GDAŃSK UNIVERSITY OF TECHNOLOGY
FACULTY OF ELECTRONICS, TELECOMMUNICATIONS AND INFORMATICS

Student's name and surname: Adam Sobczuk
ID: 188656
Cycle of studies: bachelor's degree studies
Mode of study: full-time studies
Interfaculty field of study: Data Engineering
realized at: Faculty of Electronics,Telecommunications and Informatics; Faculty of Management and Economics
Profile: Intelligent data processing

Student's name and surname: Oskar Kołoszko
ID: 188941
Cycle of studies: bachelor's degree studies
Mode of study: full-time studies
Interfaculty field of study: Data Engineering
realized at: Faculty of Electronics,Telecommunications and Informatics; Faculty of Management and Economics
Profile: Intelligent data processing

ENGINEERING DIPLOMA PROJECT

Title of the project: Building a facial expression and biosignal-based emotion recognition early fusion model for the EMBOA project dataset

Title of the project (in Polish): Budowa modelu wczesnej fuzji rozpoznawania emocji na podstawie biosygnałów i mimiki twarzy dla zbioru danych z projektu EMBOA

Supervisor: dr inż. Teresa Zawadzka

ABSTRACT

Emotion recognition, a core component of affective computing, enhances human-computer interaction by enabling systems to identify and respond to emotional states. The primary goal of this thesis was to lay grounds for developing an emotion recognition model designed and trained specifically for emotion recognition in children with autism spectrum disorder. This resulted in the development of advanced emotion recognition methods through an early-fusion multimodal framework that integrates facial embeddings and physiological signals, including heart rate (HR), electrodermal activity (EDA), and skin temperature (TEMP). The study leverages data from the EMBOA project, aimed at applying emotion recognition in robot-assisted interventions for children on the autism spectrum.

Two distinct methodologies are proposed: the first classifies dominant emotions using categorical labels, and the second predicts emotional states' intensity across six basic emotions using a continuous distribution. The implemented models are built on bidirectional Long Short-Term Memory (LSTM) networks. The first model achieved the result of 0.0208 in the mean squared error (MSE) metric, while the second achieved a 0.7896 in the F1-Score metric.

Keywords: emotion recognition, affective computing, multimodal early fusion, EMBOA project, biosignals, autism spectrum disorder

Field of science and technology in accordance with OECD requirements: Computer and Information Sciences, Information engineering

STRESZCZENIE

Informatyka afektywna-dziedzina której kluczowym elementem jest rozpoznawanie emocji, zwiększa interakcję na polu człowiek-komputer, umożliwiając systemom identyfikację i reagowanie na stany emocjonalne. Głównym celem niniejszej pracy było stworzenie podstaw dla opracowania modelu rozpoznawania emocji, zaprojektowanego i trenowanego z myślą o rozpoznawaniu emocji u dzieci z zaburzeniami ze spektrum autyzmu. Wynikiem tego było opracowanie zaawansowanych metod rozpoznawania emocji w ramach multimodalnego modelu wczesnej fuzji, integrującego wektory cech twarzy oraz biosygnały, takie jak tętno (HR), aktywność elektrodermalna (EDA) i temperatura skóry (TEMP). Badanie opiera się na danych z projektu EMBOA, którego celem było zastosowanie rozpoznawania emocji w interwencjach wspomaganych robotami dla dzieci ze spektrum autyzmu.

Zaprezentowano dwie odrębne metodologie: pierwsza klasyfikuje dominujące emocje za pomocą etykiet kategorialnych, a druga przewiduje intensywność stanów emocjonalnych w sześciu podstawowych emocjach za pomocą rozkładu ciągłego. Zaimplementowane modele opierają się na dwukierunkowych sieciach typu Long Short-Term Memory (LSTM). Pierwszy model osiągnął wynik 0,0208 w metryce średniego błędu kwadratowego (MSE), podczas gdy drugi uzyskał wynik 0,7896 w metryce F1-Score.

Słowa kluczowe: rozpoznawanie emocji, informatyka afektywna, multimodalna wczesna fuzja, projekt EMBOA, biosygnały, całościowe zaburzenia rozwojowe

Dziedzina nauki i techniki, zgodnie z wymogami OECD: Nauki o komputerach i informatyka, Inżynieria informatyczna

CONTENTS

LIST OF MOST PROMINENT ABBREVIATIONS AND SYMBOLS

EMBOA	–	Affective loop in Socially Assistive Robotics as an intervention tool for children with autism
GUT	–	Gdańsk University of Technology
ITU-YU	–	Istanbul Technical University - Yeditepe University
MAAP	–	Macedonian Association for Applied Psychology
EDA	–	Electrodermal Activity
HR	–	Heart Rate
TEMP	–	Temperature
LSTM	–	Long Short-Term Memory
CNN	–	Convolutional Neural Network
RNN	–	Recurrent Neural Network
PAD	–	Pleasure, arousal, dominance model
FaceReader	–	Emotion recognition software
BORIS	–	Behavioural Observation Research Interactive Software
TP	–	True Positive
TN	–	True Negative
FP	–	False Positive
FN	–	False Negative
MSE	–	Mean Squared Error
MAE	–	Mean Absolute Error
RMSE	–	Root Mean Squared Error

1. INTRODUCTION (ADAM SOBCZUK)

Understanding human emotions is a critical aspect of advancing human-computer interaction. The field of emotion recognition, an integral part of affective computing, seeks to enable machines to identify, interpret and respond to human emotions. This thesis addresses the challenge of improving emotion recognition systems through a multimodal approach, leveraging facial embeddings and physiological signals. By combining visual and biosignal data, this research explores methodologies that enhance the accuracy and applicability of emotion recognition systems in diverse contexts.

The primary goal of this thesis is to develop and evaluate a machine learning model capable of recognizing emotional states with a high degree of precision. The focus lies on integrating visual and physiological modalities to analyze their combined efficacy in emotion recognition. This investigation is underpinned by data collected from the EMBOA project (Affective loop in Socially Assistive Robotics as an intervention tool for children with autism)[1], a European initiative exploring the use of emotion recognition technologies in robot-assisted interventions for children with autism.

The significance of this research lies in its potential applications across various domains, including healthcare, education and human-computer interaction. By improving the ability of systems to understand emotional cues, this work contributes to creating more empathetic and effective technologies.

1.1. SCOPE AND OBJECTIVES

The scope of this research encompasses:

1. Developing a multimodal dataset combining facial embeddings and physiological signals.
2. Designing and implementing deep learning models to classify emotional states.
3. Evaluating the models using diverse metrics to determine their accuracy and robustness.

The objective of the thesis is to design and develop multimodal early-fusion models for emotion recognition from facial expressions and physiological signals (biosignals) in children on the autism spectrum based on the data gathered from the EMBOA project [1].

Thesis tasks for implementation of the objective are as follows:

- To preprocess and align multimodal data for seamless integration.
- To design neural network architectures optimized for multimodal emotion recognition.
- To assess the performance of the first model in terms of accuracy, precision, recall, and F1-Score.
- To assess the performance of the second model in terms of similarity, mean absolute error, mean squared error, and root mean squared error.
- To compare both models using hard labels, a t-test with t-statistic and p-value.
- To compare models to the FaceReader software used in the Geisler et al. [2] thesis.
- To compare models to the state-of-the-art models in literature.

1.2. METHODOLOGICAL FRAMEWORK

This study employs a systematic methodology, beginning with data preprocessing, including synchronization and normalization of facial embeddings and biosignals. Two distinct methods are used to analyze emotion recognition: one focuses on identifying the dominant emotion in a given timeframe, while the other examines the intensity of emotions across multiple categories. The models are trained and evaluated using deep learning techniques, leveraging frameworks like TensorFlow [3] and Scikit-learn [4] of the Python programming language.

1.3. STRUCTURE OF THE THESIS

The thesis is organized as follows:

1.3.1. CHAPTER 2: SCIENTIFIC BACKGROUND

This chapter introduces the theoretical foundations of emotion recognition, focusing on Ekman's discrete emotion model and various modalities of emotion detection, including visual, auditory, textual, and physiological approaches. It also explores contemporary methodologies and neural network architectures used in emotion recognition.

1.3.2. CHAPTER 3: EXPERIMENT DESCRIPTION

This chapter details the experimental framework, including the data sources from the EMBOA project, the tools used for data tagging and processing, and the methodologies employed to align and integrate multimodal data.

1.3.3. CHAPTER 4: DATASET

The dataset chapter provides a comprehensive overview of the data used in this thesis, detailing its sources, preprocessing steps, and the challenges encountered during its preparation. It also includes statistical insights into the dataset's structure and its relevance to the research objectives.

1.3.4. CHAPTER 5: MODEL

This chapter describes the architecture and implementation of the deep learning models designed for this research. It includes an explanation of the preprocessing pipeline, model training processes, and tools utilized.

1.3.5. CHAPTER 6: RESULTS

The results chapter evaluates the performance of the models using various metrics. It presents a detailed analysis of the outcomes, including comparisons betweeen the methods and insights into the models' strengths and limitations.

1.3.6. CHAPTER 7: COMPARISON

This chapter provides a comparative analysis of the models developed in this research, contextualizing their performance against the FaceReader software and existing literature, as well as highlighting their contributions to the field of emotion recognition.

1.4. WORKLOAD DISTRIBUTION

Table 1.1 outlines the tasks undertaken during the thesis, along with their respective contributors, illustrating the collaborative nature of the research. Table 1.2 provides a detailed overview of the chapters and their authorship, highlighting the specific focus of each section within the broader research framework. Both tables serve to contextualize the structure and collaborative effort underpinning this work, showcasing the alignment between the research objectives and the contributions of each chapter.

Table 1.1. Thesis tasks with people realising them

Task	**Person responsible**
Familiarization with the EMBOA project & dataset	Oskar Kołoszko, Adam Sobczuk
Data collection	Adam Sobczuk
Data wrangling	Adam Sobczuk
Data preprocessing	Oskar Kołoszko
Models conceptualization	Adam Sobczuk
Models development	Oskar Kołoszko
Models evaluation and tuning	Oskar Kołoszko
Models results visualization	Adam Sobczuk
Statistical analysis	Adam Sobczuk
Literature comparison	Oskar Kołoszko

Table 1.2. List of chapter with their authors

Chapter	**Person responsible**
Abstract	Oskar Kołoszko, Adam Sobczuk
1. Introduction	Adam Sobczuk
2. Scientific background	Oskar Kołoszko, Adam Sobczuk
2.1. Affective computing	Oskar Kołoszko
2.2. Emotion recognition	Oskar Kołoszko
2.3. Summary	Adam Sobczuk
3. Experiment description	Oskar Kołoszko, Adam Sobczuk
3.1. Information about the project	Oskar Kołoszko
3.2. Tools used	Oskar Kołoszko
3.3. First method	Oskar Kołoszko
3.4. Second method	Oskar Kołoszko
3.5. Summary	Adam Sobczuk
4. Dataset	Adam Sobczuk
5. Model	Oskar Kołoszko, Adam Sobczuk
5.1. Tools used	Oskar Kołoszko
5.2. Data preprocessing	Oskar Kołoszko
5.3. Architecture	Oskar Kołoszko
5.4. Compilation and training	Oskar Kołoszko
5.5. Summary	Adam Sobczuk
6. Results	Oskar Kołoszko, Adam Sobczuk
6.1. Evaluation metrics	Oskar Kołoszko
6.2. Method I	Oskar Kołoszko, Adam Sobczuk
6.3. Method II	Oskar Kołoszko, Adam Sobczuk
6.4. Model Comparisons	Adam Sobczuk
6.5. Summary	Adam Sobczuk
7. Comparison	Oskar Kołoszko, Adam Sobczuk
7.1. Performance of models	Adam Sobczuk
7.2. Literature comparison	Oskar Kołoszko
7.3. Summary	Adam Sobczuk
8. Summary	Adam Sobczuk
Chapters introductions	Adam Sobczuk

2. SCIENTIFIC BACKGROUND (OSKAR KOŁOSZKO, ADAM SOBCZUK)

The field of affective computing represents a multidisciplinary endeavor that brings together insights from engineering, psychology, informatics, and artificial intelligence. This chapter aims to provide the scientific foundation necessary for understanding and contextualizing the work presented in this thesis. By exploring the principles, methodologies, and tools at the intersection of these fields, the chapter introduces key concepts critical to understanding the thesis's focus on emotion recognition.

The chapter begins by delving into **affective computing**, a field dedicated to bridging the gap between humans and machines by enabling computers to identify, interpret, and respond to human emotions. The scope of affective computing is vast, encompassing applications in areas such as human-computer interaction, healthcare, security, and social media. Within this domain, two primary focuses are explored: emotion recognition and sentiment analysis, both of which aim to analyze emotional and evaluative states using various modalities such as text, speech, and physiological signals.

Building on this foundation, the chapter introduces **emotion recognition**—the core focus of this thesis. Starting with a discussion of the fundamental emotion models, it highlights two dominant approaches: Ekman's discrete emotion model, which categorizes emotions into distinct groups like happiness and anger, and dimensional models, which conceptualize emotions as points in a multidimensional space. Emphasis is placed on Ekman's model, which forms the basis of this thesis.

The chapter further explores the **channels for emotion recognition**, examining visual, auditory, and physiological modalities, each with distinct advantages and limitations. It discusses challenges such as facial occlusion in visual recognition, the ambiguity of speech signals in audio analysis, and the complexity of physiological measurements like electroencephalography, a recording of the spontaneous electrical activity in the brain (EEG) [5], and electrocardiography, a recording of the heart's electrical activity (ECG) [6].

Finally, the chapter reviews contemporary approaches to **emotion recognition modality** and **deep neural networks** for processing data and classifying emotions. Unimodal and multimodal strategies are analyzed, along with fusion techniques at the feature and decision levels. In parallel, the role of advanced neural architectures, including CNNs, RNNs, and transformers, is introduced as critical tools for achieving effective emotion recognition.

This foundational discussion sets the stage for the thesis's detailed exploration of emotion recognition methodologies and applications. The concepts outlined in this chapter serve as the building blocks for the technical and experimental contributions of the work.

2.1. AFFECTIVE COMPUTING (OSKAR KOŁOSZKO)

Affective computing is a joining term for elements from numerous fields, such as engineering, psychology, informatics, and artificial intelligence. More precisely, it encapsulates human emotion, sentiment, feelings, emotion recognition, and sentiment analysis. Its main goal is to create a bridge between computers and humans by "giving" computers the possibility of identifying and expressing emotions as well as responding to human ones. The range of applications for building such a cognitive system, which could react to human emotions and communicate its own, is wide. It could be used in various fields, such as human-computer interactions, intelligent vehicle

systems, medical health, the entertainment industry, security monitoring, physiological analysis, and social media [7].

Affective computing consists of two main topics: emotion recognition and sentiment analysis. Emotion recognition emphasizes the detection of the emotional state of human beings. To do so, it focuses on visual emotion recognition, audio/speech emotion recognition, and physiological emotion recognition. Meanwhile, sentiment analysis focuses on text evaluations and viewpoint mining, resulting in positive, negative, or neutral attitudes toward a subject. Because emotions and sentiment are strictly connected in human beings, those two fields can extend over each other [7].

2.2. EMOTION RECOGNITION (OSKAR KOŁOSZKO)

As emotion recognition is a very difficult task, it consists of many components, such as channels and modalities. Moreover, the term emotion is understood differently across the literature, creating the need for choosing a concrete emotion model to give a task. Just as approaches differ, so do the neural networks for them. On the following pages, all those components are discussed.

2.2.1. EMOTION MODELS

This thesis will focus on emotion recognition. A good understanding of emotions is the basis for emotion recognition. The most basic concept of emotion was proposed by Ekman in 1969. Nowadays, there exist two principal emotion models: the discrete emotion model and the dimensional emotion model [8].

The discrete emotion model divides human emotions into autonomous categories. Usually including happiness, fear, sadness, anger, disgust, etc. There exist from two to eight basic emotions, varying by different theories. Ekman suggested seven characteristics to distinguish different basic emotions and emotional phenomena: autonomous evaluation; specific antecedent events; also present in other primates; rapid onset; short duration; unconscious or involuntary appearance; reflected in unique physiological systems such as the nervous system and facial expressions. His emotions consist of anger, surprise, disgust, enjoyment, fear, and sadness. Meanwhile, Plutchik proposed eight emotions and discriminated them according to their intensity, creating an emotion wheel model [8].

The dimensional model notices that emotions are too complex to classify them into single categories. It views them as combinations of vectors in multidimensional space models. The most popular two-dimensional model is the valence-arousal model. In which valence reflects the evaluation of the emotion as positive or negative, while arousal corresponds to the intensity of it. However, as many emotions represent similar values of valence and arousal, there needs to be an additional dimension to distinguish them. The most popular three-dimensional model, proposed by Mehrabian and Russell, is the pleasure, arousal, and dominance (PAD) model [8].

This thesis will focus solely on Ekman's model.

2.2.2. EMOTION RECOGNITION CHANNELS

Emotion recognition focuses on visual emotion recognition, audio/speech emotion recognition, and physiological emotion recognition.

Emotion recognition based on visual sensors is the easiest and therefore the most common method. It is characterized by low cost and simplicity. However, it comes with many disadvantages. For effective emotion recognition, a clear view of the face is needed. Therefore, poor light intensity, face occlusion, or bad visual sensor angles can significantly influence the quality of

the collected data. Moreover, humans are good at showing misleading external emotions. Many social interactions require a polite smile, but that does not necessarily mean a good mood. Moreover, different skin colors, looks, and facial features can also pose difficulties to the correctness of classification. Furthermore, one emotion can be expressed by one individual in many ways. On the contrary, small changes in facial features could mean different emotions [8].

Audio/speech recognition—speech is one of the most important components of our daily communication; it is also very important for emotion recognition. The biggest difficulty of speech recognition is that the same sentence can convey different emotions. Different speaking styles of different people bring different information as well [8].

The last source of data for emotion recognition is physiological channels, which we can divide into many distinct ones, with the basic ones being EEG, ECG, EMG, GSR/EDA, BVP, EOG, ET, RES, ST, and HRV.

- Electroencephalogram (EEG) – measures the electrical signal activity of the brain through set-up electrodes on the skin surface of the head. Studies have shown that many parts of the brain, such as the prefrontal cortex, temporal lobe, and anterior cingulate gyrus, are related to the control of emotions [8],
- Electrocardiogram (ECG) - a method of electrical monitoring on the surface of the skin that detects the heartbeat controlled by the body's electrical signals. Heart rate and heart rate variability are controlled by the sympathetic and parasympathetic nervous systems [8],
- Electromyogram (EMG) - measures the degree of muscle activation by collecting differences in voltage generated during muscle contraction [8],
- Galvanic skin response/Electrodermal activity (GSR/EDA) - a method of monitoring the perspiration of the skin, thus its electrical conductivity. When a person is in an anxious or tense mood, the sweat glands usually secrete more sweat, which causes a change in current [8],
- Blood volume pulse (BVP) – monitors the pulse wave of the heart and the volume of the blood flowing through a vessel [9],
- Electrooculography (EOG) - measures the vertical and horizontal movement of the eyes.[9],
- Eye tracking (ET) – tracks the position and movement of the eyes [10],
- Respiratory rate (RES) – monitors the tempo and regularity of breaths [9],
- Skin temperature (ST) – monitors the temperature of the skin, exposing if the person is relaxed or not [9],
- Heart Rate Variability (HRV) – detects the shifts in frequency bands of heart rate [9].

EMOTION RECOGNITION MODALITY

Emotion recognition approaches different sources in two main ways: unimodal and multimodal. Unimodal emotion recognition consists of one channel, such as visual, audio, or physiological. This one channel is used solely to recognize emotion. Meanwhile, multimodal emotion recognition uses two or more channels together to better analyze emotions. The unimodal emotion recognition can also serve as a source of features for fusion to multimodal emotion recognition [9].

There are two main types of feature fusion: early (feature-level) and late (decision-level). Early fusion is characterized by joining together the features taken out of each mode. This often results

in a feature vector of bigger dimensions, which can be reduced using the dimension reduction methods. Only then are the features passed to a classifier. It profits from relationships between various modalities while having a drawback in the difficulty of time synchronization between different modalities. On the other hand, in the late fusion, all the modalities are treated independently, and their features are passed to independent classifiers. Then the results of different classifiers are algorithmically combined into one result. It is characterized by emphasizing the difference between features of different modalities and choosing an optimal classifier for each modality but loses on learning being a very time-consuming process [10].

2.2.3. DEEP NEURAL NETWORKS FOR EMOTION RECOGNITION

In the present emotion recognition tasks, the feature vector is passed to a deep neural network.

- AutoEncoder (AE) - uses back-propagation and unsupervised learning. Its architecture is built on the input layer, secret encoding layer, and decoding layer. The network initially sets up the objective result to match the input. Then it makes an effort to recreate the inputs. Then it drives the hidden layer to discover the optimum input representations [10],

- Convolutional Neural Network (CNN) – consists of the input layer, early layers for identifying features, such as edges, and late layers for recombining features using the high-level input characteristics, followed by the classification layer [10],

- Restricted Boltzmann Machine (RBM) – is made up of a fundamental sensory input layer, the hidden layer that describes data in an abstract manner, and the output layer that categorizes the network. RBM is a generative stochastic neural network that can learn a distribution of probability across its inputs [10],

- Recurrent Neural Network (RNN) – a neural network with additional connections that may deliver feedback to earlier levels, which is a key RNN feature. It is based on taking input from previous memory and modeling the problems using time series and sequences [10],

- Long Short-Term Memory (LSTM)—a type of RNN that can be trained to remember the main state of the model, which, due to the problem known as vanishing gradients, is not entirely possible in an RNN [10],

- Transformers - a neural architecture that employs an attention mechanism to encode input data into highly effective features. It tries to predict other features in the sequence [10].

2.3. SUMMARY (ADAM SOBCZUK)

This chapter provided a comprehensive overview of the scientific background necessary for understanding emotion recognition within the field of affective computing. Affective computing, a multidisciplinary field, seeks to bridge human-computer interaction by enabling machines to recognize, interpret, and respond to human emotions. Two fundamental areas, emotion recognition and sentiment analysis, were introduced as the cornerstones of this domain.

The chapter highlighted the essential theoretical foundations of emotion recognition, focusing on Ekman's discrete emotion model, which classifies emotions into distinct categories such as happiness, fear, and sadness. This model, which forms the basis of this thesis, was contrasted with dimensional emotion models like the valence-arousal framework and the PAD model, illustrating their relevance in understanding the complexity of emotional states.

A detailed exploration of the *emotion recognition channels* emphasized the various modalities used for emotion detection, including visual, auditory, and physiological signals. While visual sensors offer simplicity and low cost, their effectiveness is often hindered by external factors such as lighting. Audio signals add depth to emotion detection but face challenges such as speaker variability. Physiological modalities, encompassing EEG, ECG, and other biometric signals, provide robust emotion-related data but require more invasive or specialized setups.

The chapter also discussed *unimodal* and *multimodal emotion recognition*, emphasizing the trade-offs between simplicity and accuracy. Fusion techniques, both at the feature and decision levels, were explored as strategies to improve emotion recognition performance. Additionally, advanced deep neural networks, such as CNNs, RNNs, LSTMs, and transformers, were introduced as powerful tools for feature extraction and classification, enabling machines to process complex, multidimensional emotional data efficiently.

In conclusion, the concepts and methodologies presented in this chapter lay the groundwork for the technical contributions of this thesis. By integrating insights from diverse modalities and leveraging cutting-edge neural architectures, this work aims to advance the field of emotion recognition. The following chapters will build upon this foundation, presenting experimental methodologies, results, and a detailed analysis of the proposed approaches.

3. EXPERIMENT DESCRIPTION (OSKAR KOŁOSZKO, ADAM SOBCZUK)

This chapter provides a detailed description of the experimental setup and methodologies utilized in the research conducted by Geisler et al. [2]. The data used in this study was sourced from the *EMBOA* project, an international initiative focused on exploring emotion recognition technologies in robot-assisted interventions for children with autism.

The chapter begins by outlining the data and its structure, followed by a discussion of the tools employed for emotion analysis, namely *FaceReader* and *BORIS*. It then presents the methodologies implemented to process and compare data from automated and manual tagging approaches, including the key steps in data preparation and the metrics used to evaluate accuracy and similarity.

By describing these elements in detail, this chapter establishes the foundation for the analysis and interpretation of the results with the implemented model in the subsequent chapters.

3.1. INFORMATION ABOUT THE PROJECT (OSKAR KOŁOSZKO)

This thesis uses data provided by the project of Geisler et al. [2]. In their thesis, they used the data from the EMBOA project. EMBOA is an acronym for the project "Affective Loop in Socially Assistive Robotics as an intervention tool for Children with Autism". The project is a mixed research and didactic project under the EU Erasmus Plus Strategic Partnership for Higher Education Programme conducted in the years 2019-2022. The project was conducted by an international consortium and aimed at the development of guidelines and practical evaluation of applying emotion recognition technologies in robot-supported intervention for children on the autism spectrum [1].

They realized their thesis using the data shared by the partners of the EMBOA project. Precisely, for the analysis, they used data from four scientific centers:

- Gdansk University of Technology
- Technical University of Stambul
- Yeditepe University
- The Macedonian Association for Applied Psychology

The data provided by them was grouped into four categories:

- Camera (the video materials)
- Eye tracker
- Microphone (audio materials)
- Wristband (measurement of bio-signals from the wristband put on the children's wrist)

Each of the folders with data was structured by them in this manner:

1. Research Centre
2. Session number (written in the Sśession_number' convention),
3. Participant ID (written in the Cśession_number' convention).

As a part of their project, they also provided files with information about the participants and files with information about the scenarios of the sessions. However, those files will not be used in this thesis. Moreover, as this thesis focuses on emotion recognition based on facial expressions and bio-signals, the data obtained in eye tracker and microphone categories will not be used.

From the videos received, they deleted one file from the Technical University of Stambul, Yeditepe University, due to not being able to open the video. They also found out that every video provided by the Gdansk University of Technology Research Centre (GUT) had a duplicate with a shift of at most 2 minutes. This information was taken into account during the preprocessing of data for this thesis. The final durations of videos can be seen in Table 3.1.

Table 3.1. Girls thesis file statistics

Scientific Center	Number of films	Films duration [HH:MM:SS]
GUT	8	1:28:03
ITU-YU	13	3:18:48
MAAP	78	12:05:10
Sum	99	16:50:0

3.2. TOOLS USED (OSKAR KOŁOSZKO)

To obtain the data necessary for later analysis, Geisler et al. [2] used two softwares. The first one, FaceReader, allowed for automatic emotion recognition. The second one, BORIS, provided the possibility of manual tagging.

3.2.1. FACEREADER

Geisler et al. [2] used the FaceReader software published by Noldus Information Technology. It analyzes the human mimics and automatically labels the identified emotion based on its classifier trained on 10,000 pictures labeled by experts. The model was used by Geisler et al. [2] to analyze the videos and extract emotions from them in timestamps of one second. This meant extracting emotions from 60,604 timestamps. They have obtained two files for every video. The first one got information about the change in the participant's emotional state every second. The second one had the intensity of all emotions in every second of the video. Those two approaches will later become two different methodologies, in which the first method will focus on the leading emotion for every timestamp, while the second method will focus on the diversity of emotions for every timestamp. Due to problems that we will address later, the FaceReader software used by Geisler et al. [2] managed to identify an emotion in 20% of the provided data. This corresponded to around 11,000 tagged seconds.

3.2.2. BORIS

The second software used by Geisler et al. [2] to tag the data was an open-source program called BORIS (Behavioural Observation Research Interactive Software). BORIS is used by researchers to observe and tag events and occurrences happening in videos. Geisler et al. [2] used the BORIS program to label each video. To improve accuracy, each of them performed the labeling independently. The final product for each video was therefore three files composed of 8 columns corresponding to emotions and rows corresponding to subsequent timestamps of one second.

After obtaining the data using FaceReader and labeling the data using BORIS, the student began the analysis of the exactness of obtained emotions. Due to lacking completeness of information about the participants, the only information used was the gender and age of participants.

3.3. FIRST METHOD (OSKAR KOŁOSZKO)

Because every file was tagged three times, preliminary to the analysis, the data needed to be unified. This was done by the Geisler et al. [2] by summing all three values for each timestamp and extracting the emotion with the highest value. In this case, when in every tag a different emotion was identified, that timestamp was filled with *Unknown* values. In case the maximum value for every emotion was zero, that timestamp was filled with a *None* value. All the discrepancies in the lengths of the files were adjusted by adding rows filled with zeros.

In the next step, Geisler et al. [2] compared every timestamp of the tagged BORIS files with the corresponding files created by FaceReader. They created a new file, which stored the information about those concurrencies. For every timestamp, they were putting a value of one if the emotion from the BORIS file corresponded to the emotion from the FaceReader file and a zero value if not.

Afterwards, the values from every result table were added and divided by the length of the table. This way the accuracy value for every movie was obtained. Then they presented the accuracy based on the scientific center, the age of the participant, the gender of the participant, and the scenario. The global accuracy totaled to 53,69%. However, it is important to note that almost 27% of videos had less than 10% accuracy.

Subsequently, Geisler et al. [2] used the confusion matrix. Besides the *Happy* emotion, which had any true positive, false positive, and false negative values, all the other emotions consisted mainly of true negatives. Based on the accuracy values for individual emotions and the values of the confusion matrix, Geisler et al. [2] concluded that as the model's prediction accuracy for a given emotion increases, its detectability consistency also increases. Geisler et al. [2] also noted the model's tendency to overdetect the anger emotion. While not detecting the sadness, fear, disgust, or surprise emotion almost at all.

3.4. SECOND METHOD (OSKAR KOŁOSZKO)

While in the first method Geisler et al. [2] were centered around the leading emotion, the second method is focused on evaluating the similarity of the data acquired by them using the FaceReader in comparison to the manually tagged data using the BORIS. To evaluate this similarity, they used the absolute error, and to demonstrate the level of similarity, the value of the absolute error was additionally subtracted from one. Finally, the similarity value was computed using the formula:

$$1 - |x_1 - x_2| \tag{3.1}$$

Where:

x_1 – the value of emotion saturation manually tagged using the BORIS tool,
x_2 – the value of emotion saturation obtained automatically using the FaceReader tool.

As every file was tagged three times, before the analysis the data needed to be unified. Every BORIS-tagged file consisted of seven columns corresponding to emotions (Happy, Sad, Angry, Surprised, Scared, Disgusted, and Contempt). The result from the FaceReader tool consisted only of Ekman's six basic emotions: *Happy, Sad, Angry, Surprised, Scared, Disgusted, Contempt* and *Neutrality*, which represented a lack of any shown emotions. Due to no occurrence of the

Contempt emotion in the tables generated using the FaceReader tool, all the information about this emotion was deleted by Geisler et al. [2] from the manually tagged files. As the FaceReader files had the Neutrality column, to keep the consistency, the Neutrality column was added to the BORIS files. All the discrepancies in the lengths of the files were adjusted by adding rows filled with zeros.

To obtain the result tables, firstly, data collected in three rounds of tagging were combined. It was done by Geisler et al. [2] by first dividing the values of each row by three and then adding those values. If the row had not any values other than zeros, the value of the neutrality column was set to one. In the end, for every emotion in every timestamp, the similarity was computed using the formula 6.8. Then Geisler et al. [2] presented the similarity metrics for every scientific center, the age of the participant, the gender of the participant, and the scenario. The global similarity totaled 85.18%.

3.5. SUMMARY (ADAM SOBCZUK)

This chapter has detailed the experimental work conducted by Geisler et al. [2], providing insight into the data sources, tools, and methodologies that form the basis of this thesis. By leveraging data from the *EMBOA* project and employing tools such as *FaceReader* and *BORIS*, their work showcased the complexities of emotion recognition using both automated and manual approaches.

The comparison of methodologies highlighted the strengths and limitations of the tools used, particularly the challenges faced by automated systems in achieving accurate and consistent emotion detection. Despite these difficulties, the high similarity metrics between manual and automated tagging underscored the potential of combining these approaches to improve emotion recognition accuracy.

The findings of Geisler et al. [2] offer valuable lessons for this thesis, particularly regarding the importance of preprocessing, the careful handling of inconsistencies in data, and the selection of appropriate metrics for evaluation. These insights set the stage for the detailed analysis and advancements presented in the subsequent chapters.

4. **DATASET** (ADAM SOBCZUK)

This chapter provides a comprehensive description of the dataset used for training the emotion recognition model. The dataset originates from the *EMBOA* project, was processed and labeled by Geisler et al. [2], and then further processed for this thesis.

The chapter begins by outlining the structure and sources of the dataset, detailing the specific directories and modalities utilized in this thesis. Subsequently, it describes the preprocessing steps applied to both video data and biosignals, including the extraction of facial embeddings and the selection of relevant physiological signals. Techniques for linking visual and physiological data are also discussed, ensuring the synchronization of modalities for seamless analysis.

The chapter also includes an overview of the label generation process and the alignment of input and label vectors to create a unified dataset. Finally, it presents statistical insights into the dataset's structure and discusses the challenges encountered during preprocessing and data annotation. This detailed exploration forms the foundation for the subsequent analysis and model development.

4.1. DATASET OVERVIEW AND STRUCTURE

To understand the process of data wrangling described in later sections, first it is important to learn about the original dataset and how the contributions from it affect this thesis preprocessing tasks.

4.1.1. SOURCE OF THE DATASET

The data was collected by members of the **EMBOA** project. The Geisler et al. [2] processed and labeled the data for classification. This data was organized into six folders:

- Additional data (info about participants and scenarios)
- Camera (videos and files with emotions recognized by the FaceReader software and files with emotions labeled by Geisler et al. [2] using **BORIS** software)
- EyeTracker (videos and data about eye tracking)
- Microphone (audio recordings)
- SessionCards (session cards provided by research centers)
- Wristband (measurements from the wristband of the participant)

Within the thesis, only the *Camera* and *Wristband* directories were used.

Due to a lack of compliance with the organization of two session directories from the **MAAP** directory, they were not taken into the final dataset. The biosignals chosen for the dataset were Heart Rate (**HR**), Temperature (**TEMP**), and Electrodermal Activity (**EDA**).

4.1.2. PREPROCESSING OF THE DATASET

The dataset used for this project was derived from the **EMBOA** dataset, which includes video recordings and bio-signals for emotion recognition tasks. The extraction process involved multiple steps.

FACE EMBEDDINGS EXTRACTION

Facial feature vectors were extracted from the **EMBOA** videos using the **MTCNN** face detector and the InceptionResNetV1 model from the facenet_pytorch library[11]. FaceNet is a deep learning model for face recognition that was introduced in a paper titled "FaceNet: A Unified Embedding for Face Recognition and Clustering" by Schroff et al. [12].

facenet_pytorch is a Python library that provides a PyTorch implementation of the FaceNet model, making it easy to use FaceNet for face recognition tasks in PyTorch-based projects.

MTCNN is a prominent model designed for joint face detection and alignment [13]. InceptionResNetV1 is a convolutional neural architecture that builds on the Inception family of architectures but incorporates residual connections [14].

facenet_pytorch library provides **MTCNN** and InceptionResNetV1 models to detect faces on the video and extract facial features that are saved as vectors. These vectors serve as face embeddings, representing facial features in a high-dimensional space. If no face was detected in a given frame, the corresponding embedding values were filled with zeros.

There were several problems encountered with particular timestamps; two of them were the most prominent. In some of the recordings, children had masks on, which obstructed the view of the face, thus making it harder for the model to detect one. Another case was when multiple people were present and faced the front of the camera. In those cases, more than one face was detected in the same frame. This resulted in the problem of choosing the correct embedding—the one that corresponded to the child's face. The first approach was to use a fixed rectangle when detecting a face, knowing that the child's face is smaller on the recordings and due to the static nature of the recordings (same distance from the camera between sessions). However, this wasn't sufficient; a slight improvement was encountered, but still, there were some frames where two facial embeddings were extracted. To rectify this, another solution was applied, which was to choose, from the detected faces, the smallest rectangle. This, with the combination of the previous step, resulted in the correct use of the correct facial embedding[11].

BIO-SIGNAL CHOICE

Information about each signal was located in the *info.txt* document in the signal directory. Descriptions of each signal file are listed below. The following biosignals were collected from the wristband during each session:

- **ACC** - Data from a 3-axis accelerometer sensor. The accelerometer is configured to measure acceleration in the range [-2g, 2g],
- **BVP** - Data from photoplethysmograph,
- **EDA** - Data from the electrodermal activity sensor expressed as micro siemens (μS),
- **HR** - Average heart rate extracted from the BVP signal,
- **IBI** - Time between an individual's heartbeats extracted from the BVP signal,
- **TEMP** - Data from the temperature sensor expressed in degrees on the Celsius (℃) scale.

Out of the six biosignals collected, three were chosen: heart rate **HR**, temperature **TEMP**, and electrodermal activity **EDA**.

Sweat gland activity causes changes in the electrical characteristics of the skin, which are referred to as **EDA** signals. Sweat is linked to the sympathetic branch of the autonomic nervous system and is produced when eccrine sweat glands are activated. Furthermore, the sympathetic

nervous system is the only one that can regulate this distinct autonomic signature. Both happy and negative emotions have different **EDA** signal frequency and amplitude characteristics [15]. **EDA** biosignal is a viable choice widely used in emotion recognition models [16].

Physiological signals respond differently to various emotions. Mood swings cause changes in heart rate; specifically, fear and anger are correlated with increased heart rate. Disgust results in a decrease in heart rate; similarly, pleasantry induces a lower heart rate to a state of neutrality. The difference between the effects of fear and relaxation is of significance; the average heart rate whilst experiencing happiness is lower than that of a sad state of emotion [17]. **HR** was chosen as a second biosignal for this experiment.

Skin temperature (**TEMP**) stands out as a complementary feature to heart rate (**HR**) and electrodermal activity (**EDA**), enhancing the depth and accuracy of emotion recognition systems. **HR** and **EDA** are well-established markers in emotion recognition, reflecting autonomic nervous system (**ANS**) activity [18]. However, **TEMP** adds a unique dimension by directly linking to peripheral vascular responses governed by the sympathetic nervous system (**ANS**):

- **HR** reflects changes in cardiac activity due to emotional arousal,
- **EDA** measures sweat gland activity, providing insights into arousal levels,
- **TEMP**, on the other hand, captures vasoconstriction or vasodilation effects, which occur as part of the body's thermoregulatory response to emotional stimuli.

Together, these features offer a holistic view of physiological responses to emotions, where **TEMP** fills the gap in monitoring peripheral vascular activity [18].

TEMP complements **HR** and **EDA** by providing information on thermal regulation, adding depth to the physiological profile [18]. For instance, the standard deviation of skin temperature has been shown to significantly contribute to emotion classification, as it reflects variability tied to stimuli-triggered sympathetic nerve system (**ANS**) modulation [18].

It bridges gaps when **EDA** or **HR** alone lacks specificity, especially in low-arousal states [18]. While **HR** and **EDA** often respond with rapid fluctuations, **TEMP** exhibits slower trends that stabilize in controlled environments. This temporal behavior provides a complementary layer of information that enhances robustness in emotion detection systems [18].

While **HR** or **EDA** may be susceptible to noise or individual variability, **TEMP** provides an additional, independent marker of emotional states [18]. **HR** and **EDA** primarily reflect cardiac and sweat gland activity, whereas **TEMP** contributes insights into vascular responses, achieving a more comprehensive view of **ANS** function [18].

Integrating **TEMP** with **EDA** and **HR** emotion recognition systems gains a more nuanced understanding of physiological responses, proving to be more effective in affective computing [18].

BIO-SIGNAL NORMALIZATION

Three bio-signals: Heart Rate (**HR**), Temperature (**TEMP**), and Electrodermal Activity (**EDA**) were processed to ensure the same measurement frequency. Heart Rate was used as the baseline frequency at one measurement per second. For Temperature and **EDA**, measurements were downsampled using the mean of values in overlapping time windows to match the 1 Hz frequency. This ensured consistent alignment between all inputs while minimizing temporal discrepancies, though this process may have introduced minor data loss.

4.1.3. INPUT VECTOR

The multimodality aspect of this thesis models, described in chapter 5, requires a concatenated input vector. The process of extracting the required modalities from the source dataset and the concatenation to one input vector is described below.

LINKING FACIAL EMBEDDINGS WITH BIOSIGNALS

The integration of facial embeddings with biosignal data forms the cornerstone of this multimodal analysis[11], enabling the study of emotional states by combining visual and physiological cues. This process can be broadly divided into several stages, as outlined below.

FACIAL EMBEDDING EXTRACTION

Facial embeddings are derived from video data to capture facial features that are indicative of emotional states. To achieve this:

- A pretrained facial recognition model, such as FaceNet, is employed to generate embeddings. The Multi-task Cascaded Convolutional Networks (**MTCNN**) are utilized to detect and localize faces in video frames.
- Videos are processed frame-by-frame, with embeddings computed for frames at one-second intervals. This ensures temporal consistency and reduces computational load.
- For each frame, the smallest detected facial bounding box is cropped and processed to extract embeddings. In cases where no face is detected, a vector of zeros is used to maintain uniformity in the dataset.
- These embeddings are stored as high-dimensional vectors, representing facial features over time.

BIOSIGNAL DATA ACQUISITION AND PROCESSING

The biosignal data comprises measurements of electrodermal activity (**EDA**), temperature (**TEMP**), and heart rate (**HR**), which are indicative of physiological responses to emotional stimuli. The following steps are undertaken[11]:

- Raw data is loaded from CSV files. Each signal is aligned to its respective timestamp and sampling frequency.
- To ensure compatibility with video data, the signals are processed to unify their sampling frequencies. This involves trimming the data to ensure its length is a multiple of the frequency and calculating mean values over specific time intervals.
- Metadata, such as the recording start time and duration, is extracted to synchronize the biosignals with the video data.

SYNCHRONIZATION OF MODALITIES

The alignment of biosignals and facial embeddings is critical for linking physiological responses with visual cues. The synchronization process involves:

- Calculating the start and end times for each modality, including video and biosignals, based on their respective timestamps.

- Determining the overlapping time window across all modalities and slicing each dataset to include only the relevant segments.

- This ensures that for each second of the video, corresponding biosignal data is available for analysis.

DATA INTEGRATION

Techniques like Canonical Correlation Analysis (**CCA**) and Multiple Kernel Learning (**MKL**) are commonly employed to enhance the effectiveness of this fusion by addressing the inherent differences in scale and distribution across modalities. In its simplest form, feature concatenation is performed by aligning and merging raw or preprocessed features from each modality, facilitating seamless integration before classification. This method is particularly effective when the modalities are temporally synchronized, as it ensures that each vector encapsulates the information from all modalities for a specific time instance [19]. Due to the synchronization of modalities, this paper explores aligning and merging with a simple concatenation of all features into one vector (**MFC**).

Once the modalities are synchronized, the data is integrated into a unified dataset. Facial embeddings are combined with biosignal measurements to form a multimodal feature set. This dataset includes columns for **EDA**, **TEMP**, **HR**, and the facial embedding features. The integrated dataset is then exported for further analysis.

4.1.4. LABEL VECTOR COMPILATION

Label vectors were constructed from the **BORIS** annotation files. As there were three annotation rounds, each video had three distinct **BORIS** annotation files. Before constructing the label vectors, the three files for every video were filled with rows with only zeros to align them. The process of creating a label vector involved two distinct approaches:

- Method I: Three values of emotions for each row of the **BORIS** annotation file were added. The emotion with the higher value was evaluated to identify the dominant emotion. If no emotion was labeled, the row was marked as "None." If three distinct emotions were labeled, the row was marked as "Unknown."

- Method II: The value of each emotion for each second was normalized by dividing the raw emotion value by the number of labelers and then summed up, producing a percentage distribution of intensification across six basic emotions.

Due to content emotion not being recognized by the FaceReader software in Geisler et al. [2] thesis and the subsequent removal of information about this emotion by Geisler et al. [2] themselves in this thesis, the content emotion is also not taken into account.

4.1.5. DATA ALIGNMENT AND FINAL PROCESSING

To ensure consistent input vector and label vector lengths, all vectors were aligned using video metadata[11]. This consisted of obtaining the starting timestamp of all vectors. It was provided in the data for bio-signals and obtained using external Python libraries for face embeddings and labels based on video timestamps. Then the ending timestamp was calculated based on the length of the vector or video and its starting timestamp. The last step consisted of trimming vectors to the latest starting time and earliest ending time across all modalities.

4.1.6. LABEL DETAILS

The labels represent six basic emotions, following categorical encoding as follows:

- *Happy*: [1, 0, 0, 0, 0, 0]
- *Sad*: [0, 1, 0, 0, 0, 0]
- *Scared*: [0, 0, 1, 0, 0, 0]
- *Disgusted*: [0, 0, 0, 1, 0, 0]
- *Surprised*: [0, 0, 0, 0, 1, 0]
- *Angry*: [0, 0, 0, 0, 0, 1]

Following Method I practice, vectors with three different labels are marked as *Unknown*, and vectors with no values are marked as *None*.

4.1.7. FINAL DATASET

The final dataset includes:

- Feature Vectors: A combination of face embeddings and bio-signals, resulting in a total of 515 features per sample (3 bio-signals and 512 face embeddings).
- Label Files: Two separate labeling schemes were saved:
  - **BORIS**_method_I.csv: Contains dominant emotion labels or "None"/"Unknown."
  - **BORIS**_method_II.csv: Contains percentage distributions of emotions.

4.2. DATASET STATISTICS

This section encompasses descriptive statistics of the preprocessed dataset. Figures included in this section contain information about the distribution of files, the distribution of inputs, and labels.

4.2.1. DIRECTORY STRUCTURE

The dataset comprises three main directories representing research centers: **GUT**, **ITU-YU**, and **MAAP**. Each directory contains .csv files categorized as:

- Input files: contain facial embeddings and bio-signal data.
- BORIS_method_I.csv: Contains dominant emotion labels or "None"/"Unknown."
- BORIS_method_II.csv: Contains percentage distributions of emotions.

Figure 4.1 shows the organizational structure of the input (X) files, label (Y) files, sequential datasets, and the trained models themselves.

```
de-earlyfusionthesis
├── Datasets
│   ├── GUT_test_method_I
│   ├── GUT_test_method_II
│   ├── GUT_test_method_II_balanced
│   ├── GUT_test_method_I_balanced
│   ├── GUT_train_method_I
│   ├── GUT_train_method_II
│   ├── GUT_train_method_II_balanced
│   ├── GUT_train_method_I_balanced
│   ├── ITU_YU_test_method_I
│   ├── ITU_YU_test_method_II
│   ├── ITU_YU_test_method_II_balanced
│   ├── ITU_YU_test_method_I_balanced
│   ├── ITU_YU_train_method_I
│   ├── ITU_YU_train_method_II
│   ├── ITU_YU_train_method_II_balanced
│   ├── ITU_YU_train_method_I_balanced
│   ├── MAAP_test_method_I
│   ├── MAAP_test_method_II
│   ├── MAAP_test_method_II_balanced
│   ├── MAAP_test_method_I_balanced
│   ├── MAAP_train_method_I
│   ├── MAAP_train_method_II
│   ├── MAAP_train_method_II_balanced
│   ├── MAAP_train_method_I_balanced
│   ├── train_dataset_method_I
│   ├── train_dataset_method_II
│   ├── train_dataset_method_II_balanced
│   └── train_dataset_method_I_balanced
├── GUT
├── ITU-YU
├── MAAP
└── Models
    ├── model_method_I
    └── model_method_II
```

Figure 4.1. Directory structure of the final preprocessed data.

4.2.2. RECORDING COUNTS

Figure 4.2 shows the distribution of types of files across research centers.

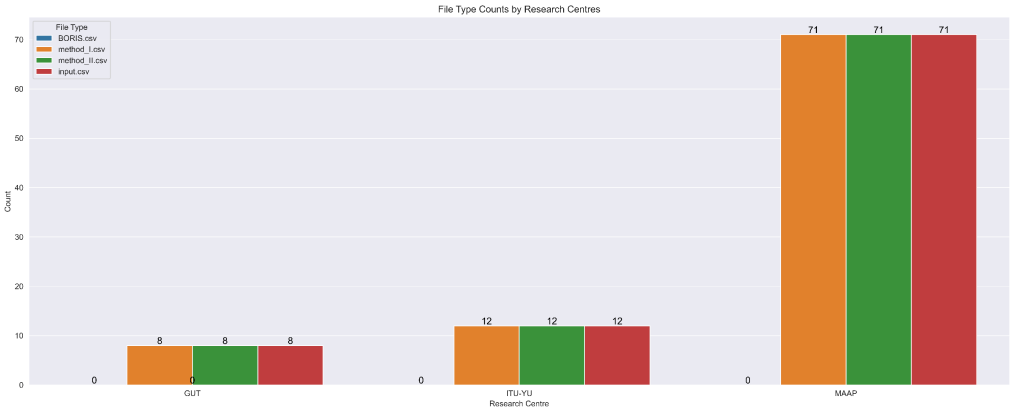

Figure 4.2. Distribution of types of files created across research centers

Most recordings (71 out of 91) were from the Macedonian Association for Applied Psychology. Recordings from Gdańsk University of Technology, Technical University of Istanbul, and Yeditepe University account for only 20 video entries in the dataset.

4.2.3. SESSION AND CAMERA COUNTS

Each directory contains subdirectories organized by session, with varying numbers of camera recordings per session. The graph below illustrates a breakdown of videos per session by each research center. There are up to 11 sessions from **MAAP**, three sessions from **GUT**, and two sessions from **ITU-YU**.

Figure 4.3 shows how many recordings were taken for each session at Gdańsk University of Science and Technology.

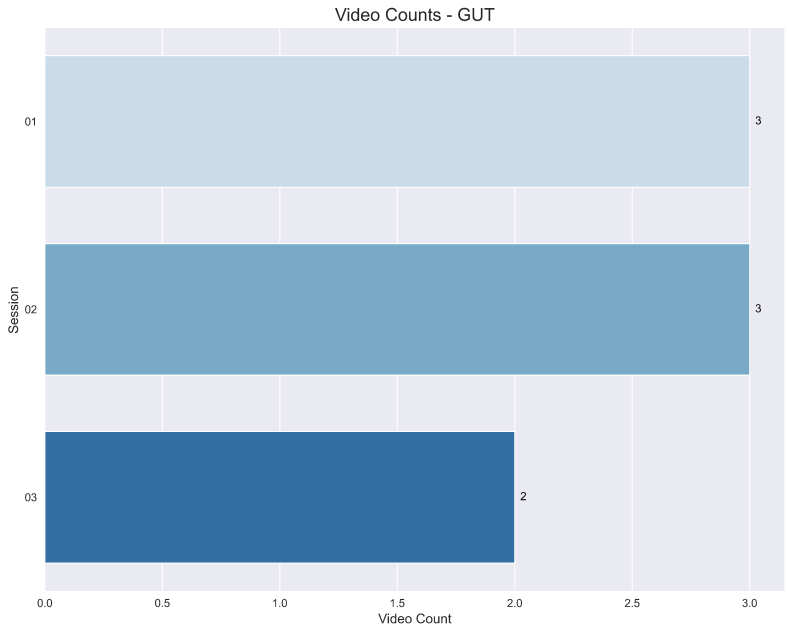

Figure 4.3. Camera recordings per session in GUT

Figure 4.4 shows how many recordings were taken for each session at the Technical University of Stambul and Yeditepe University.

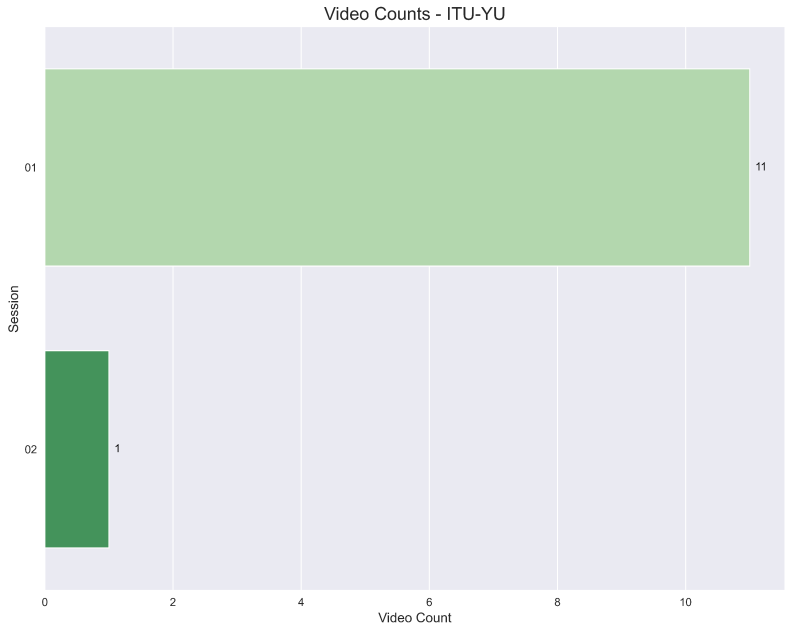

Figure 4.4. Camera recordings per session in ITU-YU

Figure 4.5 shows how many recordings were taken for each session at the Macedonian Association for Applied Psychology.

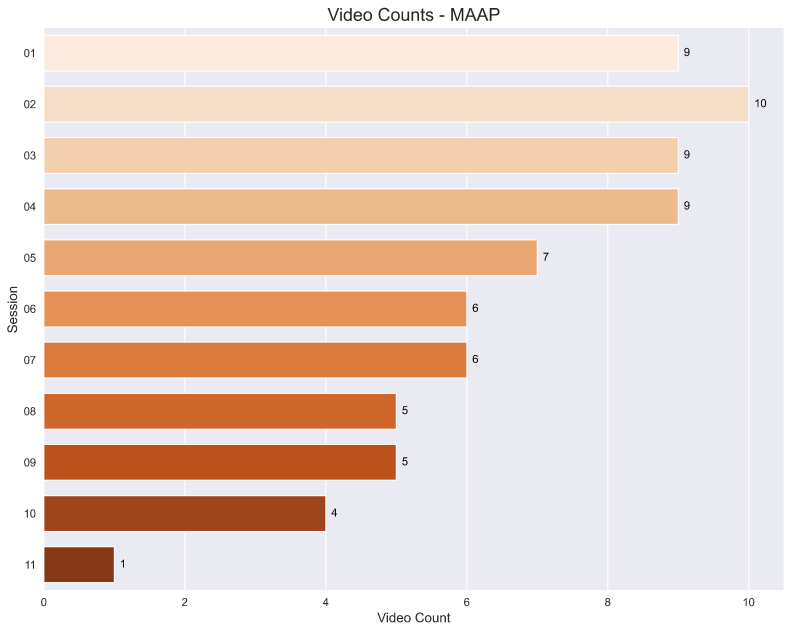

Figure 4.5. Camera recordings per session in MAAP

4.2.4. DATASET STATISTICS

Figure 4.6 shows the distribution of labels labeled as 'None' and 'Happy', with the remaining emotions shown together under the 'Other' label.

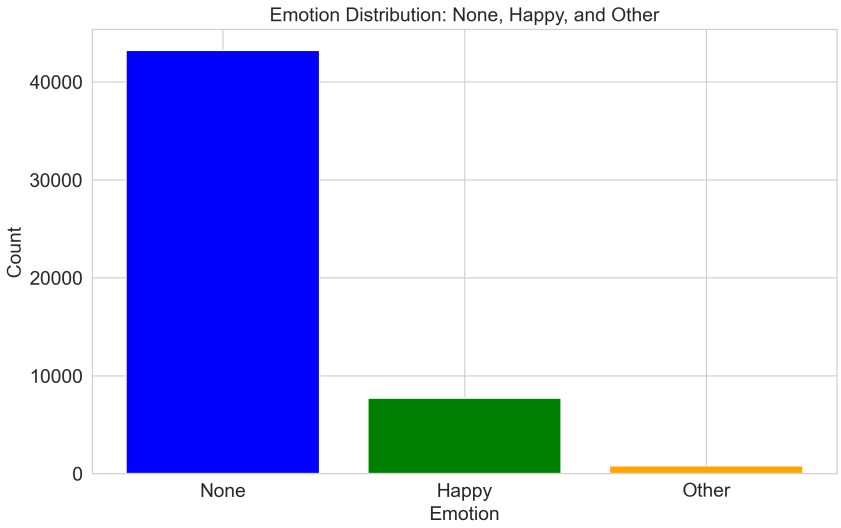

Figure 4.6. Distribution of labels None, Happy, and Others in the Method I labels

Most of the labels, an astounding amount of over 40,000, are not applicable for emotion recognition. However, they serve as valuable insights into seconds preceding and after an emotion appears. This is crucial for the model proposed in the later chapter. However, there is also a discrepancy that needs to be resolved, as Happiness is the prevalent emotion this dataset needs to be normalized, which is discussed further in later sections of the thesis.

Figure 4.7 shows the distribution of other emotions.

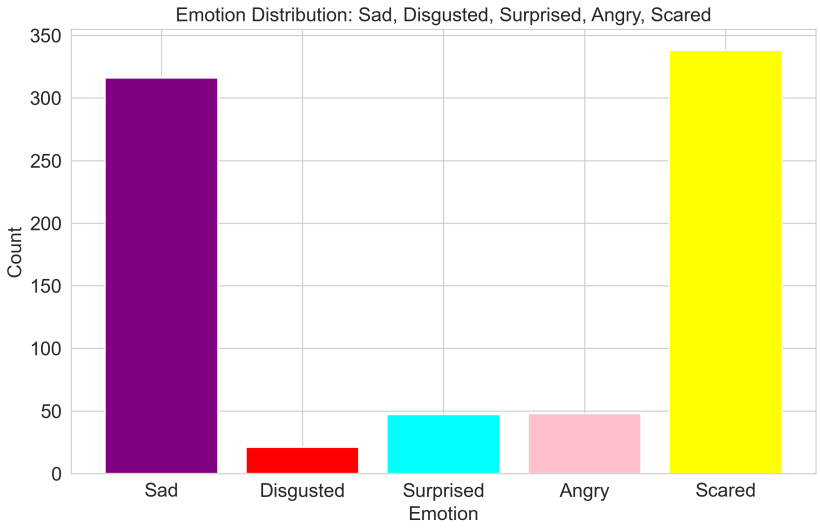

Figure 4.7. Distribution of other emotion labels in the Method I labels

Out of the remaining five emotions, sadness and fear take second place in terms of appearance in the dataset. Figure 4.8 shows the distribution of labels labeled as 'None' and 'Happy', with the remaining emotions shown together under the 'Other' label.

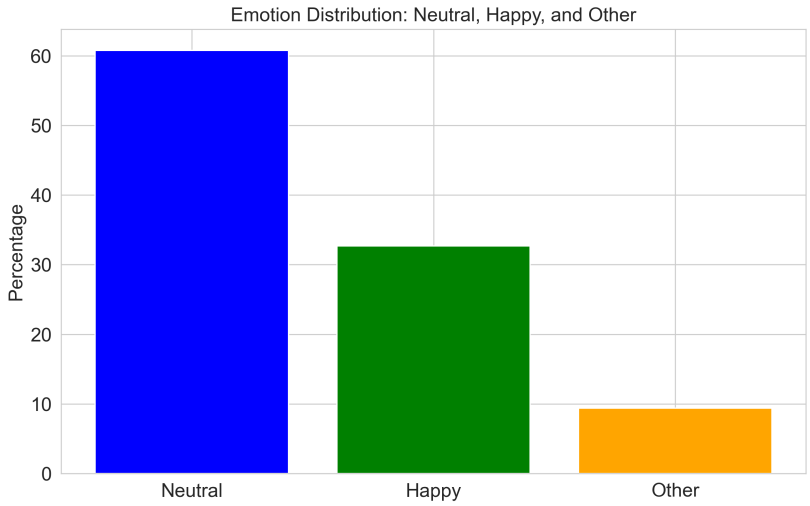

Figure 4.8. Percentage distribution of labels None, Happy, and Others in the Method II labels

The labels for the second method show a similar distribution, with Neutral rows representing 60% of the data. Happiness is also the most prevalent emotion in this dataset, with over 30% of labels.

Figure 4.9 shows the distribution of other emotions.

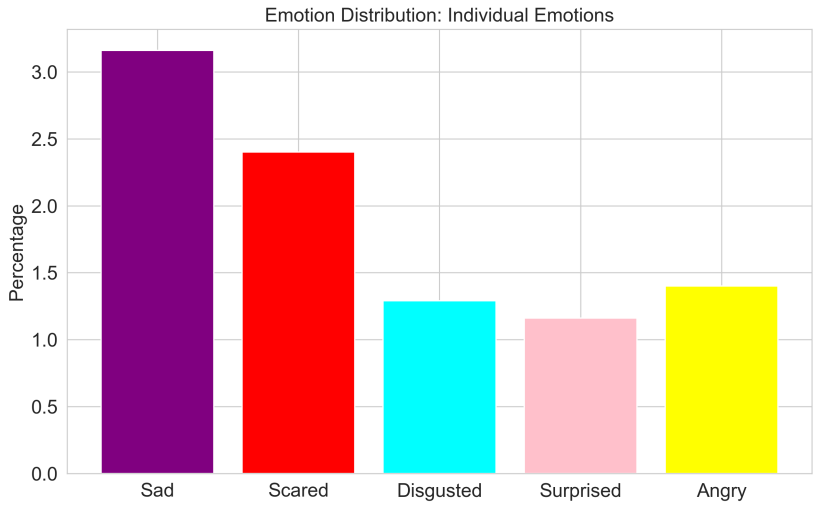

Figure 4.9. Distribution of other emotion labels in the Method II labels

Other emotions, accounting for almost 10% of the dataset, show a similar distribution to the first method. However, the percentages of labels for the remaining three emotions after sadness and fear increase. This dataset also requires normalization efforts, which will be discussed later. Figure 4.10 shows the percentage of timestamps with identified face embeddings.

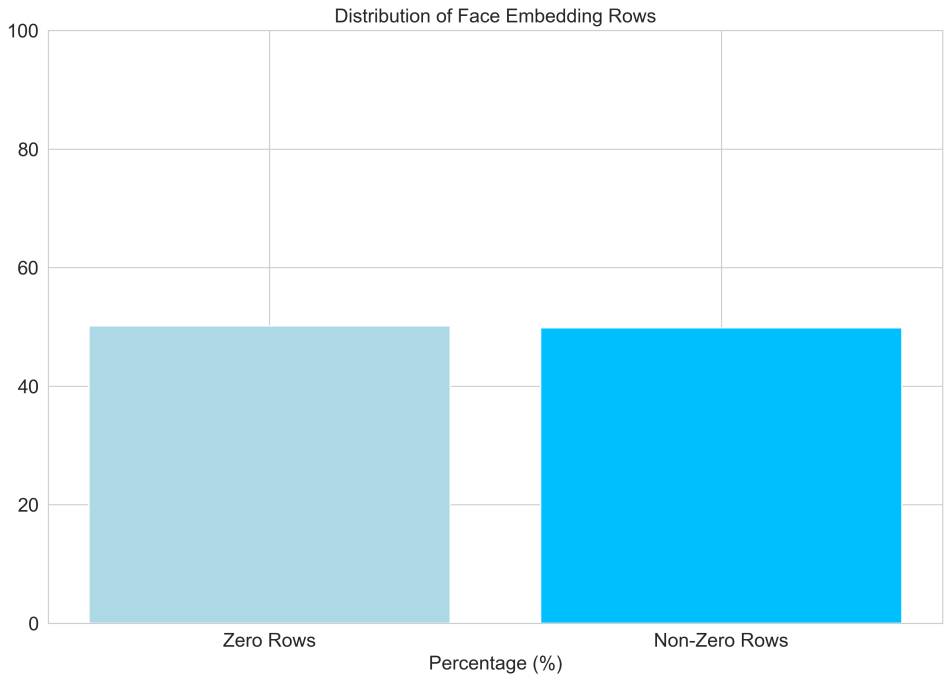

Figure 4.10. Percentage of timestamps with identified face embeddings

The resulting dataset is halved due to issues with face detection discussed in the Problems section.

Figure 4.11 shows the emotional states distribution with and without embeddings in Method I.

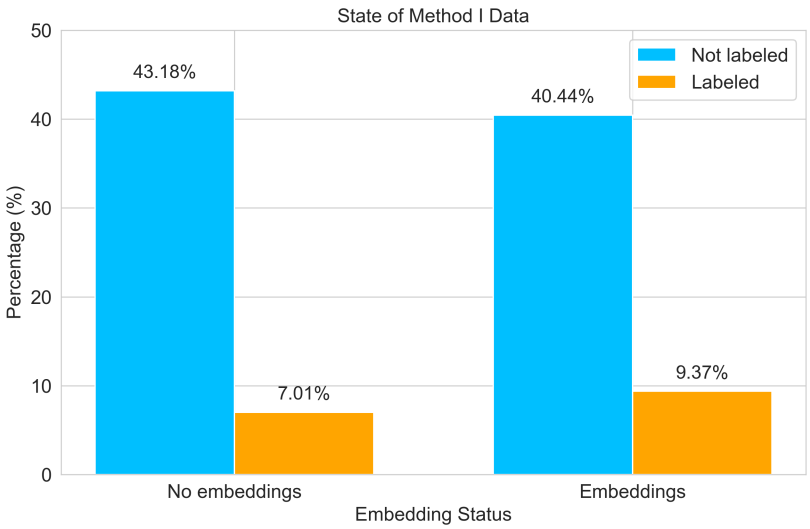

Figure 4.11. Distribution of Emotional States With and Without Embeddings in Method I

This figure illustrates the limitations of facial embedding extraction resulting from issues with face detection. Unfortunately, due to the 50% of timestamps with no detected faces, 7% of labeled data is not included in this analysis.

Figure 4.12 shows the emotional states distribution with and without embeddings in Method II.

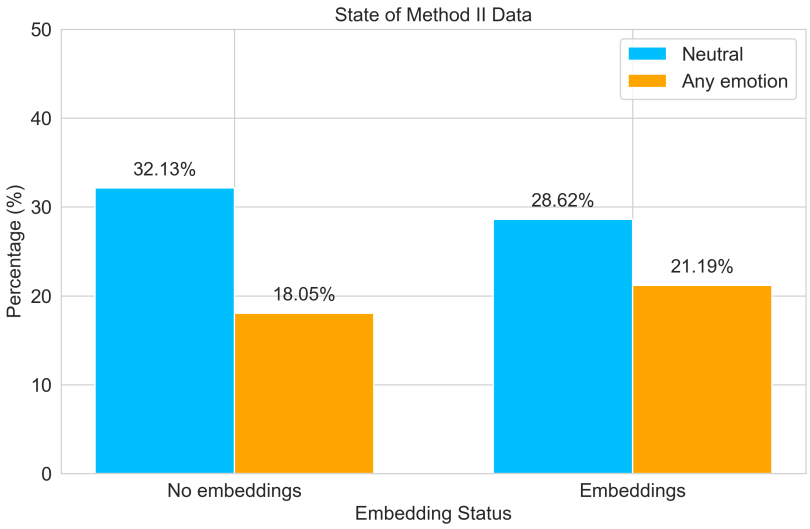

Figure 4.12. Distribution of Emotional States With and Without Embeddings in Method II

The second method's loss of data is significantly higher (18%), as more rows are taken into analysis due to a different approach in combining labels.

4.3. PROBLEMS

In 50% of cases, the facenet_pytorch was unable to correctly recognize the face and extract face embeddings from it. This was caused by many possible factors. In many of the videos provided by **ITU-YU**, participants were wearing facemasks, which makes it quite difficult for a

face recognition model to work properly. Moreover, in the video provided by **MAAP**, participants were sitting at a long distance from the camera, which made the identification harder, as emotion recognition relies on many face details unobtainable from a long distance. In many videos, other people besides participants were visible, making it harder to extract the proper face. From the side of labeling, labelers were not trained in recognizing feelings, especially in autistic people.

4.4. SUMMARY

This chapter outlined the dataset used in this thesis, detailing its structure, preprocessing, and the methodologies applied to integrate multimodal data. Originating from the **EMBOA** project, processed by Geisler et al. [2], and processed for the purpose of this thesis, the dataset provides a valuable foundation for emotion recognition research.

The chapter described the extraction of facial embeddings using state-of-the-art tools and the selection of key biosignals (**EDA**, **TEMP**, **HR**) to capture physiological responses. Synchronization and alignment processes were critical in ensuring the temporal consistency of the multimodal data. Two labeling methodologies were introduced, focusing on dominant emotions and percentage distributions to provide versatile labels for analysis.

Despite its strengths, the dataset presented several challenges, including difficulties with face detection due to masks, long distances from cameras, and the presence of multiple individuals in recordings. Additionally, inconsistencies in labeling highlighted the complexity of annotating emotions, particularly in neurodiverse populations.

Overall, the dataset, combined with the preprocessing strategies and fusion techniques described in this chapter, establishes a robust basis for the development and evaluation of emotion recognition models, which will be explored in subsequent chapters.

5. MODELS (OSKAR KOŁOSZKO, ADAM SOBCZUK)

This chapter presents the deep learning model used in this thesis, detailing its architecture, preprocessing pipeline, training process, and the tools employed for implementation. The model is designed to perform emotion recognition using multimodal data, combining facial embeddings and physiological signals.

The chapter begins by introducing the computational and software tools used to implement the model, including TensorFlow, Scikit-learn, and other supporting libraries. Following this, it describes the data preprocessing steps, such as normalization, sequence creation, and label encoding, which were critical for ensuring data suitability for training.

Next, the chapter delves into the architecture of the model, highlighting its key components, including the bidirectional LSTM layer and dense layers for dimensionality reduction and classification. The compilation and training process, including the use of the Adam optimizer and a custom loss function, are also detailed.

By describing the methodological and technical aspects of the model, this chapter provides a comprehensive understanding of its design and implementation, setting the stage for its evaluation in the subsequent chapters.

5.1. TOOLS USED (OSKAR KOŁOSZKO)

The model was running on an Intel(R) Core(TM) i5-8250U CPU with 8GB of RAM processor of a personal computer. To effectively implement the LSTM model, several software tools and frameworks were utilized. The choice of those tools was led by their efficiency in handling deep learning problems, their availability, and them being state-of-the-art technologies.

5.1.1. TENSORFLOW

TensorFlow is a software library for machine learning and artificial intelligence developed by Google. The system is flexible and can be used to create a broad number of algorithms and models, including training and inference algorithms for deep neural network models, such as LSTM [3]. TensorFlow offers APIs available for several languages, such as Python, JavaScript, C++, and Java. The Python API is presented as the most complete and the easiest to use, and it is used for the implementation of the model in this thesis [20]. Additionaly TensorFlow provides a user-friendly API, TensorFlow Keras, which allows to use an easy-to-understand, highly productive interface for solving deep learning problems [21].

5.1.2. SCIKIT-LEARN

Scikit-learn is a machine learning library for Python. Its compatibility with TensorFlow and many other libraries and ease of its tensor manipulation for operations such as one-hot encoding tensors, splitting data, and computing label weights made it a perfect choice for using in the preprocessing stage [4].

5.1.3. ADDITIONAL TOOLS

To facilitate even better data processing and to create an environment where all other programs could collaborate, additional tools were utilized.

JUPYTER NOTEBOOK

Jupyter Notebook is a fast interactive environment for prototyping and explaining code during both preprocessing and model training [22].

NUMPY

Numpy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices [23].

5.2. DATA PREPROCESSING (OSKAR KOŁOSZKO)

Data preprocessing is a critical step in the machine learning (ML) pipeline, involving the transformation of raw data into a format that is suitable for training ML models. In this study, the preprocessing was performed using Scikit-learn and Numpy libraries. The acquisition of the dataset was described in chapter 4.

5.2.1. DATA STANDARIZATION

Z-score data normalization was implemented using NumPy to scale all features to have a mean of 0 and standard deviation of 1. The Z-score method is characterized by handling the outliers better due to the way the standard deviation is calculated in the first place. The normalization was implemented according to the formula:

$$z = \frac{x - \mu}{\sigma} \tag{5.1}$$

Where:

$\mu -$ the mean of the population,
$\sigma -$ the standard deviation of the population.

5.2.2. BATCH CREATION

LSTM networks require input in the form of sequences to properly capture temporal dependencies. The data was therefore segmented into overlapping sequences of fixed length of 20 timestamps with a sliding window of 10.

5.2.3. LABEL MANIPULATION

For better extraction of dependencies in data and to make it compatible with specific functions, the data was encoded and normalized.

ONE-HOT ENCODING

As the labels for the Method I were provided as categorical labels, they needed to be transformed into numerical representations. For this, the one-hot encoding technique was used. One-hot encoding is an algorithm that converts categorical, labeled data into integers and then transforms those integers into binary encoded values [24].

DATA NORMALIZATION

The labels provided for Method II were already numerical. However, as their values were large, to prevent the possible problem of gradient explosion, all label values were divided by 100.

5.2.4. DATA SAMPLING

As presented in Figure 4.6, Figure 4.7, Figure 4.8 and Figure 4.9 in the chapter 4 the data is not evenly distributed. The *Neutral* and *Happy* classes are making up the significant part of the dataset. As the *Neutral* class is not an emotion and thus is not going to be taking part in the classification process as a label, its imbalance is not considered a problem as long as it is nearby any emotion and can be considered useful information. However, the *Happy* label can create a strong bias of the model towards it. To prevent this from happening, a decision of sampling and undersampling was made. In the first place, only the sequences that would contain any other emotion than *Happy* were sampled. This proved sufficient for Method II, in which the discrepancy between the *Happy* class and others was not as high as in Method I. Additionally, for the Method I all the sequences in which the *Happy* class was the only emotion class and the frequency of its occurance was higher than one, were ommited. This excluded all the sequences without any emotion class and downsampled the *Happy* class to the level of two next in order.

5.2.5. DATA LOADING

After preprocessing, the data was converted into tensors by the TensorFlow library to allow for efficient computation and batch processing.

5.2.6. SPLITTING THE DATA

The data was split into training and testing data to ensure robust model evaluation. A widely used 70/30 split was performed.

5.3. ARCHITECTURE (OSKAR KOŁOSZKO)

Both methods employ an LSTM-based deep neural network to optimally handle sequential data. Due to the differences in problem labeling and complexity, those two models differ slightly.

5.3.1. METHOD I

Figure 5.1 shows the architecture of the deep neural model created for Method I.

INPUT LAYER

The input layer is not shown in Figure 5.1. It accepts data in three-dimensional format with a shape of *BATCH_SIZE, SEQUENCE_LENGTH, INPUT_SIZE*. The *BATCH_SIZE* is the size of the batch, which refers to the number of training samples used in one iteration of model training. It is changeable and was set to 32. *SEQUENCE_LENGTH* is the length of the sequence represented by timestamps chosen in the dataset creation step described in chapter 4 and equal to 20. *INPUT_SIZE* represents the dimensionality of each feature in every timestamp and is equal to 515.

BIDIRECTIONAL LAYER

Bidirectional recurrent neural networks (BRNNs) work on the basis of presenting each training sequence both forwards and backwards to two distinct recurrent nets that are both connected to

the same output layer. This indicates that the BRNN has comprehensive, sequential knowledge about all points before and after each point in a specific sequence [25]. Here a bidirectional layer was implemented with the core of an LSTM layer with 64 units.

DENSE LAYERS

To reduce the dimensionality of the features extracted by LSTM, there were used 3 fully connected layers. The first two, with respectively 64 and 32 units, are using the ReLU activation function, defined as $f(x) = max(x,0)$ [26]. The function was chosen as being the default in state-of-the-art deep neural networks [27].

OUTPUT LAYER

The last dense layer has 6 units corresponding to the number of emotion labels and is using a Softmax activation function that converts a vector of numbers into a vector of probabilities [28].

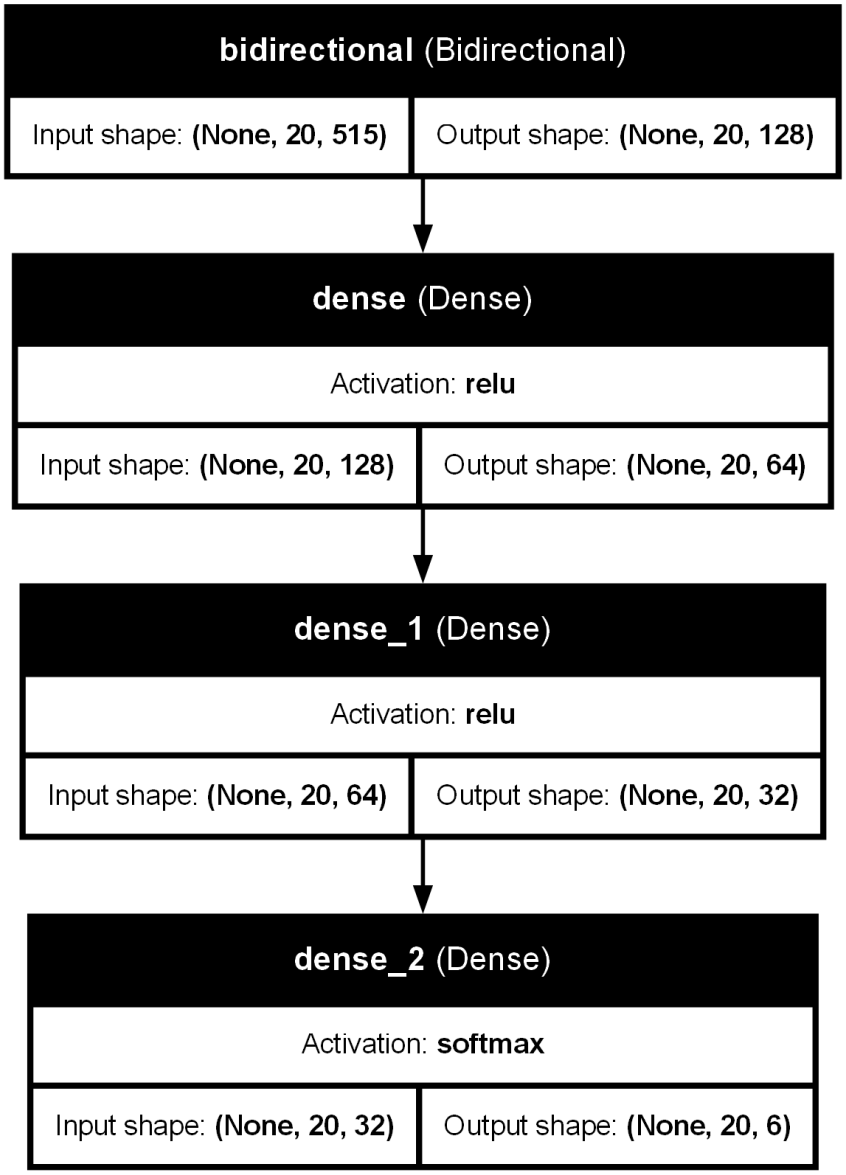

Figure 5.1. Architecture of model for Method I

5.3.2. METHOD II

Figure 5.2 shows the architecture of the deep neural model created for Method II.

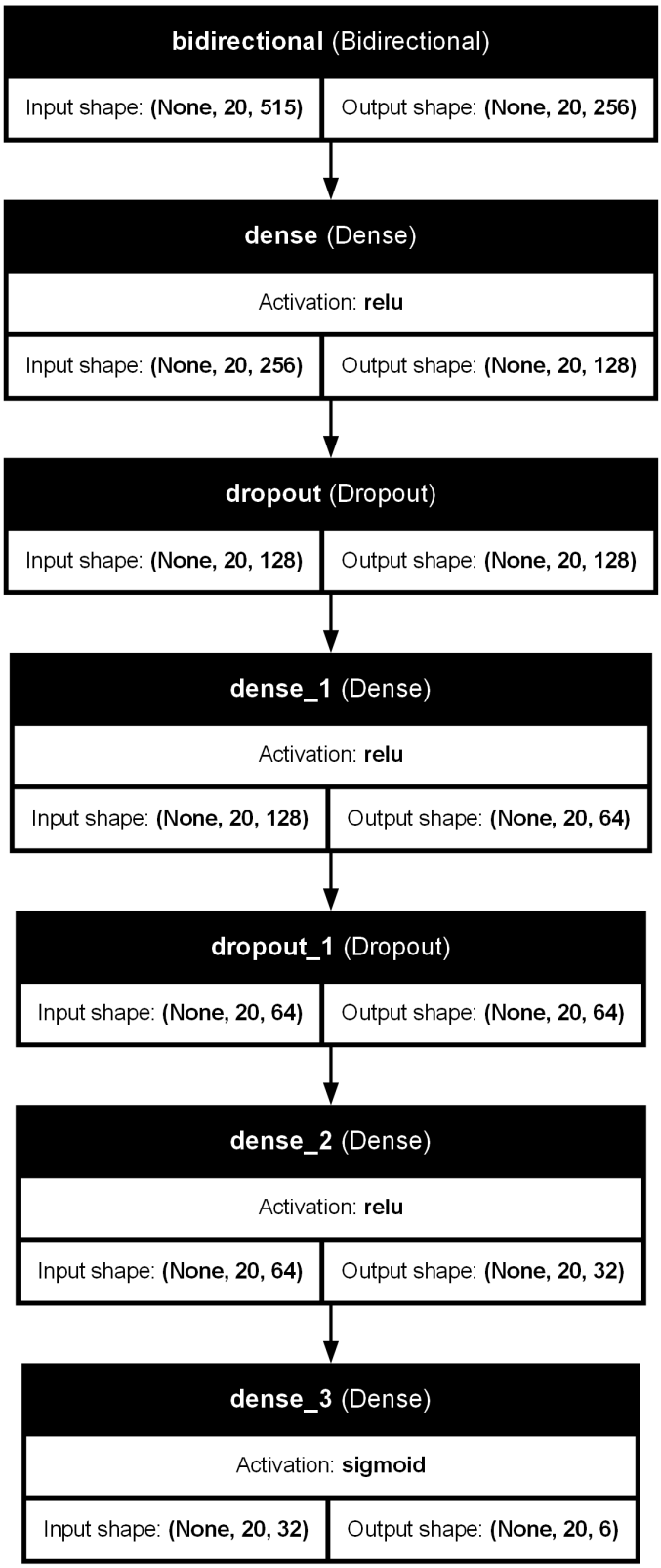

Figure 5.2. Architecture of model for Method I

INPUT LAYER

The architecture of the model for Method II is similar to the architecture for Model I, with small differences. The input layes were left the same as in the previous model.

BIDIRECTIONAL LAYER

As the model for Method II needs to predict values for every class, not only choose the best one, it needs additional LSTM layers for better understanding of the temporal data. Therefore, the bidirectional layer was implemented with the core of an LSTM layer with 128 units in comparison to 64 units for the Method I model.

DENSE LAYERS

As the dimensionality increased with additional LSTM units in the Bidirectional layer, for better reduction, there was an additional fully connected layer. Therefore, the model architecture consists of 4 fully connected layers. The first 3 have, respectively, 128, 64, and 32 units, are using the ReLU activation function.

DROPOUT LAYERS

As the complexity of the model increased, there was a need for a layer that would prevent overfitting. Therefore, two additional dropout layers with a parameter set to 0.1 were introduced between the first and second dense layers and between the second and third layers.

OUTPUT LAYER

The last dense layer has 6 units corresponding to the number of emotion labels and is using a Sigmoid activation function that converts a continuous real number into a range from 0 to 1. [29].

5.4. COMPILATION AND TRAINING (OSKAR KOŁOSZKO)

To compile and train the model, an optimizer, loss function, and evaluation metrics were chosen. Additionally, for better performance, a class weighting algorithm was implemented. Moreover, a decision was made to save the model for further use.

5.4.1. OPTIMIZER

Both models were compiled using the Adam optimizer. Adam is an extension of the Stochastic Gradient Descent method, being a technique for optimizing stochastic objective functions using gradients of the first order based on lower-order moment adaptive estimations [30]. It was chosen due to its strong, empirically proven performance.

5.4.2. LOSS FUNCTION

Due to differences in label types, a different loss function was implemented for each model.

METHOD I

The model uses a custom loss function built on the Categorical Cross Entropy (CCE) loss. The CCE loss was chosen as it was proven to have fast convergence, improved classification accuracy and good compatibility with the Softmax activation function [31]. The customization consists of applying a mask to all *Neutral* labels encoded as [0,0,0,0,0,0] to ensure that the model would learn to predict only the 'emotions' label.

METHOD II

The model uses a custom loss function built on the Mean Squared Error (MSE) loss. The MSE loss was chosen as it was proven to be a common choice in regression tasks [32]. The customization remains the same and consists of applying a mask to all *Neutral* labels encoded as [0,0,0,0,0,0] to ensure that the model would learn to predict only the 'emotions' label.

5.4.3. METRICS

As well as loss functions, different label types motivated the usage of different metrics.

METHOD I

As the model works on a classification problem, the accuracy metric was chosen to train the model.

METHOD II

As the model works on a multiclass regression problem, the MSE metric was chosen.

5.4.4. CLASS WEIGHTING

As the classes in Method II could not be fully equalized during preprocessing, an additional step of adding weights to classes using Scikit-learn was performed. This assigned weights to every class based on its frequency and made the learning process less biased.

5.4.5. SAVING

After training, both models were saved to a directory in the TensorFlow *.keras* format.

5.5. SUMMARY (ADAM SOBCZUK)

This chapter outlined the design, implementation, and training process of the deep learning models used in this thesis for emotion recognition. Models were built using state-of-the-art deep learning frameworks, such as TensorFlow and Scikit-learn, which were chosen for their efficiency and compatibility with the specific requirements of this project. Models integrate facial embeddings and biosignal data, leveraging a bidirectional Long Short-Term Memory (LSTM) network to capture temporal dependencies in the data.

The preprocessing steps, including data normalization, sequence creation, and label encoding, were critical for ensuring that the data was appropriately formatted for model training. Additionally, techniques like Z-score normalization and data undersampling were applied to mitigate issues such as class imbalance and improve model robustness.

The architecture of the model was designed to handle sequential data, with a bidirectional LSTM layer allowing the model to learn both forward and backward temporal dependencies. The dense layers that followed the LSTM layer help reduce the dimensionality of the features and ensure the model can efficiently process the extracted information. A custom loss function, built on categorical cross-entropy or mean squared error loss, was used to optimize models, and the Adam optimizer was chosen for its superior performance in training deep neural networks.

In terms of evaluation, the model for Method I uses accuracy as the primary metric to assess its classification performance, whereas the model for Method II uses mean squared error. Models were successfully trained and saved in the TensorFlow *.keras* format for further analysis and testing.

In conclusion, the design and implementation of models demonstrate the effective application of modern deep learning techniques to emotion recognition. The use of bidirectional LSTM layers, coupled with data preprocessing and advanced techniques like class balancing through masking, provides a robust framework for handling the challenges posed by multimodal emotion recognition tasks. This chapter has laid the technical foundation for the subsequent evaluation of the model's performance and its potential contributions to the field of affective computing.

6. **RESULTS** (OSKAR KOŁOSZKO, ADAM SOBCZUK)

This chapter presents the evaluation of the emotion recognition model and provides a detailed analysis of the results. The performance of the model is assessed based on two distinct methods: Method I, which focuses on classifying the dominant emotion at each timestamp, and Method II, which examines the intensity of emotions across multiple categories.

The chapter begins by introducing the evaluation metrics used to assess the models. For Method I, metrics such as Accuracy, Precision, Recall, and F1-Score are employed to evaluate the model's classification performance. For Method II, regression-based metrics including Mean Squared Error (MSE), Mean Absolute Error (MAE), and the custom Similarity metric are utilized to assess the intensity of the model's predictions. This allows for a comparative analysis of the models' performance on different tasks.

The results for both methods are then presented, including a quantitative evaluation of the emotion classification performance, confusion matrices, and visual analysis. Finally, comparisons between the two methods are made, highlighting their strengths and providing insights into the model's overall performance.

6.1. EVALUATION METRICS (OSKAR KOŁOSZKO)

Both models serve for multiclass classification. However, due to Method I being focused on the leading emotion and Method II having a more holistic approach and focusing on the state of all emotions in each second, two different approaches for evaluation had to be used.

6.1.1. MODEL I

As the first model's focal point is the main emotion in each second, it requires metrics that will focus on the ratio of classified emotions to misclassified ones. There are many metrics suited for this task. In this thesis, there were used Accuracy, Precision, Recall and F1-Score.

BINARY CONTINGENCY TABLE

Before explaining other metrics, let's introduce four basic outcomes of a classification task. Let's define an experiment for a condition using P positive instances and N negative instances. The Table 6.1 is an example of how to formulate the outcomes in a n x n contingency table. For this analysis, the $n = 2$, where n is equal to the number of classes [33].

Table 6.1. Binary contingency table [33]

Total population (P + N)	**Predicted Positive (PP)**	**Predicted negative (PN)**
Positive (P)	True positive (TP)	False negative (FN)
Negative (N)	False positive (FP)	True negative (TN)

The indicators used in Table 6.1 have the following meaning [33]:

- True positive (TP): number of cases in which the model correctly predicted the positive class,
- True negative (TN): number of cases in which the model correctly predicted the negative class,
- False positive (FP): number of cases in which the model wrongly predicted the positive class,

- False negative (FN): number of cases in which the model wrongly predicted the negative class.

In the context of the prepared model for exemplary emotion, *Happy* the indicators were understood as follows:

- True positive (TP): number of cases in which the model predicted the *Happy* class, while the true value was also *Happy*,
- True negative (TN): number of cases in which the model predicted any other emotion than *Happy*, while the true value was also other than *Happy*,
- False positive (FP): number of cases in which the model predicted the *Happy* class, while the true value was other than *Happy*,
- False negative (FN): number of cases in which the model predicted any other emotion than *Happy*, while the true value was *Happy*.

ACCURACY

Accuracy is the proportion of correctly predicted labels to the total population. It indicates the overall effectiveness of a classifier [34]. It is defined as [34]:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{6.1}$$

PRECISION AND RECALL

Precision denotes the portion of Predicted Positive cases that are correctly Real Positives or, in other words, the fraction of relevant instances among the retrieved instances. Inversely, Recall is the portion of Real Positive cases that are correctly Predicted Positive or, paraphrasing, the fraction of relevant instances that were retrieved [35]. Precision focuses on class agreement between the classifier's positive labels and the data labels, while recall of the effectiveness of a classifier to identify positive labels [34]. Precision and recall are then defined as [36]:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6.2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6.3}$$

F1-SCORE

F1-Score is the harmonic mean of Precision and Recall that focuses on relationships between a classifier's positive labels and those of the data [34]. It is defined as [37]:

$$\text{F1-score} = \frac{2 * TP}{2 * TP + FP + FN} \tag{6.4}$$

MICRO AND MACRO AVERAGED RESULTS

It is easily spotted that calculating metrics for every label results in a huge number of metrics, making understanding of the model harder. To make the interpretation of the result less challenging and get a condensed synopsis of the performance, the use of averaging can be implemented. There exist two variances of averaging: macro-averaging and micro-averaging [38]. The formula for macro-averaging a metric is defined as [38]:

$$\text{macro-averaged metric} = \frac{\text{metric}_A + \text{metric}_B + ...\text{metric}_N}{N} \tag{6.5}$$

Where:

metric_A,metric_B,metric_N – the values of metric for each class,
N – the number of classes.

The formulas for micro-averaging consist of adding metric components together so they vary acroos metrics. Taking as example two previously defined metrics of Precision and Recall, their following micro-averaged metric are defined as [38]:

$$\text{micro-averaged Precision} = \frac{\text{TP}_{Total}}{\text{TP}_{Total} + \text{FP}_{Total}} \tag{6.6}$$

$$\text{micro-averaged Recall} = \frac{\text{TP}_{Total}}{\text{TP}_{Total} + \text{FN}_{Total}} \tag{6.7}$$

Where:

TP_{Total} – the total count of True Positives,
FP_{Total} – the total count of False Positives,
FN_{Total} – the total count of False Negative.

Therefore, it can be concluded that macro-averaging demonstrates average performance across all classes, giving each class equal weight, while micro-averaging displays average performance across all classes and assigns equal weight to each instance [38].

6.1.2. MODEL II

The second method focuses on the intensity of emotion perception in each second. Therefore, there is a need for metrics that centre on how far away from predicting the intensity was the model. For this problem, the Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, R-Squared were chosen. Additionally, the Similarity metric defined in Geisler et al. [2] was used for easier comparison.

MEAN SIMILARITY

Similarity is a custom metric introduced by Geisler et al. [2]. Is using the absolute error, and to demonstrate the level of similarity, the value of the absolute error is additionally subtracted from one. To make it compatible with the model, a change of averaging was added. The new mean similarity value for the model can be computed using the formula in the Equation 6.8 located on the next page:

$$\text{Mean similarity}(y, x) = \frac{\sum_{i=0}^{N-1} 1 - |y_i - x_i|}{N} \tag{6.8}$$

Where:

$N -$ the number of samples,
$y_i -$ the predicted value,
$x_i -$ the true value.

MEAN ABSOLUTE ERROR (MAE)

Mean Absolute Error is a metric that calculates the average magnitude of the absolute errors between the predicted and actual values. MAE is calculated as the sum of absolute errors divided by sample size and is given by the formula [39]:

$$MAE(y, x) = \frac{\sum_{i=0}^{N-1} |y_i - x_i|}{N} \tag{6.9}$$

Where:

$N -$ the number of samples,
$y_i -$ the predicted value,
$x_i -$ the true value.

MEAN SQUARED ERROR (MSE)

Mean Squared Error is a metric that attaches significance to outliers, as it doubles the error. MSE is measured as the mean squared differences between actual output and predicted output, which is defined as [40]:

$$MSE(y, x) = \frac{\sum_{i=1}^{N} (y_i - x_i)^2}{N} \tag{6.10}$$

Where:

$N -$ the number of samples,
$y_i -$ the predicted value,
$x_i -$ the true value.

ROOT MEAN SQUARED ERROR (RMSE)

As the MSE introduces square to error, making it less intuitive to understand, the Root Mean Squared Error was introduced to make the error have the same units as predictions. RMSE is given by the formula [41]:

$$RMSE(y, x) = \sqrt{\frac{\sum_{i=1}^{N} (y_i - x_i)^2}{N}} \tag{6.11}$$

Where:

$N -$ the number of samples,
$y_i -$ the predicted value,
$x_i -$ the true value.

*6.2. **METHOD I*** (OSKAR KOŁOSZKO, ADAM SOBCZUK)

This section contains a description of the first method's model results. The tables included in this section contain precision, recall and F1-Score metrics, as well as their micro- and macro-averaged equivalents. The second part of the section provides a more precise insight by providing different confusion matrices and showing results for different emotions and scientific centers.

6.2.1. QUANTITATIVE RESULTS

The following paragraphs contain computed metrics, showing results for the Method I model.

EMOTION CLASSIFICATION METRICS

The classification metrics for the emotion detection model are presented in the tables below. The metrics were computed based only on the data in which any emotion was recognised. The data labeled as *Neutral* was used only in the learning process and was not utilized during evaluation of the model. These include Accuracy, Precision, Recall, and F1-Score for specific emotions, providing insights into the model's performance on the full testset. Additionally, micro-average and macro-average scores are reported to summarize overall performance.

LABEL COUNT

In Table 6.2 the count of individual emotions from the joint test set is presented. As the dataset is imbalanced, it has been decided to use other metrics beyond accuracy.

Table 6.2. Number of emotion classes for joint test set of GUT, ITU-YU, and MAAP for Method I

Emotion	Labels
Happy	90
Sad	271
Scared	130
Disgusted	13
Surprised	25
Angry	41
Total	572

METRICS FOR COMBINED DATASET

The Table 6.3 table summarizes the combined performance across all datasets (GUT, ITU-YU, MAAP).

Table 6.3. Metrics for Emotion Classification for GUT, ITU-YU and MAAP

Emotion	Accuracy	Precision	Recall	F1-Score
Happy	0.9441	0.8718	0.7556	0.8095
Sad	0.8829	0.9087	0.8388	0.8724
Scared	0.8969	0.7205	0.8923	0.7973
Disgusted	0.9790	0.5333	0.6154	0.5714
Surprised	0.9790	0.7600	0.7600	0.7600
Angry	0.9895	0.9268	0.9268	0.9268

It can be observed that every emotion has an accuracy score of over 88%. Although this result may be pleasing, because of the class imbalance, it is not that important. The *Happy* class has relatively high precision with a slightly lower recall reflected in the 0.81 value of the F1-Score, meaning it both classified some not *Happy* labels as such while also missing some of the true *Happy* labels. The *Sad* class has the best results of all, probably because of its frequency in the dataset. Its F1-Score shows a good balance between precision and recall. The last emotion with more than 10% of the total labels is the *Scared* class. Its F1-Score is similar to the *Happy* F1-Score, however, it has lower precision and higher recall, meaning that although it captures most of the *Scared* labels, it has many false positives. The *Disgusted* class, being the least numerous, also has the worst scores, with all the metrics below 0.65, meaning that the model was not able to properly capture the features of this class. The *Surprised* has average metrics, meaning a reasonably good performance. The *Angry* label, despite having a low number of labels, has a very high precision and recall, meaning that the model easily recognises this emotion. For better understanding the Table 6.4 assesses the overall model performance by computing additional micro and macro averages of precision recall and F1-Score.

Table 6.4. Micro and mcro-averaged metrics for Emotion Classification for GUT, ITU-YU, and MAAP

Average	Precision	Recall	F1-Score
Micro-average	0.8357	0.8357	0.8357
Macro-average	0.7869	0.7982	0.7896

The micro-average aggregates the contributions of all classes by summing true positives, false positives, and false negatives. It gives equal weight to each instance. The 0.84 value for precision, recall and F1-Score shows that the model achieves strong performance considering all classes proportionally. The macro-average, on the other hand, calculates metrics independently for each class and averages them, giving equal weight to each class. The value around 0.79 for all metrics reflects at the lower performance of minority classes.

6.2.2. VISUAL ANALYSIS

The following paragraphs contain computed figures and their descriptions, showing the distribution of predicted emotions for the first method model.

CONFUSION MATRIX

Figure 6.1 shows the number of predictions and true values of the Method I model. On the figure, it is visible that most of the values are concentrated on the diagonal. This indicated that the model correctly classifies the majority of the instances for each class. Additionally, it is observable that the model tends to confuse *Sad* class with *Scared* class. Moreover, it can be assumed that the model tends to missclassify the *Happy* class and the *Disgusted* class. However, the *Angry* class has a very good performance with very little missclassification.

Figure 6.1. Confusion matrix for Model I

BINARY CONFUSION MATRICES

For better understanding of emotion distribution and possibility for further comparison with Geisler et al. [2] work additional confusion matrix metrics were computed. They are providing insights into the model's distribution of true positives, true negatives, false positives and false negatives across various datasets: GUT, ITU-YU, MAAP, and their combined evaluation.

Figure 6.2 illustrates the emotion metrics distribution for the test set of Gdańsk University of Technology (GUT).

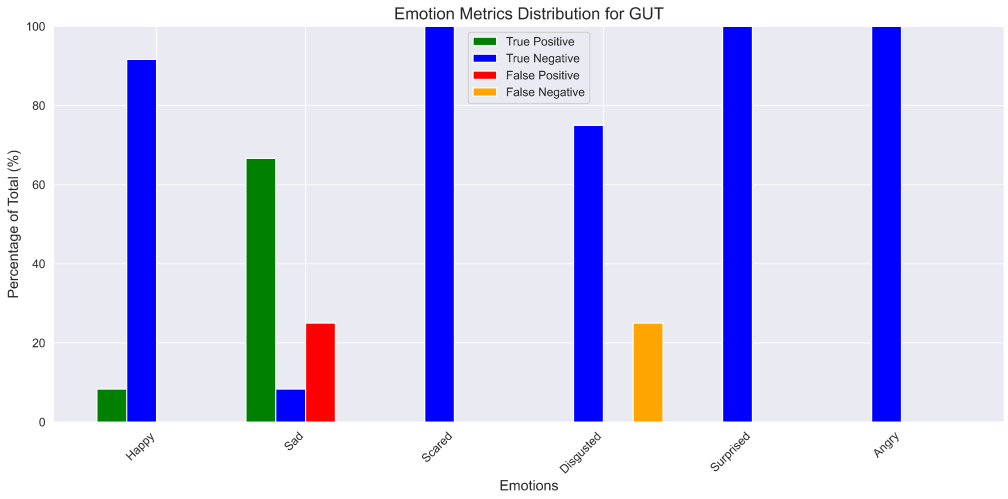

Figure 6.2. Emotion metrics distribution for GUT

It can be observed that the test set of GUT did not have any classes labeled as *Scared*, *Surprised* and *Angry*.

Figure 6.3 shows the emotion metrics distribution for the İstanbul Teknik Üniversitesi (ITU) - Yeditepe University (YU) test dataset.

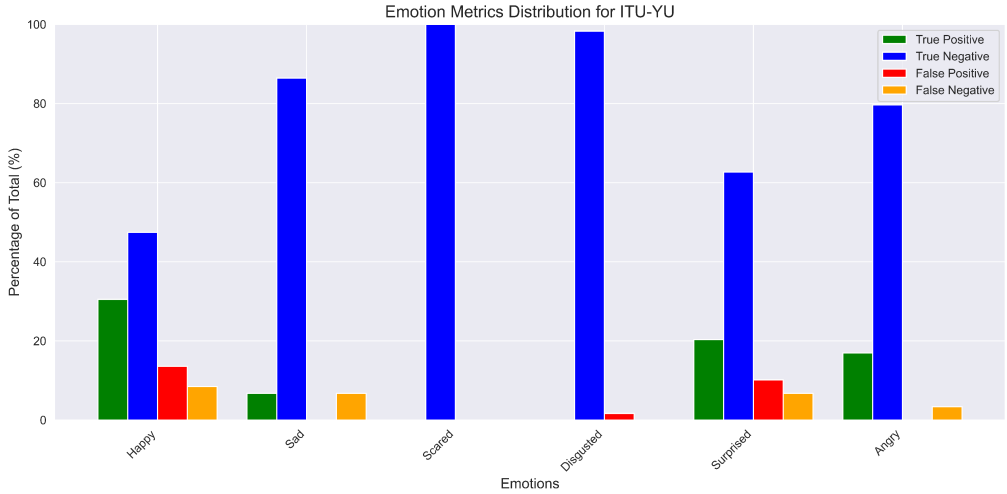

Figure 6.3. Emotion metrics distribution for ITU-YU

It can be observed that ITU-YU test set did not have any classes labeled as *Scared*.

Figure 6.4 presents the emotion metrics distribution for the Macedonian Association for Applied Psychology (MAAP) test dataset.

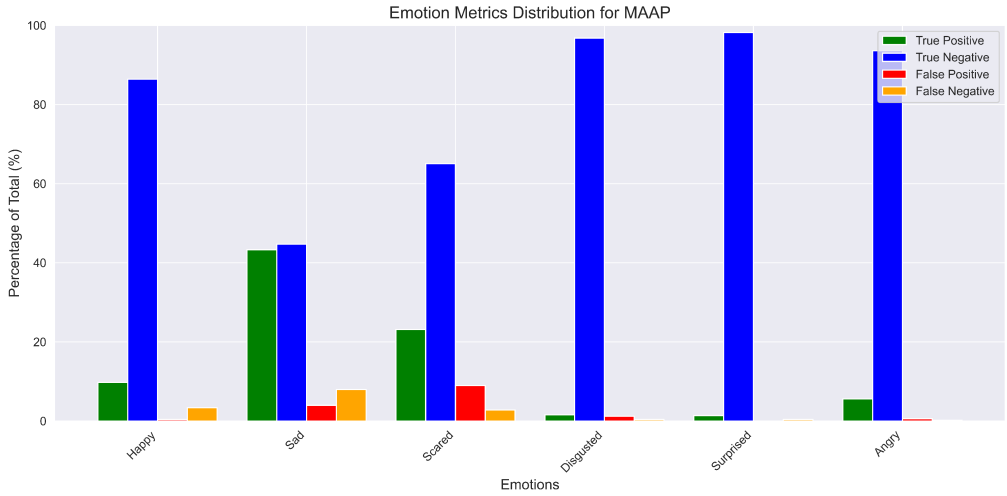

Figure 6.4. Emotion metrics distribution for MAAP

Additionally, a joint graph was computed for the collective test dataset of all scientific centers. Figure 6.5 summarizes the overall emotion metrics distribution across all test datasets.

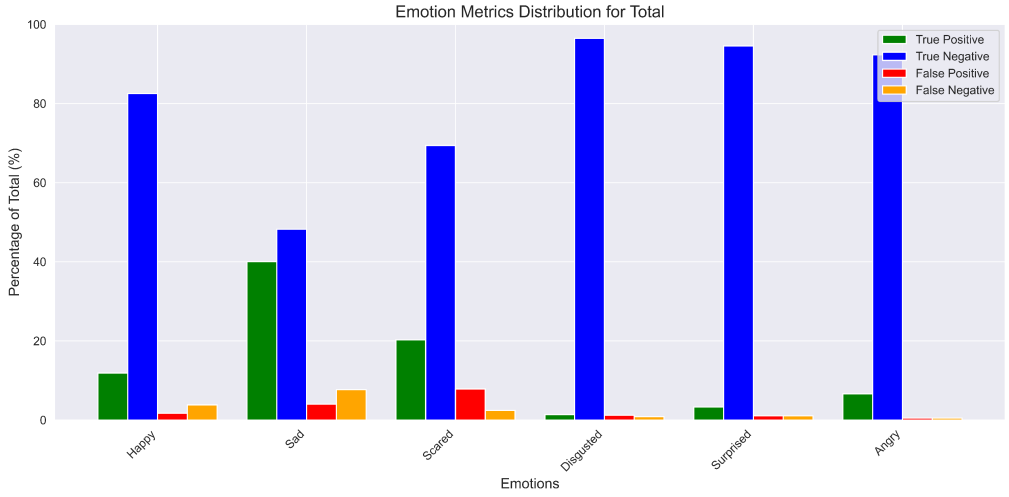

Figure 6.5. Overall emotion metrics distribution across all datasets

Based on the preceding figures, it can be inferred that the True Positive rate grows with the number of labels of the given emotion in the dataset. It can be observed that the *Sad* class performs the best, while the *Happy* and *Angry* also have moderate results. The *Scared* emotion has also good performance, but the higher rate of False Positives and False Negatives means that the model's ability to correctly identify *Scared* is impacted. The low count of all metrics beside True Negative of *Disgusted* class and *Surprised* class suggests a significant under-detection of those classes, likely due to the imbalance of those classes.

6.3. METHOD II (OSKAR KOŁOSZKO, ADAM SOBCZUK)

This section contains a description of the second method's model results. The tables included in this section contain similarity metrics, mean squared errors (MSE), mean absolute errors (MAE), and root mean squared errors.

6.3.1. QUANTITATIVE RESULTS

The process behind computing the metrics is described below, and subsequently, the model's performance is described through label counts and the similarity metric.

EMOTION CLASSIFICATION METRICS

The classification metrics for the emotion detection model for Method II are presented in the tables below. The metrics were computed based only on the data in which any emotion was recognized. The data labeled as [0,0,0,0,0,0] corresponding to *Neutral* class was not utilized during the evaluation of most of the model metrics. The metrics include Similarity, MSE, MAE, and RMSE for specific emotions, providing insights into the model's performance on the full test set. Additionally, the Neutral Similarity metric was computed. It is the only metric that utilizes the *Neutral* class and corresponds to average prediction for the [0,0,0,0,0,0] label.

LABEL COUNT

Table 6.5 shows the count of rows in which the given emotion has a non-zero value.

Table 6.5. Number of non-zero emotion classes for joint test set of GUT, ITU-YU, and MAAP for Method II

Emotion	Labels
Happy	1733
Sad	1020
Scared	764
Disgusted	426
Surprised	286
Angry	409
Total	4638

METRICS FOR COMBINED DATASET

Table 6.6 summarizes the combined performance across all datasets (GUT, ITU-YU, MAAP).

Table 6.6. Metrics for Emotion Classification for GUT, ITU-YU, and MAAP

Emotion	Similarity	MSE	MAE	RMSE	Neutral Similarity
Happy	0.8001	0.0709	0.1999	0.2662	0.7784
Sad	0.8286	0.0558	0.1714	0.2362	0.9283
Scared	0.8435	0.0380	0.1565	0.1950	0.9583
Disgusted	0.8140	0.0488	0.1860	0.2209	0.9389
Surprised	0.8229	0.0525	0.1771	0.2290	0.9552
Angry	0.8066	0.0544	0.1934	0.2332	0.9512
Overall	0.8193	0.0534	0.1807	0.2301	0.9184

Before discussing every emotion individually, let's introduce what the exemplary value of custom metrics means. The similarity value of 0.80 means that on average the predicted value differs from the true value by 0.20, which value, on the other hand, is equal to the 0.20 value of MAE. The Neutral similarity value of 0.90 means that on average the predicted value for a [0,0,0,0,0,0] label equals 0.1.

The Similarity and Neutral Similarity values for all emotions are relatively high, being an indicator of good model performance, with predicted values close to true values and values close to 0 for situations in which no emotion was detected. The only outlier is the *Happy* class, which tends to overstate the *Happy* emotion intensity in cases in which it is not observed. The average MAE value for all the classes is 0.18, meaning that on average the predictions are off by 0.18. However, as it was observed in the Neutral Similarity case, the *Happy* has a slightly bigger error compared to other emotions. This trend can be observed also both with MSE and RMSE. The average MSE equals 0.05, and the average RMSE equals 0.23. Class *Scared* appears to have the best performance, having lower values of all metrics.

6.4. MODEL COMPARISONS (ADAM SOBCZUK)

The comparison of the two methods (Method I and Method II) is presented in Table 6.7. The results demonstrate the performance of each model across various metrics, highlighting their respective strengths and weaknesses.

Table 6.7. Evaluation Metrics for Emotion Recognition Models

Metric	Method I	Method II	Subset Comparison
Full Dataset Metrics			
Accuracy	0.8357	N/A	0.4000 (Subset)
Precision	0.7869	N/A	N/A
Recall	0.7982	N/A	N/A
F1-Score	0.7896	N/A	0.2857 (Subset)
Mean Squared Error (MSE)	N/A	0.0208	0.0118 (Subset)
Similarity Score	N/A	0.9264	0.9335 (Subset)
Statistical Comparison		**Method I vs. Method II**	
T-Statistic		1.1613	
P-Value		0.2458	

6.4.1. METHOD I METRICS

Method I focuses on classifying the leading emotion in each instance, leveraging traditional classification metrics such as Accuracy, Precision, Recall, and F1-Score. Accuracy is calculated as a micro-average to incorporate evaluation of the model as a whole, while the Precision, Recall, and F1-score are calculated as macro-average to take into account differences in data distribution. As shown in Table 6.7, Method I achieves an overall accuracy of 83.57%, correctly predicting the dominant emotion in the majority of cases. Precision (78.69%) and Recall (79.82%) indicate a balanced performance, with the F1-Score (78.96%) reinforcing the robustness of this approach. These metrics suggest that the model performs reliably, particularly for well-represented emotions in the dataset.

6.4.2. METHOD II METRICS

Method II evaluates the intensity of emotions using regression-based metrics, such as Mean Squared Error (MSE) and a custom Similarity Score. Method II in model comparisons has to take into account the dominant emotion resulting in the so-called hard labels, to be evaluated on the same grounds as the model from the first method. The low MSE value of 0.0208 indicates that the model predictions are close to the true probabilities. Additionally, the Similarity Score of 92.64% demonstrates that the predicted distributions align closely with the ground truth, highlighting the model's capability to represent emotional states comprehensively.

6.4.3. SUBSET COMPARISON

To further analyze the performance of both methods, a subset comparison was conducted. As seen in Table 6.7, Method I's performance on the subset drops significantly, with an accuracy of 40.00% and an F1-Score of 0.2857. This decline suggests variability in the model's ability to generalize to smaller or imbalanced datasets. In contrast, Method II maintains high performance on the subset, achieving a reduced MSE of 0.0118 and a Similarity Score of 93.35%. These results indicate that Method II is more consistent and reliable across varying data sizes and distributions.

6.4.4. STATISTICAL COMPARISON

A statistical comparison of the predictions from both methods was performed using a t-test. The t-statistic (1.1613) and p-value (0.2458) suggest no statistically significant difference between the methods in dominant emotion prediction when Method II's probabilistic outputs are converted

to hard labels. This result highlights the comparability of the methods in this context while acknowledging their distinct approaches.

6.4.5. INSIGHTS AND RECOMMENDATIONS

Strengths of each method and the comparison are listed below:

- **Strengths of Method I**: High accuracy and balanced precision-recall scores make this approach suitable for tasks requiring discrete emotion classification. The straightforward interpretability of single-label predictions is advantageous in applications where identifying the leading emotion is critical.

- **Strengths of Method II**: Superior performance in predicting emotion intensities and handling probabilistic outputs offers a nuanced understanding of emotional states. Consistent performance on subsets suggests robustness to data variability, making it well-suited for scenarios requiring reliability across diverse datasets.

- **Comparison**: While Method I excels in single-label classification, Method II provides richer, probabilistic outputs with comparable classification performance. The choice of method should align with the specific requirements of the application, such as whether a discrete or continuous representation of emotions is needed.

6.4.6. SUMMARY OF RESULTS

The analysis of the two methods underscores their complementary strengths. Method I's ability to classify dominant emotions with high accuracy and Method II's capability to model nuanced emotional distributions present distinct advantages. The choice between these methods depends on the specific objectives and constraints of the task at hand.

6.5. SUMMARY (ADAM SOBCZUK)

This chapter provided a comprehensive evaluation of the emotion recognition model, focusing on the performance of both Method I and Method II. The evaluation metrics and results highlighted the strengths and limitations of each approach.

Method I, which focused on classifying the dominant emotion, achieved an overall accuracy of 83.57%, with high performance on emotions such as *Sad* and *Angry*, which were well-represented in the dataset. However, Method I struggled with the under-represented emotions like *Disgusted* and *Surprised*, leading to lower scores for these classes. The confusion matrix confirmed that the model tended to confuse certain emotions, such as *Sad* and *Scared*, and had difficulty with distinguishing *Happy* and *Disgusted*.

On the other hand, Method II, which evaluated emotion intensities, demonstrated strong performance in terms of similarity and regression metrics. The model's Mean Squared Error (MSE) was low (0.0208), indicating that the predictions were close to the true values. Additionally, the Similarity metric showed that the model was able to predict emotion intensity with a high degree of accuracy (92.64%).

A direct comparison between the two methods revealed that while Method I was more suited for discrete emotion classification, Method II performed better in capturing emotion intensity and was more robust across varying dataset sizes and distributions. This makes Method II more reliable for tasks that require nuanced emotional analysis, while Method I is more appropriate for applications where clear, dominant emotion identification is needed.

The statistical analysis between both methods confirmed no significant difference in performance when Method II's probabilistic outputs were converted to hard labels, underscoring the complementary nature of the two approaches. In conclusion, the results highlight the strengths of both methods, and the choice between them should be based on the specific requirements of the application, such as whether a discrete or continuous representation of emotions is needed.

7. **COMPARISON** (OSKAR KOŁOSZKO, ADAM SOBCZUK)

This chapter examines the comparative performance of the proposed emotion recognition models, Model I and Model II, against the FaceReader software used in Geisler et al. [2] paper and benchmarks established in relevant literature. The focus lies in evaluating the methodological advancements introduced in the thesis, with particular emphasis on how these models address the challenges of recognizing nuanced emotional states in children on the autism spectrum.

The evaluation spans multiple datasets from three distinct research centers—GUT, ITU-YU, and MAAP—offering a comprehensive view of the models' performance. Metrics such as accuracy, precision, recall, F1-score, and similarity are employed to assess Model I's and Model II's capabilities. The chapter also juxtaposes these findings with FaceReader's results to highlight the strengths and limitations of each method in various emotional categories.

Finally, the performance of the proposed models is situated within the broader context of emotion recognition research, leveraging insights from state-of-the-art literature. This analysis underscores the practical implications of multimodal approaches and highlights areas for future refinement.

7.1. PERFORMANCE OF MODELS IN RELATION TO THE FACEREADER SOFTWARE (ADAM SOBCZUK)

This section is focused on evaluating both methods models against the FaceReader software's results computed in Geisler et al. [2] thesis. The comparison of the models is drawn on the whole dataset and each individual institution.

7.1.1. MODEL I

The first method model's evaluation is described through the visual analysis of heatmaps for each institution, contrasting this thesis model's metrics with the results from FaceReader software on the whole test set.

GUT DATASET

The classification performance of Model I on the GUT dataset is detailed in Table 7.1.

Table 7.1. Metrics for Emotion Classification for GUT

Emotion	Accuracy	Precision	Recall	F1-Score
Happy	1.0000	1.0000	1.0000	1.0000
Sad	0.7500	0.7273	1.0000	0.8421
Scared	1.0000	0.0000	0.0000	0.0000
Disgusted	0.7500	0.0000	0.0000	0.0000
Surprised	1.0000	0.0000	0.0000	0.0000
Angry	1.0000	0.0000	0.0000	0.0000
Micro-average	N/A	0.7500	0.7500	0.7500
Macro-average	N/A	0.2879	0.3333	0.3070

This performance is further illustrated in Figure 7.1, which visualizes the distribution of emotion metrics for this dataset.

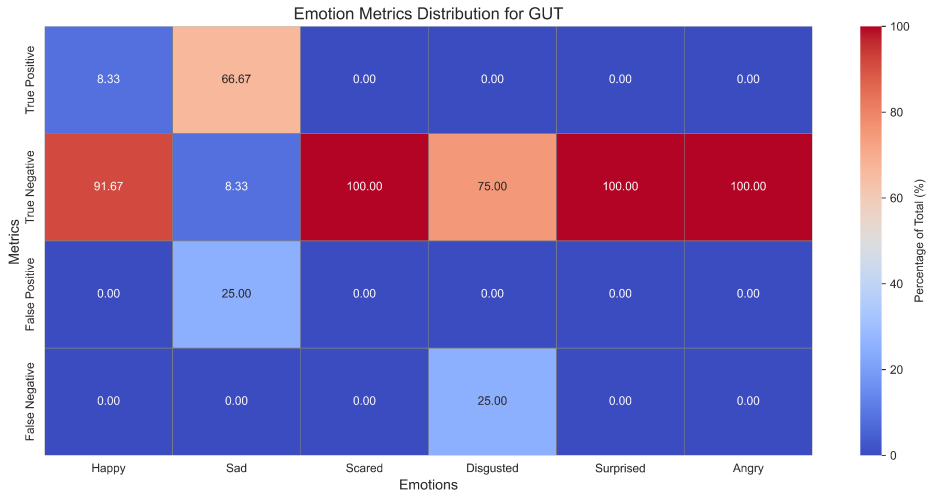

Figure 7.1. Emotion metrics distribution for GUT

Model I demonstrated a significant improvement over the FaceReader model employed by Geisler et al. [2]. Notably, the accuracy for predicting *Happiness* was 100%, a marked improvement. For the *Sadness* emotion, Model I achieved an accuracy of 75%, albeit with a slight disadvantage of an increased number of false positives.

ITU-YU DATASETS

The classification metrics for the ITU-YU dataset are summarized in Table 7.2.

Table 7.2. Metrics for Emotion Classification for ITU-YU

Emotion	Accuracy	Precision	Recall	F1-Score
Happy	0.7797	0.6923	0.7826	0.7347
Sad	0.9322	1.0000	0.5000	0.6667
Scared	1.0000	0.0000	0.0000	0.0000
Disgusted	0.9831	0.0000	0.0000	0.0000
Surprised	0.8305	0.6667	0.7500	0.7059
Angry	0.9661	1.0000	0.8333	0.9091
Micro-average	N/A	0.7458	0.7458	0.7458
Macro-average	N/A	0.5598	0.4777	0.5027

Figure 7.2 visually represents the distribution of emotion metrics for this dataset.

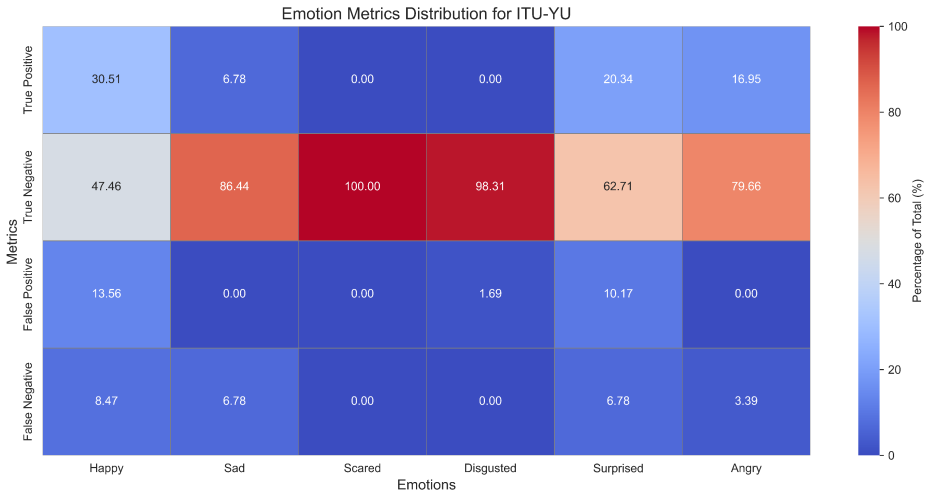

Figure 7.2. Emotion metrics distribution for ITU-YU

On the ITU-YU dataset, Model I significantly outperformed the FaceReader model, achieving 93% accuracy in predicting *Sadness*. The model also showed substantial improvement in detecting *Anger*, with a resulting accuracy of 96%. Similarly, the accuracy for the *Surprise* emotion was 83%, further demonstrating the robustness of the model compared to the baseline.

MAAP DATASET

The performance of Model I on the MAAP dataset is summarized in Table 7.3.

Table 7.3. Metrics for Emotion Classification for MAAP

Emotion	Accuracy	Precision	Recall	F1-Score
Happy	0.9621	0.9608	0.7424	0.8376
Sad	0.8802	0.9156	0.8444	0.8785
Scared	0.8822	0.7205	0.8923	0.7973
Disgusted	0.9840	0.5714	0.8000	0.6667
Surprised	0.9960	1.0000	0.7778	0.8750
Angry	0.9920	0.9032	0.9655	0.9333
Micro-average	N/A	0.8483	0.8483	0.8483
Macro-average	N/A	0.8453	0.8371	0.8314

The corresponding distribution of emotion metrics is depicted in Figure 7.3.

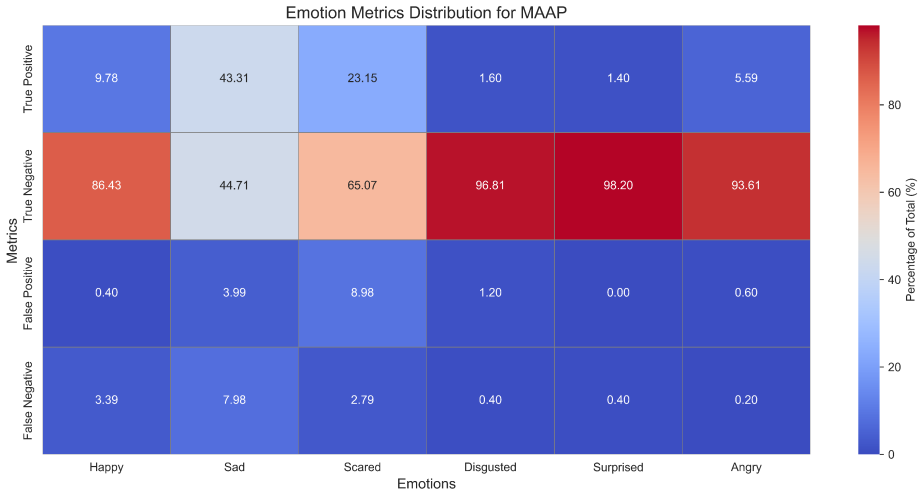

Figure 7.3. Emotion metrics distribution for MAAP

On the MAAP test dataset, Model I demonstrated improved performance in predicting *Fear*, with an accuracy increase of 2.79% compared to the FaceReader model, which exhibited 4.77% false negatives for this emotion.

7.1.2. PERFORMANCE ON THE FULL TEST SET

The analysis with the comparison on the full test set (GUT, ITU-YU, and MAAP) is described below. Figure 7.4 presents the overall emotion metrics distribution for the full test dataset, encompassing all centers.

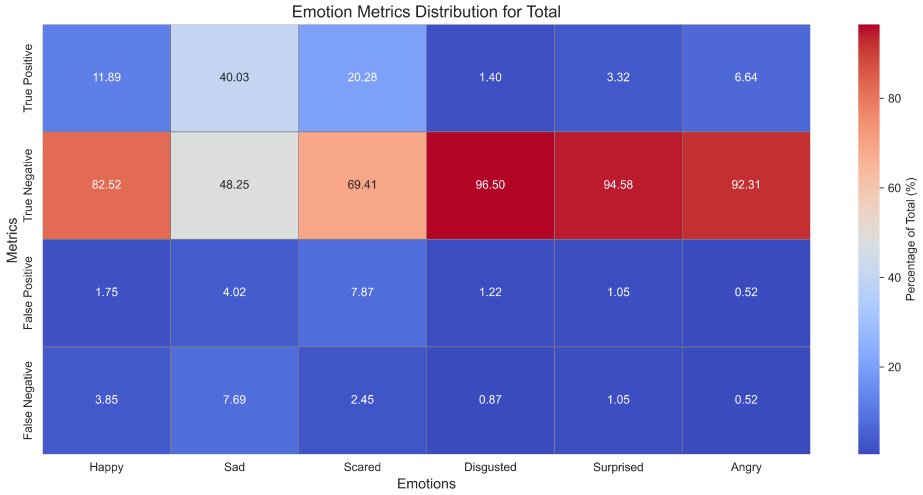

Figure 7.4. Emotion metrics distribution for the Full Test dataset

Model I introduced in this thesis showed a marked decrease in false positives for *Happiness*, *Surprise*, and *Anger* compared to the FaceReader software utilized in the experiments conducted by Geisler et al. [2]. However, it exhibited an increase in false positives for *Sadness*, *Fear*, and *Disgust*. Similarly, false negatives were reduced for the *Happiness* and *Surprise* emotions, but an increase was observed for *Sadness*, *Fear*, *Disgust*, and *Anger*.

Despite these trade-offs, the overall performance of Model I surpassed that of the FaceReader software. Model I achieved an accuracy of 83.57%, a substantial improvement over the 53.69% accuracy reported for FaceReader. This enhancement in performance underscores the efficacy of the methodological improvements and will be further compared against findings in the literature.

7.1.3. MODEL II

Then, the second method's model is compared to the relevant results of the FaceReader software gathered by Geisler et al. [2].

Figure 7.5 presents the performance of Model II on the full test set based on the similarity metric.

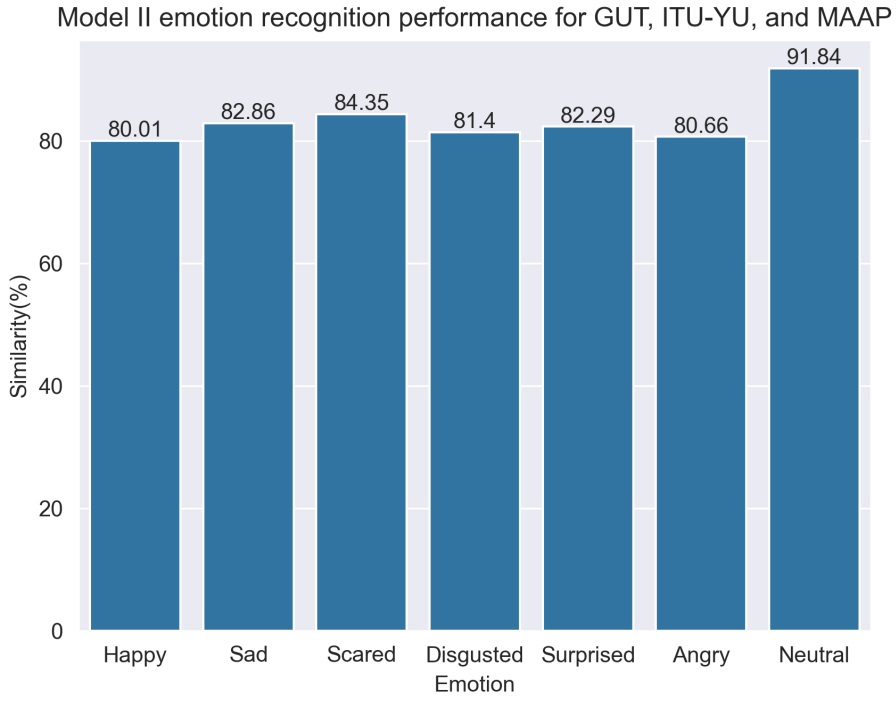

Figure 7.5. Average similarity of the whole test set

Table 7.4 shows the average similarity for each emotion.

Table 7.4. Model II Similarity metric across whole test set

Emotion	Similarity
Happy	80.01%
Sad	82.86%
Scared	84.35%
Disgusted	81.40%
Surprised	82.29%
Angry	80.66%
Neutral	91.84%

The comparison of emotion similarity metrics between the FaceReader software and Model II highlights the performance differences across various emotions.

Happy: The FaceReader software achieved a similarity score of 73.95%, while Model II demonstrated a higher performance with a similarity score of 80.01%. This improvement indicates the robustness of Model II in capturing the subtleties of *Happiness*.

Sad: For the *Sad* emotion, FaceReader achieved 90.06%, outperforming Model II, which scored 82.86%. This suggests that while Model II performs well, there is room for improvement in recognizing *Sadness*.

Scared: FaceReader achieved a near-perfect similarity score of 95.07% for the *Scared* emotion, significantly outperforming Model II's score of 84.35%. This indicates that Model II struggles slightly in accurately modeling this emotion.

Disgusted: The similarity score for the *Disgusted* emotion was 95.10% for FaceReader, compared to 81.40% for Model II. This difference underscores the challenges Model II faces in accurately detecting *Disgust*.

Surprised: For *Surprise*, FaceReader achieved a higher similarity score of 94.76% compared to Model II's 82.29%. This suggests that FaceReader excels in detecting the characteristic features of *Surprise*.

Angry: FaceReader attained 89.44% similarity for the *Angry* emotion, surpassing Model II's score of 80.66%. The discrepancy might reflect challenges in distinguishing *Angry* expressions in multimodal data.

Neutral: Conversely, Model II significantly outperformed FaceReader in detecting the *Neutral* state, achieving a similarity score of 91.84%, compared to FaceReader's 52.46%. This demonstrates Model II's superior capability in accurately identifying the absence of emotional expression.

In summary, while Model II excels in recognizing *Neutral* expressions and shows competitive performance for emotions like *Happy* and *Scared*, FaceReader outperforms it in detecting *Sadness*, *Anger*, *Surprise*, and *Disgust*. These results underline the complementary strengths of the two approaches and suggest potential areas for further refinement in Model II to match or exceed FaceReader's performance in specific emotional categories. Figure 7.6 displays the average similarity for each emotion on each institute test set.

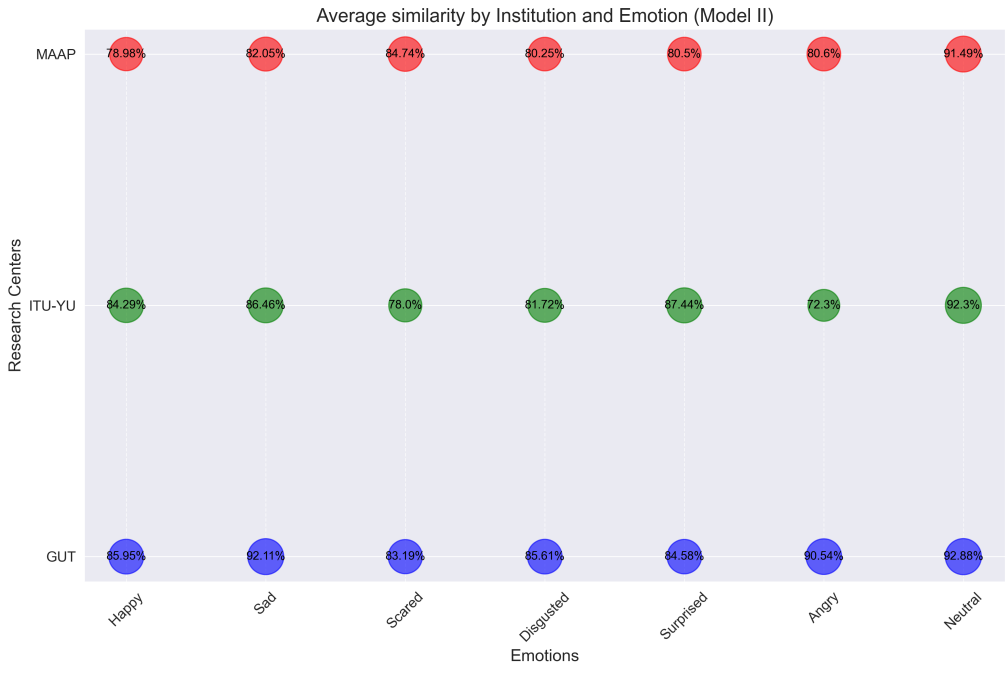

Figure 7.6. Average Model II similarity by Institution for each emotion.

The evaluation of emotion similarity across research centers for Model II highlights its performance trends compared to FaceReader, showcasing strengths and areas for improvement.

GUT

A comparison of results on the GUT dataset for each emotion is described in the following list.

- **Happy**: Model II achieved a similarity of 85.95%, surpassing FaceReader's 75.37%.
- **Sad**: Model II scored 92.11%, exceeding FaceReader's 86.47%.
- **Scared**: Model II achieved a similarity of 83.19%, compared to FaceReader's 97.81%.
- **Disgusted**: Model II reported 85.61%, compared to FaceReader's 95.79%.
- **Surprised**: Model II scored 84.58%, underperforming FaceReader's 95.67%.
- **Angry**: Model II reported 90.54%, marginally better than FaceReader's 88.87%.
- **Neutral**: Model II significantly outperformed FaceReader, achieving 92.88% compared to 47.33%.

ITU-YU

Subsequently, this list contains the comparison of results on the ITU-YU dataset.

- **Happy**: Model II achieved 84.29%, surpassing FaceReader's 82.96%.
- **Sad**: Model II scored 86.46%, closely aligned with FaceReader's 86.61%.
- **Scared**: Model II reported 78.00%, significantly lower than FaceReader's 99.16%.
- **Disgusted**: Model II achieved 81.72%, compared to FaceReader's 97.21%.
- **Surprised**: Model II scored 87.44%, compared to FaceReader's 95.74%.
- **Angry**: Model II reported a similarity of 72.30%, underperforming compared to FaceReader's 85.46%.
- **Neutral**: Model II achieved 92.30%, significantly better than FaceReader's 47.70%.

MAAP

Lastly, the comparison of results on the MAAP dataset is listed below.

- **Happy**: Model II achieved a similarity of 78.98%, surpassing FaceReader's 71.65%.
- **Sad**: Model II scored 82.05%, lower than FaceReader's 89.57%.
- **Scared**: Model II achieved 84.74%, compared to FaceReader's 93.03%.
- **Disgusted**: Model II reported 80.25%, underperforming FaceReader's 93.47%.
- **Surprised**: Model II scored 80.50%, compared to FaceReader's 93.32%.
- **Angry**: Model II reported 80.60%, marginally lower than FaceReader's 89.01%.
- **Neutral**: Model II significantly outperformed FaceReader, achieving 91.49% compared to 52.96%.

SUMMARY

Model II demonstrates notable strengths in detecting *Happy*, *Neutral*, and *Angry* emotions, particularly excelling in the identification of *Neutral* expressions, where FaceReader consistently underperforms. However, FaceReader outshines Model II in recognizing emotions such as *Scared*, *Disgusted*, and *Surprised*, particularly in datasets from ITU-YU and MAAP.

These results indicate that Model II offers balanced performance across datasets and is particularly effective in capturing subtle emotional states, such as *Neutral*. In contrast, FaceReader remains highly competitive for more pronounced emotions, such as *Scared* and *Disgusted*. This will be further contested in the literature review.

7.2. LITERATURE COMPARISON (OSKAR KOŁOSZKO)

Emotion recognition for children on the autism spectrum requires a robust model capable of understanding multimodal, often sequential data. In the literature, many different approaches are found that include CNNs, autoencoders, or LSTMs. This section compares those state-of-the-art models with proposed solutions.

7.2.1. LITERATURE SURVEY

The search string used during the literature research is presented in Listing 7.1. Furthermore, the results were filtered to a maximum of 9 years old. However, only one was actually older than 5 years.

Listing 7.1. Search string segments used during literature search

```
a) emotion recognition OR affective computing OR emotion detection
b) autism OR autistic children OR asd OR autism spectrum disorder
c) model OR algorithm OR framework OR approach
d) multimodal OR multi-modal OR multiple modalities OR fusion
e) evaluation OR results OR performance OR validation
OR accuracy OR experiment
```

Key studies in the field of emotion recognition for people with autism spectrum disorder have employed different approaches and methodologies to tackle the problem. Most of them are using a unimodal approach as opposed to the proposed model. This gives a chance to check if adding additional modalities makes a difference in the performance of the model. The modes used in unimodal do not vary much, being mostly facial expressions, with Heart Rate being an exception introduced in Ali, Shah, and Hughes [42]. As it comes to labeling, half of the models work with children, while the other half uses the help of adult participants. The smaller part of papers uses only basic labeling for emotions stated as *Positive*, *Negative* and *Neutral* while the bigger part leans towards the six basic emotions of Ekman's model or a variation of it.

The biggest diversity can be observed in models used across the literature. The largest part is created by various CNN models [43], [44], [45], [46], [47], [48]. However, there are also approaches using the Generative Adversarial Networks (GAN) [42], Support Vector Machines (SVM) [49] or even approaches based on incorporating Principal Component Analysis with an eigenvector-based algorithm [50]. This makes the proposed model using LSTM have a diverse benchmark.

7.2.2. RESULTS COMPARISON

The Table 7.5 presents the comparison between the proposed model and the benchmarks from the literature.

Table 7.5. Literature benchmarks

Source	Model	Modality	Modes	Participant age	Labels	Accuracy
M. Talaat et al. [43]	Xception CNN	Unimodal	Facial expressions	Children	Ekmans emotions	95.23%
Rani [49]	Support Vector Machine	Unimodal	Facial expressions	Children	Ekmans emotions	90.00%
Jingjing et al. [47]	Spatial-Temporal Graph Convolutional Network	Multimodal early-fusion	Skeleton pose, facial expressions	Children	Only positive, negative, neutral emotions	85.40%
Smitha and Vinod [50]	PCA + eigenvectors	Unimodal	Facial expressions	Children	Ekmans emotions	82.30%
Syed et al. [45]	CNN	Unimodal	Facial expressions	Children	Only positive, negative, neutral emotions	80.07%
Abu-Nowar et al. [46]	Ensemble CNN	Unimodal	Facial expressions	Adult	Ekmans emotions	66.00%
Fuentes-Alvarez et al. [44]	CNN	Unimodal	Facial expressions	Adult	Ekmans emotions	63.60%
Arabian et al. [48]	Graphical Convolution Network (GCN)	Unimodal	Facial expressions	Adult	Ekmans emotions	58.76%
Ali, Shah, and Hughes [42]	Transfer-based Expression Recognition Generative Adversarial Network (TER-GAN)	Unimodal	Heart Rate	Children	Only positive, negative, neutral emotions	39.80%
FaceReader	CNN	Unimodal	Facial expressions	Children	Ekmans emotions	56.69%
Proposed model	LSTM	Multimodal late-fusion	Facial expressions, EDA, Heart Rate, Temperature	Children	Ekmans emotions	83.57%

What can be inferred from the figure above is that the performance of the models varies significantly across different papers. Model in which participants are children tend to obtain higher performance. The multimodal approach, including the proposed model, is doing quite well in the overall comparison. Further, the model using an uncovential modality of Heart Rate achieves poor performance, underscoring the need for adding facial expression to emotion recognition models.

Overall, despite two unimodal approaches being the most successful ones, the proposed model shows qualities to become useful in the future. Further examination is recommended.

*7.3. **SUMMARY*** (ADAM SOBCZUK)

The comparative analysis presented in this chapter underscores the distinct strengths and limitations of the proposed models, Model I and Model II, relative to FaceReader software and state-of-the-art approaches in the literature. Model I demonstrated a significant improvement in accuracy, achieving 83.57% compared to FaceReader's 53.69%, particularly excelling in recognizing *Happy*, *Sad*, and *Angry* emotions. However, it exhibited increased false positives and false negatives in specific emotional categories such as *Fear* and *Disgust*.

Model II, employing a similarity-based metric, revealed robust performance in capturing *Neutral*, and *Happy* expressions, where it significantly outperformed FaceReader. Additionally, Model II achieved competitive results for *Scared* emotions, although it lagged behind in accurately detecting *Disgusted*, *Surprised*, *Sad* and *Angry*.

When benchmarked against the literature, the proposed models highlighted the potential advantages of multimodal emotion recognition frameworks. While unimodal models leveraging advanced architectures like CNNs occasionally outperformed the proposed models, the integration of biosignals and facial embeddings demonstrated promise in achieving a more comprehensive emotional understanding.

In conclusion, this chapter illustrates the effectiveness of the proposed models in advancing emotion recognition for nuanced datasets while identifying key areas for future enhancement, particularly in addressing class imbalances and improving performance for underrepresented emotions.

8. SUMMARY (ADAM SOBCZUK)

This chapter synthesizes the key findings, contributions, and implications of the thesis, providing a cohesive overview of the research conducted. The work focused on advancing the field of emotion recognition through a multimodal approach, integrating visual and physiological data to enhance the precision and applicability of machine learning models. By leveraging data from the EMBOA project, the research addressed challenges unique to neurodiverse populations and provided innovative solutions for dataset preparation, model design, and evaluation.

The chapter begins with a concise recap of the research objectives, followed by a detailed summary of the main findings, highlighting the strengths and limitations of the developed methods. It then discusses the theoretical, practical, and methodological contributions made by the research, emphasizing its impact on the field of affective computing. Finally, the chapter reflects on the broader implications of the study, proposing future directions to expand on the foundation established by this thesis. Through this synthesis, the chapter underscores the significance of the research in advancing emotion recognition technologies and fostering inclusive, adaptive human-computer interactions.

8.1. RECAP OF RESEARCH OBJECTIVES

This thesis addressed a critical challenge in advancing emotion recognition technologies-enhancing accuracy and applicability through a multimodal approach. The research integrated facial embeddings and physiological signals (EDA, HR, and TEMP) to improve the ability of machines to recognize and interpret human emotional states. By leveraging data from the EMBOA project, the study focused on developing and evaluating machine learning models suitable for neurodiverse contexts, such as robot-assisted interventions for children on the autism spectrum. This work is grounded in the broader field of affective computing, aiming to bridge human-computer interaction by enabling machines to exhibit empathetic responses.

8.2. SUMMARY OF KEY FINDINGS

This section encompasses summaries of dataset and model development, challenges encountered during implementations, as well as results and insights drawn.

8.2.1. DATASET DEVELOPMENT AND CHALLENGES

Multimodal Dataset Preparation: The thesis processed and integrated video data (facial embeddings) and biosignals, creating a dataset suitable for emotion recognition tasks. Significant effort was dedicated to resolving challenges such as:

- **Data Misalignment:** Synchronization of timestamps between modalities to ensure consistency.
- **Class Imbalance:** Neutral and happy states were overrepresented, necessitating undersampling and label weighting.
- **Face Detection Issues:** Problems arising from masks, distance from cameras, and multiple individuals in recordings were mitigated through advanced preprocessing techniques, such as bounding-box prioritization.

8.2.2. MODEL DEVELOPMENT AND IMPLEMENTATION

- Two approaches:
  - **Method I:** Focused on dominant emotion classification per timestamp using categorical labels.
  - **Method II:** Quantified emotional intensity across categories using continuous-valued distributions.
- Model Architecture:
  - Both methods employed bidirectional Long Short-Term Memory (LSTM) networks, with tailored dense layers and output configurations to align with task-specific objectives.
  - Method II utilized additional dropout layers to mitigate overfitting due to increased complexity.

8.2.3. RESULTS AND INSIGHTS

- **Method I Performance:**
  - Achieved high classification accuracy for dominant emotions.
  - Highlighted model strengths in detecting pronounced emotional states while revealing limitations in handling subtle variations.
- **Method II Performance:**
  - Demonstrated strong alignment with manually annotated data, achieving an 81.93% global similarity metric.
  - Outperformed Method I in capturing nuanced emotional expressions, benefiting from its regression-based evaluation framework.
- **Comparative Analysis:**
  - Benchmarked both models against the FaceReader software and existing literature, revealing comparable or superior performance in emotion recognition tasks.
  - Method II offered deeper insights into emotional states at the cost of increased computational complexity.

8.3. RESEARCH CONTRIBUTIONS

This section contains a description of the practical, and theoretical contributions of the proposed models, and solutions to data preprocessing problems. Possible applications of this thesis work are presented at the end of the section.

8.3.1. PRACTICAL CONTRIBUTIONS

- **Integration of Modalities:** Demonstrated the potential of combining visual and physiological data to enhance emotion recognition, especially in neurodiverse populations.
- **Data Processing Innovations:** Introduced robust methods for dataset alignment and handling imbalanced emotional distributions, laying a foundation for future multimodal datasets.

8.3.2. THEORETICAL CONTRIBUTIONS

- **Methodological Advances:** Highlighted the advantages of bidirectional LSTM architectures in sequential emotion recognition with insights into optimal configurations for multimodal inputs.
- **New Evaluation Metrics:** Utilized a validated custom similarity metric from Geisler et al. [2] for regression-based emotional intensity modeling, enriching the evaluation landscape.

8.3.3. APPLICATIONS

- **Assistive Technologies:** The study contributes directly to the development of robot-assisted therapies for children on the autism spectrum, enabling adaptive systems that respond to emotional cues.
- **Generalizable Insights:** Offers methodologies applicable to broader domains, including healthcare, education, and adaptive learning systems.

8.4. IMPLICATIONS OF THE RESEARCH

- **Broader Applications:** The models and methodologies can inform the design of systems in domains like security, entertainment, and human-computer interaction, where understanding emotional states is vital.
- **Addressing Neurodiversity:** By focusing on data from children on the autism spectrum, this research emphasizes the inclusivity and robustness of emotion recognition systems.

8.4.1. POSSIBLE FUTURE DIRECTIONS

1. New model:
    - Designing a different model based on the conclusions drawn from this thesis.
2. Additional Modalities:
    - Incorporating audio signals or text-based sentiment analysis to create even richer multimodal datasets.
3. Advanced Architectures:
    - Exploring state-of-the-art architectures like transformers to enhance performance on both classification and regression tasks.
4. Real-World Validation:
    - Deploying the models in real-world environments, particularly in therapeutic or educational settings, to asses practical efficacy.

8.5. CONCLUDING REMARKS

This thesis contributes meaningfully to the field of affective computing, particularly in advancing multimodal emotion recognition systems. By integrating visual and physiological data, addressing critical preprocessing challenges, and designing robust machine learning models, it establishes a foundation for empathetic human-computer interaction. The insights gained from this research not only highlight the potential of multimodal approaches but also underscore the importance of inclusive and adaptive designs in emotion recognition systems. Future work should

build on these foundations to refine methodologies, broaden applicability, and deepen the impact of these technologies on society.

BIBLIOGRAPHY

[1] *Emboa Project Website*. https://emboa.eu/. Accessed: 2024-12-08.

[2] A. Geisler et al. *Przygotowanie publikowalnego zbioru danych na podstawie danych zebranych w projekcie EMBOA*. 2023.

[3] M. Abadi et al. "Tensorflow: Large-scale machine learning on heterogeneous distributed systems". In: *arXiv preprint arXiv:1603.04467* (2016).

[4] *Scikit-learn Documentation*. https://scikit-learn.org/stable/. Accessed: 2024-12-08.

[5] D. L. Schomer et al. "20C2Cellular Substrates of Brain Rhythms". In: *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields*. Oxford University Press, Nov. 2017. ISBN: 9780190228484. DOI: 10.1093/med/9780190228484.003.0002. eprint: https://academic.oup.com/book/0/chapter/305245903/chapter-ag-pdf/44502193/book_35515_section_305245903.ag.pdf. URL: https://doi.org/10.1093/med/9780190228484.003.0002.

[6] A. Feather, D. Randall, and M. Waterhouse. *Kumar and Clark's Clinical Medicine E-Book: Kumar and Clark's Clinical Medicine E-Book*. Elsevier, 2020. ISBN: 9780702078705. URL: https://books.google.pl/books?id=sl3sDwAAQBAJ.

[7] Y. Wang et al. "A systematic review on affective computing: emotion models, databases, and recent advances". In: *Information Fusion* 83-84 (2022), pp. 19–52. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2022.03.009. URL: https://www.sciencedirect.com/science/article/pii/S1566253522000367.

[8] Y. Cai, X. Li, and J. Li. "Emotion Recognition Using Different Sensors, Emotion Models, Methods and Datasets: A Comprehensive Review". In: *Sensors* 23.5 (2023). ISSN: 1424-8220. DOI: 10.3390/s23052455. URL: https://www.mdpi.com/1424-8220/23/5/2455.

[9] B. Pan et al. "A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods". In: *Neurocomputing* 561 (2023), p. 126866. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2023.126866. URL: https://www.sciencedirect.com/science/article/pii/S092523122300989X.

[10] K. Ezzameli and H. Mahersia. "Emotion recognition from unimodal to multimodal analysis: A review". In: *Information Fusion* 99 (2023), p. 101847. ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2023.101847. URL: https://www.sciencedirect.com/science/article/pii/S156625352300163X.

[11] A. Sobczuk and O. Kołoszko. The repository containing the source code of this thesis proceedings. URL: https://github.com/Azirral/de-face-recognition.

[12] F. Schroff, D. Kalenichenko, and J. Philbin. "FaceNet: A Unified Embedding for Face Recognition and Clustering". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015.

[13] K. Zhang et al. "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks". In: *IEEE Signal Processing Letters* 23.10 (Oct. 2016), pp. 1499–1503. ISSN: 1558-2361. DOI: 10.1109/lsp.2016.2603342. URL: http://dx.doi.org/10.1109/LSP.2016.2603342.

[14] C. Szegedy et al. *Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning*. 2016. arXiv: 1602.07261 [cs.CV]. URL: https://arxiv.org/abs/1602.07261.

[15] K. H. e. a. Ganapathy N. Veeranki Y.R. "Emotion Recognition Using Electrodermal Activity Signals and Multiscale Deep Convolutional Neural Network." In: *J Med Syst 45, 49* (2021). URL: https://doi.org/10.1007/s10916-020-01676-6%7D.

[16] D. Yu and S. Sun. "A Systematic Exploration of Deep Neural Networks for EDA-Based Emotion Recognition". In: *Information* 11.4 (2020). ISSN: 2078-2489. DOI: 10.3390/info11040212. URL: https://www.mdpi.com/2078-2489/11/4/212.

[17] L. Shu et al. "Wearable Emotion Recognition Using Heart Rate Data from a Smart Bracelet". In: *Sensors* 20.3 (2020). ISSN: 1424-8220. DOI: 10.3390/s20030718. URL: https://www.mdpi.com/1424-8220/20/3/718.

[18] G. Cosoli et al. "Measurement of multimodal physiological signals for stimulation detection by wearable devices". In: *Measurement* 184 (2021), p. 109966. ISSN: 0263-2241. DOI: https://doi.org/10.1016/j.measurement.2021.109966. URL: https://www.sciencedirect.com/science/article/pii/S026322412100899X.

[19] X. Huang et al. "Multi-modal emotion analysis from facial expressions and electroencephalogram". In: *Computer Vision and Image Understanding* 147 (2016). Spontaneous Facial Behaviour Analysis, pp. 114–124. ISSN: 1077-3142. DOI: https://doi.org/10.1016/j.cviu.2015.09.015. URL: https://www.sciencedirect.com/science/article/pii/S1077314215002106.

[20] *TensorFlow API Documentation*. https://www.tensorflow.org/api_docs. Accessed: 2024-12-02.

[21] *TensorFlow Keras API Documentation*. https://www.tensorflow.org/guide/keras. Accessed: 2024-12-02.

[22] *Jupyter Documentation*. https://docs.jupyter.org/en/latest/. Accessed: 2024-12-02.

[23] *NumPy Documentation*. https://numpy.org/doc/stable/. Accessed: 2024-12-02.

[24] J. Brownlee. *Why One-Hot Encode Data in Machine Learning?* https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/. Accessed: 2024-12-02.

[25] M. Schuster and K. Paliwal. "Bidirectional recurrent neural networks". In: *Signal Processing, IEEE Transactions on* 45 (Dec. 1997), pp. 2673–2681. DOI: 10.1109/78.650093.

[26] P. Ramachandran, B. Zoph, and Q. V. Le. *Searching for Activation Functions*. 2018. URL: https://openreview.net/forum?id=SkBYYyZRZ.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Ed. by F. Pereira et al. Vol. 25. Curran Associates, Inc., 2012. URL: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

[28] J. Ren and H. Wang. "Chapter 3 - Calculus and optimization". In: *Mathematical Methods in Data Science*. Ed. by J. Ren and H. Wang. Elsevier, 2023, pp. 51–89. ISBN: 978-0-443-18679-0. DOI: https://doi.org/10.1016/B978-0-44-318679-0.00009-0. URL: https://www.sciencedirect.com/science/article/pii/B9780443186790000090.

[29] Y. Chen et al. “Chapter 2 - Fundamentals of neural networks”. In: *AI Computing Systems*. Ed. by Y. Chen et al. Morgan Kaufmann, 2024, pp. 17–51. ISBN: 978-0-323-95399-3. DOI: https://doi.org/10.1016/B978-0-32-395399-3.00008-1. URL: https://www.sciencedirect.com/science/article/pii/B9780323953993000081.

[30] D. P. Kingma and J. Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG]. URL: https://arxiv.org/abs/1412.6980.

[31] Z. Zhang and M. R. Sabuncu. *Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels*. 2018. arXiv: 1805.07836 [cs.LG]. URL: https://arxiv.org/abs/1805.07836.

[32] W. Strawderman. “Theory of Point Estimation by E. L. Lehmann; George Casella”. In: *Journal of the American Statistical Association* 95 (Mar. 2000). DOI: 10.2307/2669560.

[33] S. Visa et al. “Confusion Matrix-based Feature Selection.” In: vol. 710. Jan. 2011, pp. 120–127.

[34] M. Sokolova and G. Lapalme. “A systematic analysis of performance measures for classification tasks”. In: *Information Processing & Management* 45.4 (2009), pp. 427–437. ISSN: 0306-4573. DOI: https://doi.org/10.1016/j.ipm.2009.03.002. URL: https://www.sciencedirect.com/science/article/pii/S0306457309000259.

[35] D. Powers. “Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation”. In: *Mach. Learn. Technol.* 2 (Jan. 2008).

[36] D. Olson and D. Delen. *Advanced Data Mining Techniques*. Jan. 2008, p. 138. ISBN: 978-3-540-76916-3. DOI: 10.1007/978-3-540-76917-0.

[37] A. A. "Taha and A. Hanbury. “"Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool"”. In: *"BMC Med. Imaging"* 15.1 (2015). DOI: https://doi.org/10.1186/s12880-015-0068-x. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC4533825/.

[38] *A complete guide to classification metrics in machine learning*. https://www.evidentlyai.com/classification-metrics. Accessed: 2024-12-06.

[39] C. Willmott and K. Matsuura. “Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance”. In: *Climate Research* 30 (Dec. 2005), p. 79. DOI: 10.3354/cr030079.

[40] K. Tyagi et al. “Chapter 4 - Regression analysis”. In: *Artificial Intelligence and Machine Learning for EDGE Computing*. Ed. by R. Pandey et al. Academic Press, 2022, pp. 53–63. ISBN: 978-0-12-824054-0. DOI: https://doi.org/10.1016/B978-0-12-824054-0.00007-1. URL: https://www.sciencedirect.com/science/article/pii/B9780128240540000071.

[41] T. O. Hodson. “Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not”. In: *Geoscientific Model Development* 15.14 (2022), pp. 5481–5487. DOI: 10.5194/gmd-15-5481-2022. URL: https://gmd.copernicus.org/articles/15/5481/2022/.

[42] K. Ali, S. Shah, and C. E. Hughes. “In-the-Wild Affect Analysis of Children with ASD Using Heart Rate”. In: *Sensors* 23.14 (2023). ISSN: 1424-8220. DOI: 10.3390/s23146572. URL: https://www.mdpi.com/1424-8220/23/14/6572.

[43] F. M. Talaat et al. “Real Time Facial Emotion Recognition model Based on kernel Autoencoder and Convolutional Neural Network for Autism Childrens”. In: (Dec. 2022). DOI: 10.21203/rs.3.rs-2387030/v1. URL: https://link.springer.com/article/10.1007/s00500-023-09477-y#citeas.

[44] R. Fuentes-Alvarez et al. “Energetic optimization of an autonomous mobile socially assistive robot for autism spectrum disorder”. In: *Frontiers in Robotics and AI* 9 (2023). ISSN: 2296-9144. DOI: 10.3389/frobt.2022.1053115. URL: https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2022.1053115.

[45] A. J. Syed et al. “Expression Detection Of Autistic Children Using CNN Algorithm”. In: *2023 Global Conference on Wireless and Optical Technologies (GCWOT)*. 2023, pp. 1–5. DOI: 10.1109/GCWOT57803.2023.10064653.

[46] H. Abu-Nowar et al. “SENSES-ASD: a social-emotional nurturing and skill enhancement system for autism spectrum disorder”. In: *PeerJ Computer Science* 10 (2024), e1792.

[47] L. Jingjing et al. “Multimodal Emotion Recognition for Children with Autism Spectrum Disorder in Social Interaction”. In: *International Journal of Human–Computer Interaction* 40.8 (2024), pp. 1921–1930. DOI: 10.1080/10447318.2023.2232194. URL: https://doi.org/10.1080/10447318.2023.2232194.

[48] H. Arabian et al. “Emotion Recognition beyond Pixels: Leveraging Facial Point Landmark Meshes”. In: *Applied Sciences* 14.8 (2024). ISSN: 2076-3417. DOI: 10.3390/app14083358. URL: https://www.mdpi.com/2076-3417/14/8/3358.

[49] P. Rani. “Emotion Detection of Autistic Children Using Image Processing”. In: *2019 Fifth International Conference on Image Information Processing (ICIIP)*. 2019, pp. 532–535. DOI: 10.1109/ICIIP47207.2019.8985706. URL: https://ieeexplore.ieee.org/document/8985706.

[50] K. Smitha and A. Vinod. “Facial emotion recognition system for autistic children: a feasible study based on FPGA implementation”. In: *Medical and biological engineering and computing* 53 (Aug. 2015). DOI: 10.1007/s11517-015-1346-z.

LIST OF FIGURES

LIST OF TABLES