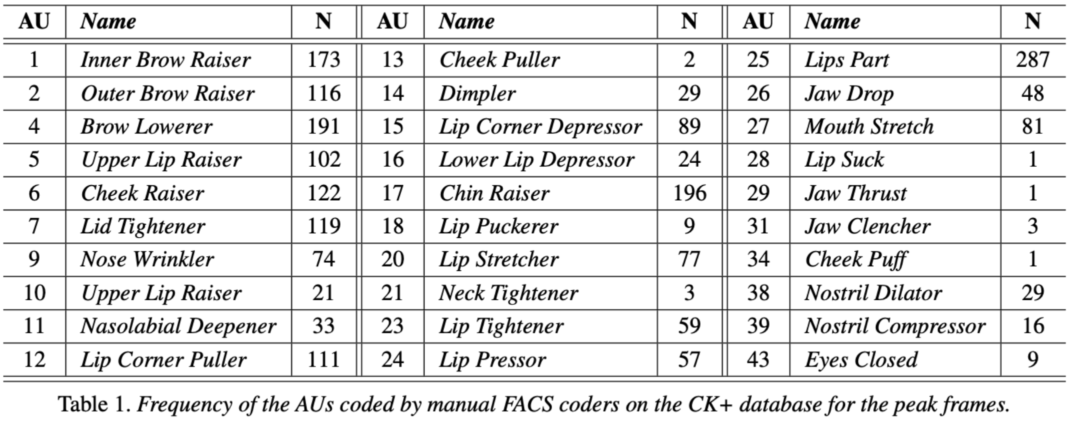
() Sequence 28

() Sequence 29

Figure A.-3: Distribution of sequences from the human validation

Appendix B

FACS Coding Table



AU	*Name*	N	AU	*Name*	N	AU	*Name*	N
1	*Inner Brow Raiser*	173	13	*Cheek Puller*	2	25	*Lips Part*	287
2	*Outer Brow Raiser*	116	14	*Dimpler*	29	26	*Jaw Drop*	48
4	*Brow Lowerer*	191	15	*Lip Corner Depressor*	89	27	*Mouth Stretch*	81
5	*Upper Lip Raiser*	102	16	*Lower Lip Depressor*	24	28	*Lip Suck*	1
6	*Cheek Raiser*	122	17	*Chin Raiser*	196	29	*Jaw Thrust*	1
7	*Lid Tightener*	119	18	*Lip Puckerer*	9	31	*Jaw Clencher*	3
9	*Nose Wrinkler*	74	20	*Lip Stretcher*	77	34	*Cheek Puff*	1
10	*Upper Lip Raiser*	21	21	*Neck Tightener*	3	38	*Nostril Dilator*	29
11	*Nasolabial Deepener*	33	23	*Lip Tightener*	59	39	*Nostril Compressor*	16
12	*Lip Corner Puller*	111	24	*Lip Pressor*	57	43	*Eyes Closed*	9

Table 1. *Frequency of the AUs coded by manual FACS coders on the CK+ database for the peak frames.*

Figure B.1: Table of Full FACS coding (©Jeffrey Cohn)