Multimodal Emotion Recognition Using Physiological Signals

by

Mehakdeep Kaur Dhothar

A thesis submitted to the University of Ottawa
in partial fulfillment of the requirements for the
Master of Applied Science
in
Electrical and Computer Engineering with Concentration in Applied Artificial Intelligence

School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

Declaration of Authorship

I hereby certify that this thesis is entirely my own original work, except where otherwise indicated. The practical implementation of signal quality assessment for ECG and RSP signals was conducted by Dr. Hitham Jleed. This section was instrumental and has been appropriately included in the thesis. I am aware of the University's regulations concerning plagiarism, including the associated disciplinary actions. Any use of the work of other authors, in any form, has been properly acknowledged at the point of use.

Abstract

Affective computing aims to develop systems capable of recognizing and interpreting human emotions, yet existing multimodal datasets frequently suffer from limitations such as poor signal quality, high inter-subject variability, and inconsistent evaluation protocols. To address these gaps, this thesis develops and validates a comprehensive framework for multimodal emotion recognition using physiological signals—Electrocardiogram (ECG), Electrodermal Activity (EDA), and Respiration (RSP)—augmented with speech-based representations. The goal was to establish standardized preprocessing workflows, rigorous signal quality assessment (SQA), and reproducible baseline experiments to support the development and technical validation of a large-scale physiological dataset. This framework was applied to a dataset collected from 99 participants, containing synchronized physiological recordings, speech responses, and self-reported emotional annotations during exposure to validated video stimuli.

To ensure data integrity, a rigorous SQA and artifact-removal pipeline was applied across modalities, integrating established ECG and respiration metrics with newly designed EDA-specific indicators. Using this refined dataset, multiple emotion-classification experiments were conducted under a strict subject-independent evaluation protocol, comparing fixed 30-second windows with emotion-triggered temporal segments. Across all tasks—binary arousal, binary valence, and multiclass emotion recognition—trigger-based segments consistently produced clearer and more discriminative physiological patterns. Random Forest achieved the strongest overall performance, including 78.8% multiclass accuracy using physiological features alone.

To explore multimodal enhancement, speech embeddings were fused with handcrafted physiological features. This early-fusion approach led to substantial improvements across all tasks, most notably increasing multiclass accuracy from 78.8% to 97% when using trigger-based segments. These findings demonstrate that speech provides complementary affective information that enhances physiological representations. A subject-wise evaluation was also conducted to examine emotion separability across individuals and to identify video-specific misclassification patterns that reveal how different stimuli elicit varying physiological responses. Overall, this thesis delivers a validated multimodal dataset, reproducible processing pipelines, and strong baseline benchmarks that provide a solid foundation for future research in physiological and multimodal emotion recognition. [1]

[1] Text shown in blue throughout this thesis indicates hyperlinks to external or internal sources.

Acknowledgements

I would like to express my deepest gratitude to my supervisor, Professor Miodrag Bolic, for his unwavering support, guidance, and belief in my potential throughout every stage of my academic journey. His thoughtful mentorship, patience, and genuine willingness to help have been instrumental in my progress. I feel truly fortunate to have had the opportunity to learn from him.

I would also like to thank Dr. Hitham Jleed, who was involved in the initial phase of the project as a Research Associate. His valuable advice during the early stages helped shape the direction of this work and gave me the confidence to move forward.

My sincere thanks go to the members of the Psychology team, Vincent Francoeur and Professor Sylvain Gagnon, whose collaboration and support were vital to the success of this project. The regular biweekly meetings were especially helpful in overcoming challenges and maintaining steady progress.

I am also grateful to Orbmedic, Ottawa, Canada, the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Ontario Centres of Innovation (OCI) for their financial support throughout this project. I sincerely appreciate their contribution to making this work possible.

Finally, I am profoundly grateful to my family, whose constant encouragement, understanding, and emotional support have sustained me throughout this journey. Their belief in me has been the foundation of my strength and perseverance.

Table of Contents

List of Tables

List of Figures

Abbreviations

ANS Autonomic Nervous System 2

basSQI Relative Power in the Baseline 19, 21

BW Baseline Wander 16

ECG Electrocardiogram iii, 2, 5, 7, 16, 71

EDA Electrodermal Activity iii, 2, 5, 7, 71

EEG Electroencephalogram 7

EMA Electrode Motion Artifacts 16, 17

EMG Electromyography 7

ER-SCR Event-Related Skin Conductance Response 15

GLU Gated Linear Units 49, 77

HCI Human Computer Interaction 1

kSQI Kurtosis 18, 21

LOSOCV Leave-One-Subject-Out Cross Validation 11

MA Muscle Activity 16, 17

MFCC Mel-frequency Cepstral Coefficients 45

ML Machine Learning 17

NS-SCR Non-Specific Skin Conductance Response 15

PLI Powerline Interference 16

pSQI Relative Power of QRS Complex 19

RSP Respiration iii, 2, 5, 7, 8, 71

SCL Tonic Skin Conductance Level 8

SCR Skin Conductance Response 8

skSQI Skewness 18

SNR Signal-to-Noise Ratio 18

SNS Sympathetic Nervous System 15

SQA Signal Quality Assessment 14, 16, 21, 38, 78

SQIs Signal Quality Indices 17, 21

SSQIs Statistical Signal Quality Indicators 17, 18

SVM Support Vector Machine 11

UMAP Uniform Manifold Approximation and Projection 56

Chapter 1

Introduction

1.1 Background and Motivation

Emotions play a fundamental role in shaping human behavior, influencing how we think, act, and interact with others. They are fundamental to human experience and serve as a critical communication channel in social interactions, often conveying more than words alone. As a result, the ability to recognize and interpret emotional states has become central to many fields such as psychology, neuroscience, and, more recently, computer science. The desire to inculcate machines with the ability to understand and respond to human emotional states has given rise to the field of affective computing. The applications of such technology are vast and transformative, ranging from enhanced Human Computer Interaction (HCI) and intelligent user interfaces to advanced driver-assistance systems, personalized healthcare, and mental health monitoring.

The literature has vastly relied on using facial expressions and speech patterns to detect emotional cues [1–3]. While these modalities have proven effective in many contexts, they are not without limitations. Facial expressions and tone of voice can be consciously controlled or suppressed, making them potentially unreliable indicators of an individual's true emotional state [4]. This limitation becomes particularly critical in medical or clinical settings, where an individual may consciously mask their emotions due to stress, trauma, or discomfort. For instance, during a medical examination, a patient experiencing anxiety or distress might outwardly appear calm. In such situations, understanding the patient's true emotional state can be vital for accurate diagnosis and care, making physiological signals a more dependable alternative to observable behavioral cues.

This has led researchers to explore alternative, more objective sources of emotional information. Physiological signals, which are regulated by the Autonomic Nervous System (ANS), offer a promising alternative [5]. The ANS controls involuntary bodily functions such as heart rate, respiration, and perspiration. When an individual experiences an emotion, the ANS triggers a cascade of physiological changes that are largely beyond conscious control. These changes can be captured and analyzed to provide a more direct and unadulterated window into a person's affective state [6].

Commonly used physiological signals in emotion recognition include the ECG, which measures the heart's electrical activity; EDA, which reflects changes in skin conductance due to sweat gland activity; and RSP, which captures breathing patterns. The advent of wearable sensors and devices has made the collection of these signals increasingly non-invasive and suitable for real-world applications. Wearable devices such as smartwatches and rings (e.g., Apple Watch, Oura Ring) now routinely measure heart rate, skin conductance, and other physiological metrics. This accessibility enables large-scale emotion recognition and monitoring without the need for specialized laboratory setups. In contrast, using modalities like facial expressions or speech would require continuous camera or microphone access, which is often impractical in everyday scenarios and in contexts where privacy is a critical concern.

This technological advancement opens up exciting possibilities for continuous and unobtrusive emotion monitoring, paving the way for systems that can adapt and respond to users' emotional states in real-time.

1.2 Problem Statement

Despite the significant potential of physiological signals for emotion recognition, several challenges hinder the development of robust and generalizable models. One of the primary obstacles is the inherent variability of physiological signals, both within and between individuals. A person's baseline physiological state can fluctuate due to a variety of factors, including time of day, physical activity, and overall health. Moreover, different individuals can exhibit markedly different physiological responses to the same emotional stimulus, a phenomenon known as inter-subject variability [7, 8].

Another significant challenge is the susceptibility of physiological recordings to noise and artifacts. Motion artifacts, sensor displacement, and environmental interference can corrupt the signals, making it difficult to extract meaningful emotional patterns. If not properly addressed, these artifacts can lead to inaccurate feature extraction and, consequently, poor model performance [9–11].

Furthermore, the complexity of emotional experiences themselves presents a challenge. Emotions are not always discrete and well-defined; they can be subtle, mixed, and highly context-dependent [12]. One person may feel anger while watching a video stimulus, while another person might not feel angry at all, and thus emotions are also linked to human's past experiences. Accurately labeling emotional states for the purpose of training supervised machine learning models is therefore a non-trivial task that often relies on subjective self-reports.

Finally, while many studies have demonstrated the feasibility of emotion recognition from physiological signals in controlled laboratory settings, the generalizability of these models to real-world scenarios remains a key concern. The transition from the lab to everyday life introduces a host of new variables and complexities that can significantly impact the performance and reliability of emotion recognition systems.

1.3 Contributions

This thesis aims to address some of the aforementioned challenges by making the following key contributions to the field of multimodal emotion recognition using physiological signals:

1. **Curation of a signal quality assessment and artifact removal pipeline:** A significant portion of this work is dedicated to developing and validating a pipeline for assessing the quality of ECG, EDA, and RSP signals and for removing artifacts. This ensures that the data used in this thesis for model training and evaluation is of high quality, thereby enhancing the reliability of the results.

2. **Systematic evaluation of classical machine learning and deep learning approaches:** We conduct a thorough investigation into the performance of both traditional machine learning models and deep learning methods with handcrafted features for emotion recognition. This comparative analysis provides valuable insights into the strengths and weaknesses of different modeling approaches for classification tasks.

3. **Investigation into the impact of speech features:** To further enhance the accuracy and robustness of our emotion recognition models, we investigate the benefits of incorporating speech features. We explore the fusion of physiological data with linguistic and paralinguistic features extracted from speech, demonstrating the potential of a truly multimodal approach that combines both physiological and behavioral cues.

4. **Presenting a curated, annotated multimodal dataset:** This research is based on a dataset collected in collaboration with the Department of Psychology at the University of Ottawa. While the data collection was conducted by members of the Psychology department, I was responsible for organizing the data, performing signal cleaning, and conducting signal quality assessment. The dataset includes synchronized physiological signals (ECG, EDA, and RSP), emotional self-reports, and speech recordings from 99 participants. The experimental protocol was carefully designed to elicit authentic emotional responses using validated video stimuli. This comprehensive dataset forms the foundation of the analyses presented in this thesis and constitutes a valuable new resource for the affective computing research community.

1.4 Organization of Thesis

The remainder of this thesis is organized as follows:

- **Chapter 2: Background and Related Works** provides a comprehensive review of the existing literature on emotion models, emotion recognition using physiological signals, and current state-of-the-art methods in the field.
- **Chapter 3: Signal Quality Assessment and Artifacts Removal Methods** details the methodologies developed and employed for assessing the quality of the physiological signals and for removing noise and artifacts.
- **Chapter 4: Classification of Emotions based on Handcrafted Features** describes the extraction of handcrafted features from the physiological signals and the evaluation of various classical machine learning models for emotion classification.
- **Chapter 5: Conclusion and Future Works** concludes the thesis with a summary of the main contributions and provides recommendations for future research directions in the field of multimodal emotion recognition.

Chapter 2

Background and Related Works

Emotion recognition using physiological signals is trending area of research in affective computing, human-computer interaction (HCI), and mental health monitoring. There are different ways of emotion recognition – visual, speech, and physiological. Unlike visual or speech-based emotion recognition, physiological signal-based methods capture the internal, involuntary responses of the body, offering a more objective insight into a person's emotional state. Commonly used physiological signals include EDA, ECG, RSP, and skin temperature. These signals are typically collected using wearable devices, making the process non-invasive and suitable for continuous monitoring [5].

This chapter provides a comprehensive overview of the literature relevant to Emotion Recognition (ER) systems. It begins with a discussion of prominent emotion representation models used in current research, followed by an examination of the physiological signals commonly employed for emotion inference. The chapter concludes by highlighting the specific technical methodologies that form the foundation of this thesis.

2.1 Understanding Emotions

In psychology, emotions are often described as subjective experiences that initiate psychological shifts, which subsequently result in observable physical changes influencing an individual's behavior and cognition during that emotional state [13]. Emotions are also defined as complex psychological and physiological responses triggered by internal or external stimuli [14] [15]. They influence not only mental well-being but also physical health and behavioral patterns. Positive emotions can enhance focus, motivation, and productivity, whereas negative emotions may contribute to stress-related illnesses or diminished

cognitive performance [16]. Due to their central role in human experience, emotions have become a major focus across disciplines such as neuroscience, psychology, and computer science. Researchers continue to explore how emotional states are formed, expressed, and recognized through various observable and measurable indicators.

2.1.1 Emotions Models

In the existing literature, emotions have been represented through a variety of models [3, 17–20]. These models can be categorized into two major types - categorical, and dimensional. The categorical model makes use of six basic emotion classes namely - anger, disgust, fear, joy, sadness, and surprise [3] or domain-specific expressive classes such as boredom, confusion [18]. In this model, emotions are represented by a single label that represents human emotions and thus are easy to understand. However, they may oversimplify emotional complexity by grouping diverse emotions under limited categories. Moreover, cultural, linguistic, and individual differences can influence how emotions are expressed and interpreted, making consistent labeling across populations challenging [21].

In dimensional model, the emotions are represented by a point in a continuous multidimensional space [17]. The most widely adopted model is Russell's Circumplex Model, which maps emotions onto two primary dimensions: valence, indicating the degree of pleasantness (positive vs. negative), and arousal, representing the intensity of activation (high vs. low) [22]. The extensions of this model include a third dimension — dominance — which captures the level of control or influence an individual feels over their emotional state [23]. Plutchik's Wheel of Emotions [6] arranges eight primary emotions—joy, trust, fear, surprise, sadness, disgust, anger, and anticipation—around a color wheel structure.

2.2 Emotion Recognition using Physiological Signals

Physiological signals are a widely used user-input for the task of emotion recognition. Emotion recognition using physiological signals involves analyzing the body's involuntary responses, such as heart activity, skin conductance, and breathing patterns, to infer emotional states. These signals are indicators of the autonomic nervous system's response to emotional stimuli, making them valuable for affective computing [24]. Since these physiological responses are controlled by the autonomic nervous system, they are largely involuntary, and cannot be consciously manipulated. And therefore, they are an almost true indicator of a person's emotional state [4].

The physiological signals that are typically used for emotion classification are ECG, Electroencephalogram (EEG), EDA, Electromyography (EMG), RSP, and others [16]. Among these, ECG and EDA are especially popular, other than EEG, when classifying emotions by arousal and valence levels. This is mainly because ECG and EDA are easier to record and more comfortable for participants compared to EEG. The following sections explain each of these signal types in more detail.

2.2.1 Electrocardiogram

ECG is a fundamental physiological signal that captures the electrical activity of the heart over time [25]. Each heartbeat in an ECG signal is characterized by distinct waveforms: the P wave, the QRS complex, and the T wave, each corresponding to different phases of the cardiac cycle, such as atrial depolarization, ventricular depolarization, and ventricular repolarization, respectively [25, 26]. These waveforms carry significant information about an individual's cardiac state. Research has established strong correlations between cardiovascular activity and emotional states, as emotions directly influence the autonomic nervous system, which in turn regulates heart rhythms [27, 28]. This link makes ECG a reliable source of information for emotion recognition systems within affective computing.

Extracting useful information from heartbeats for affective computing has involved various feature extraction techniques. Common approaches include deriving time-domain and frequency-domain features, often after detecting fiducial points like P, Q, R, S, and T components from the ECG waveform [28]. Specifically, the R-peak within the QRS complex is of particular interest, as it is the most prominent component, enabling the calculation of RR-intervals and Heart Rate Variability (HRV), both valuable features for affective computing [27, 29].

2.2.2 Electrodermal Activity

The Electrodermal Activity (EDA), also known as galvanic skin conductance (GSC) or galvanic skin response (GSR), is a physiological signal that reflects changes in the skin conductance due to sweat gland activity, modulated by the sympathetic nervous system [30]. It is sensitive to emotional arousal and has been shown to effectively distinguish between high-arousal and low-arousal emotional states, making it a primary modality in emotion recognition systems. An aroused state leads to the production of more sweat, and sweat is a weak electrolyte making it a good conductor [31]. When sweating increases, the sweat ducts fill, creating many low-resistance pathways, and increasing the conductivity.

EDA is typically measured by placing two electrodes on the skin, often on the palmar surface of the hands or plantar surface of the feet, which are rich in eccrine sweat glands. The electrical conductance of the skin is then recorded, usually in microSiemens (μS) [32, 33].

There are two main components of EDA signal: the Tonic Skin Conductance Level (SCL), which reflects gradual and sustained shifts occurring over periods of several seconds to minutes, and skin conductance responses (SCRs), which are rapid, transient changes in skin conductance superimposed on the tonic level that occur in response to specific stimuli or events, often characterized by their amplitude, rise time, and recovery time [30, 32]. By convention, a Skin Conductance Response (SCR) is defined as a rapid increase in skin conductance of at least 0.05 μS within approximately 1–3 seconds, followed by a slower return to the baseline over the subsequent 3–15 seconds [34–36]. Features are then extracted from both EDA components for emotion classification. Table 4.2 provides a complete list of all features derived from these components in this work.

2.2.3 Respiration

Respiration involves the intake of oxygen-rich air into the lungs (inhalation) and the expulsion of carbon dioxide-rich air from the lungs (exhalation). This exchange of gases occurs through the respiratory system, which includes the nose, trachea, bronchi, and lungs [37]. The respiration signal (RSP) reflects an individual's breathing patterns, closely linked to the emotional states. Faster, shallow breathing may be indicative of anxiety or fear, while slower, deeper breathing is typically associated with calm or relaxed states [38,39]. Usually, respiration belts (thoracic and abdominal bands that measure changes in circumference) are used to record the respiration rates [40]. Traditional analysis of respiration signals involves extracting various features that characterize breathing patterns. These include respiratory rate (breaths per minute), breath amplitude, inspiratory and expiratory durations, and the inspiratory-to-expiratory duration ratio (I/E ratio).

2.3 Related Works

2.3.1 Existing Datasets

Over recent years, several publicly available multimodal emotion recognition datasets have been introduced, significantly advancing research in affective computing. In Table 2.1, a

Table 2.1: Comparison of Existing Emotion Recognition Datasets and Signal Quality Assessment (SQA)

Dataset	Year	Number of Participants	Signals Collected	Stimuli Used	Annotation Type	SQA Performed
AMIGOS [41]	2017	40	ECG, EEG, EDA	Videos (short clips), personality questionnaires	Self-reported (Valence, Arousal, Personality, Mood)	✗
DEAP [42]	2012	32	EEG, EDA, EMG, Respiration	Music videos (60 seconds)	Self-reported (Valence, Arousal, Liking, Dominance) SAM scale[1, 9]	✗
MAHNOB-HCI [43]	2012	27	EEG, ECG, EDA, Respiration	Emotional video clips	Self-reported, Expert Ratings	✗
ASCERTAIN [44]	2016	58	ECG, EEG, EDA	Movie trailers (emotional content)	Self-reported (Valence [-3, 3], Arousal [0, 6])	✗
RECOLA [45]	2013	46	ECG, EDA, Audio, Video	Dyadic video chat (collaborative task)	Continuous Expert Ratings (Valence, Arousal)	✗
WESAD [46]	2018	15	ECG, EDA, EMG, Respiration, ACC	Stress induction protocol (baseline, amusement, stress)	Self-reported (Binary - stress, non-stress, Three-class - baseline, stress, amusement)	✗
CASE [47]	2021	30	ECG, EDA, PPG, ACC, Respiration	Video clips	Continuous self-reported (Valence, Arousal) using joystick	✗
Our Dataset	2025	99	ECG, EDA, Respiration, Speech	Neutral and emotional videos (6 emotions × 2 repetitions)	Self-reported (Valence, Arousal, 6 Discrete Emotions)	✓

✓: *Signal quality assessment performed;* ✗: *No explicit SQA reported.*

comprehensive comparison of some of these datasets have been provided. The AMIGOS dataset [41] contains data from 40 participants exposed to 16 short emotional video clips in both individual and group settings. It includes EEG (14 channels, 128 Hz), ECG, and GSR signals (both at 1000 Hz). Participants reported valence and arousal ratings using a 9-point Likert scale, a commonly used psychometric scale in affective studies that allows individuals to quantify the intensity of their emotional state (1 being the lowest, and 9 the highest). Additional metadata such as Big Five personality traits (a standardized

assessment measuring five major dimensions of personality: openness, conscientiousness, extraversion, agreeableness, and neuroticism [48]) and PANAS mood scores (Positive and Negative Affect Schedule, which quantifies current positive and negative emotional states) were also collected. However, explicit physiological SQA was not performed. The DEAP dataset [42] features physiological recordings from 32 participants watching 40 one-minute music videos selected to evoke specific affective responses. Signals include EEG (32 channels), EDA (GSR), EMG, and respiration, initially sampled at 512 Hz and downsampled to 128 Hz (to reduce processing time). Participants provided ratings on valence, arousal, liking, and dominance using a 9-point Self-Assessment Manikin (SAM) scale. Despite basic artifact removal and filtering, no explicit physiological SQA is reported. The MAHNOB-HCI dataset [43] comprises multimodal recordings from 27 participants viewing 20 emotional videos, each lasting 35 to 117 seconds. The dataset includes EEG (32 channels), ECG, EDA (GSR), and respiration signals sampled at 256 Hz, along with synchronized facial video recordings. Participants rated their emotional experiences using SAM scales for valence, arousal, and dominance, supplemented by external annotations categorizing emotions into six discrete labels. Explicit physiological SQA is not documented, potentially limiting studies that require artifact-free signals. The ASCERTAIN dataset [44] includes physiological signals (ECG at 250 Hz, EDA at 100 Hz, and EEG at 32 Hz) collected from 58 participants watching 36 emotion-inducing movie trailers. Participants self-reported their emotional states on valence scales (-3 to +3) and arousal scales (0 to 6). It additionally contains personality and affect questionnaire scores for comprehensive affective and trait analyses. However, the dataset lacks explicit physiological SQA. The RECOLA dataset [45] was developed for multimodal affect recognition during naturalistic dyadic interactions. It includes ECG, EDA, audio, and video data from 46 participants engaged in collaborative tasks. Valence and arousal were annotated continuously by experts, enabling detailed temporal affective analysis. Basic filtering and normalization steps were applied, but explicit SQA was not conducted, and artifact-affected segments were not flagged. The CASE dataset [47] captures eight physiological signals—ECG, BVP, GSR (EDA), respiration, skin temperature, and EMG from three muscle groups—all sampled at 1000 Hz and recorded from 30 participants while they watched emotional videos. Continuous valence-arousal annotations were recorded using a joystick interface. While this study focuses on validating physiological and annotation alignment, explicit physiological SQA was not reported, assuming signal reliability based purely on consistency with annotations. The WESAD dataset [46] is focused on stress and affect detection, capturing physiological signals using wearable devices (RespiBAN and Empatica E4) from 15 participants undergoing three experimental protocols - baseline, stress, and amusement. Physiological signals include ECG, EDA, EMG, respiration, body temperature, and accelerometer data sampled between 4–700 Hz. Despite comprehensive multimodal data collection, explicit

physiological SQA is absent.

In contrast, our dataset significantly improves upon existing datasets by offering a larger and more extensive participant pool of 99 individuals. Additionally, it includes explicit, detailed procedures for physiological signal quality assessment, enhancing the integrity and reliability of the data. The methodology for signal cleaning and artifact removal is described in detail in Section 3, positioning the dataset as a valuable resource for multimodal emotion recognition research.

2.3.2 Classical Machine Learning, and Deep Learning Approaches

Hsu et al. [49] explored categorical emotion recognition (joy, tension, sadness, peace), and dimensional classification (valence, and arousal) using Support Vector Machine (SVM) classifier on features extracted from ECG signals collected during music-listening sessions. They tested their approach using Leave-One-Subject-Out Cross Validation (LOSOCV) cross-validation achieving around 82% accuracy for valence, and around 72% for valence, and 61% for categorical classification.

Gjoreski et al. [7] applied classical machine learning techniques to classify arousal using physiological data from three publicly available datasets: AMIGOS, ASCERTAIN, and MAHNOB-HCI. They focused on manually extracting time-domain features from both ECG and EDA signals. Seven classical ML models were evaluated on each dataset separately to determine the optimal classifier for each signal type. In the AMIGOS dataset, ECG features performed best with a k-nearest neighbors (k-NN) classifier, achieving an accuracy of 53%, while EDA features achieved their highest performance using an AdaBoost classifier with decision trees. For the ASCERTAIN dataset, a support vector machine (SVM) yielded the best results for both ECG and EDA, achieving 66% accuracy. In the MAHNOB-HCI dataset, ECG-based classification was most accurate with SVM (62%), while EDA features were best classified using a Naive Bayes model, also with 62% accuracy. They used both intra-dataset and inter-dataset (cross-dataset) evaluation methods, and the cross-dataset validation revealed a noticeable drop in performance.

Anderson et al. [50] implemented the multimodal fusion on EDA, ECG, EEG, electrooculogram (EOG), and photoplethysmogram (PPG). After manually extracting 98 features covering both time and frequency domains, they used wrapper-based sequential forward selection and used 21 features to train a arousal-based classification model using SVM. They employed leave-one-subject-out (LOSO) cross-validation, allowing each participant to serve once as the test set while the remaining participants formed the training set. The highest performance, 89%, was observed when all five modalities were combined.

Feng et al. [51] proposed an emotion classification method that relies solely on wavelet features extracted from the EDA signal. They used wavelet packet decomposition approach to extract time-frequency features from the EDA signal, capturing both transient and sustained physiological responses. Their findings suggest that wavelet-based EDA features can yield satisfactory classification performance for affect recognition.

Kawde et al. [52] used Deep Belief Networks (DBNs) to classify emotions across arousal, valence, and dominance dimensions based on EDA, EEG, EMG, and EOG. A Deep Belief Network (DBN) is composed of multiple layers of Restricted Boltzmann Machines (RBMs) stacked one on top of the other. Each RBM is a two-layer neural network that can be used to learn representations (hidden patterns or features) from the signals. The output from the last RBM was fed into a softmax classifier to classify in two-class, and three-class settings for each emotional dimension. The reported accuracies were 78.28% for valence, 70.33% for arousal, and 70.16% for dominance. Although promising, the study did not elaborate on whether subject-independent evaluation strategies were used.

2.3.3 Limitations in the Current Literature

Emotion recognition using physiological signals has made considerable progress, particularly with the increasing availability of wearable devices capable of capturing data in real-world settings. This opens exciting opportunities for deploying emotion-aware systems beyond laboratory environments. However, current approaches still face several limitations—such as limited generalizability to new users, weak or inconsistent evaluation protocols, small and heterogeneous datasets, and the absence of proper signal quality assessment—that can affect the validity and interpretability of reported results. These limitations are discussed in the remainder of this section.

One important challenge is the generalizability of trained models to unseen data. Many existing works report performance based on cross-validation strategies like k-fold. While these methods help estimate model performance, they often rely on validation accuracy within a closed dataset. Consequently, the ability of these models to generalize to entirely new users or environments remains uncertain, particularly when exposed to previously unseen signal patterns.

Related to this is the concern of evaluation protocol design. Although some studies adopt subject-independent validation strategies, others use intra-subject splits or random data divisions that fail to ensure participant-level separation between training and testing sets. This introduces the risk of data leakage, where overlapping segments from the same individual are inadvertently used for both model development and validation. Such leakage

can result in inflated performance metrics that do not reflect true model robustness. In contrast, the classification tasks in this thesis are evaluated using strictly subject-independent strategies—ensuring that each participant's data is either entirely in the training or testing fold—which provides a more realistic estimate of model generalizability in practical applications.

The limitations in dataset size and composition also need consideration. Many widely used datasets include recordings from a relatively small number of participants—typically fewer than 30—making it difficult to account for inter-subject variability in physiological responses. Furthermore, heterogeneity in emotional stimuli, elicitation protocols, and annotation schemes across datasets introduces additional inconsistencies. By comparison, the dataset used in this thesis comprises 99 participants, offering a considerably larger and more diverse sample. This enhances the ability to reliably detect meaningful patterns and differences among emotions, improves generalization, and enables more rigorous cross-participant analysis.

Another shortcoming in the current literature is the lack of proper signal quality assessment. Most studies apply basic preprocessing steps like filtering, but few check whether parts of the signal are too noisy or affected by artifacts like motion or poor sensor contact. These bad-quality segments can reduce the accuracy of the models. In this thesis, we developed a detailed signal quality assessment process for ECG, EDA, and respiration signals, which allowed us to detect and remove low-quality data before feature extraction and classification (see Section 3).

Finally, many studies only perform basic classification tasks and do not explore how different factors—like signal types, data segments, or fusion methods—impact performance. In contrast, this thesis presents a series of experiments in Chapter 4 that cover different segment types, baseline correction methods, fusion of physiological and speech signals, and classification of specific emotions. These experiments provide a more complete picture of how different approaches affect model performance and help us understand the practical strengths and weaknesses of each method.

In summary, while existing studies have contributed important insights, they are often limited by small datasets, weak evaluation methods, and lack of signal quality control. This thesis addresses these issues by introducing a larger and cleaner dataset, applying subject-independent evaluation protocols, and conducting extensive experiments to better understand the dynamics of emotion recognition using physiological data.

Chapter 3

Signal Quality Assessment and Artifacts Removal Methods

Before using physiological signals for emotions classification, it is important to ensure that the data is of good quality. Physiological recordings are often susceptible to noise, motion artifacts, sensor disconnections, and environmental interference. If not properly handled, these artifacts can distort relevant emotional patterns and degrade model performance. This chapter outlines methods for assessing signal quality and removing artifacts from the electrocardiogram (ECG), electrodermal activity (EDA), and respiration (RSP) signals used in this thesis.

3.1 Background

Emotion recognition models rely on subtle physiological variations linked to emotional states. These changes are often small in magnitude and can be masked by noise or artifacts. Thus, maintaining high-quality signals is essential for enabling accurate feature representation and effective emotion classification.

Signal Quality Assessment (SQA) involves quantifying the reliability and usability of physiological signals. Good-quality signals typically show characteristic physiological waveforms, low baseline drift, and minimal motion artifacts.

3.1.1 Signal Quality Assessment of Electrodermal Activity

As discussed in the Section 2.2.2, EDA measures the fluctuations in the electrical conductance of the skin that result from Sympathetic Nervous System (SNS) activity, and has two components: the skin conductance level, and the Skin Conductance responses. SCRs can be further classified into Non-Specific Skin Conductance Response (NS-SCR), which occur spontaneously without known stimulus, and Event-Related Skin Conductance Response (ER-SCR), which is triggered by a specific external event [30, 53]. An ER-SCR typically represents a physiological reaction to a stimulus that provokes arousal or startle, reflecting the body's orienting or 'fight-or-flight' response. The latency between stimulus onset and the SCR peak is usually around three seconds, with some variability across individuals and settings. [5, 54].

The NS-SCR peaks occur periodically without any external stimuli. Elevated levels of arousal and stress can lead to an increased frequency of these NS-SCRs [55, 56]. However, artifacts caused by movements [57], electrode shifts, or change in breathing patterns [58,59] have similar characteristics as SCRs and can make it difficult to distinguish physiological responses from noise. This overlap between authentic signals and highlights the need for careful artifact detection and correction.

In their work on automated quality assessment procedures for EDA data, Kleckner et al. [36] described a set of four simple rules to identify invalid data segments, that we adopt in our work. These rules are based on the common artifacts that are seen in EDA signals.

1. **EDA is out of range** (0.05 μS - 45 μS): Data points where the EDA signal falls outside a physiologically plausible and device-specific valid range are considered invalid. This can occur when an electrode loses contact with the skin (e.g., EDA approaches 0), and "ceiling" artifacts, which happen when the measurement circuit becomes overloaded due to excessively high conductance. The minimum 0.05 μ is the accepted lowest detectable value for EDA [60]. The upper threshold is specific to the Mindware device used to record the data used in this work (refer sec: 3.2), as recommended by the manufacturer [53].

2. **Rapid changes/spikes in the signal**: Data points where the rate of change (slope) of the EDA signal exceeds a predefined threshold are deemed invalid. This is to prevent high-frequency or "jump" artifacts, which often result from sudden movements or pressure on the electrodes rather than genuine physiological changes.

3. **Transitional data surrounding invalid portions**: Data points immediately surrounding segments identified as invalid by Rules 1 or 2 are also invalidated to account

for "transition effects" or periods of instability that often occur directly before or after a clear artifact. These transitions might not meet the criteria of the first three rules, but are still compromised by adjacent invalid data. We invalidate the data within a 3-second window (both before and after) for any segment identified as invalid by the preceding rules. This helps to ensure that any potentially compromised data surrounding a clear artifact is also excluded.

Kleckner et al. [36] also defined a temperature-based rule to detect improper sensor contact. If the sensor temperature is too low (such as ambient room temperature) or too high (for example, direct sunlight exposure), this suggests that the device may not be in stable contact with the skin. However, in this study, the temperature recordings were relatively stable because data collection occurred in a controlled laboratory environment. Therefore, we did not consider this temperature-based rule necessary in our analysis, as the controlled conditions minimized the likelihood of improper sensor contact.

3.1.2 Signal Quality Assessment of Electrocardiography

SQA in the context of ECG focuses on evaluating the reliability and clarity of electrocardiogram signals. It involves identifying and quantifying noise, artifacts, and other distortions that may obscure or alter the heart's electrical activity [61]. A thorough SQA process ensures that the ECG data is suitable for further analysis, supporting accurate interpretation and downstream tasks such as feature extraction or classification. Among the most important considerations in ECG SQA is understanding the specific types of noise that can degrade the signal. Various sources of interference can affect recordings, each with unique characteristics and challenges for detection and removal. These include Powerline Interference (PLI), Baseline Wander (BW), Electrode Motion Artifacts (EMA), and Muscle Activity (MA), all of which must be accounted for to ensure high-quality ECG measurements [10].

- PLI: This is electrical interference from the AC power supply, usually appearing as a sinusoidal signal at 50 or 60 Hz [10, 62]. Because it has a narrow and well-defined frequency, it is generally easier to remove using filters.
- BW: Baseline wander is a low-frequency fluctuation of the ECG signal baseline, typically in the range of 0.5–1 Hz. It arises due to factors like breathing or slow body movements [63]. This drift can be significant enough to overshadow important components such as the QRS complex.

- EMA: Electrode movement occurs when the contact between the electrodes and the skin changes, for example, due to body motion or poor attachment [62]. This results in abrupt shifts in impedance, producing large-amplitude distortions often in the 1–10 Hz range. These artifacts can sometimes resemble normal heart activity, which complicates their identification.

- MA: Also known as electromyographic (EMG) noise, muscle artifacts come from electrical activity generated by skeletal muscles during contraction or movement [63]. This type of noise covers a broad frequency spectrum, roughly 0.01–100 Hz, and often overlaps with the ECG's own frequency components, especially the QRS complex [10, 62]. Consequently, it is particularly challenging to separate muscle activity from cardiac signals.

Recognizing and understanding these different noise sources is a necessary first step before applying systematic quality assessment techniques. To ensure the reliability of ECG signals, many Signal Quality Indices (SQIs) has been developed [64]. These metrics capture various aspects of the signal, including time-domain properties, frequency-domain characteristics, QRS complex morphology, and nonlinear dynamics [65, 66].

Signal Quality Indices

SQIs are numerical or categorical metrics that quantify the degree to which an ECG segment is free of noise and suitable for further processing. Essentially, SQIs act as filters at the preprocessing stage, enabling automatic rejection or correction of low-quality signal segments [67].

Over the years, researchers have proposed a variety of methods to compute ECG SQIs. As proposed by Rahman et al. [65], SQIs can be classified into three main categories based on their underlying principles and computational strategies: Statistical, Template-based, and Machine Learning (ML)-based. The Statistical Signal Quality Indicators (SSQIs) rely on statistical properties of the ECG waveform, such as variance, skewness, kurtosis, and signal-to-noise ratio, to estimate signal quality. SSQIs are widely used due to their simplicity and low computational cost, making them especially attractive for real-time monitoring on resource-constrained devices [68–71]. In template-based approaches, each detected heartbeat is compared to a reference template, assessing alignment via correlation metrics. The reference template is typically generated by averaging multiple clean, well-aligned cycles of a signal from the same subject, producing a representative waveform for comparison. High similarity suggests high-quality segments, while low correlation indicates

noise or misalignment. [72–76]. More recent methods have started using supervised learning algorithms that learn discriminative features of clean vs. noisy segments from labeled datasets [67–69, 77–79]. Each of these categories has distinct advantages and limitations. Template-based approaches can be highly accurate when a good reference template is available, but often lack generalizability across subjects [65]. Machine learning methods can model subtle, non-linear noise patterns but depend on large volumes of annotated data and can be computationally intensive. Statistical SQIs, in contrast, remain appealing because they can be calculated efficiently, are interpretable, and can often serve as input features to higher-level classifiers.

SSQIs quantify the regularity and amplitude distribution of the ECG waveform. Several established metrics are as follows:

- Signal-to-Noise Ratio (SNR): SNR measures the relative strength of the ECG signal compared to the noise [80, 81].

$$\text{SNR} = 10 \log_{10} \left(\frac{P_{\text{signal}}}{P_{\text{noise}}} \right) \tag{3.1}$$

  where P_{signal} and P_{noise} are the estimated powers of the clean signal and noise, respectively. Bemmel et al. [81] proposes that an ECG signal with an SNR $\geq$ 80 dB is of good quality or acceptable for further processing.

- Kurtosis (kSQI): Kurtosis quantifies the "peakedness" of the signal's amplitude distribution:

$$kSQI = \frac{1}{N} \left| \sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma} \right)^4 \right| \tag{3.2}$$

  where x_i is the ECG sample, μ is the mean, σ is the standard deviation, and N is the total number of samples. [71] uses kSQI as an ECG signal quality indicator.

- Skewness (skSQI): Skewness measures asymmetry of the amplitude distribution:

$$skSQI = \frac{1}{N} \left| \sum_{i=1}^{N} \left(\frac{x_i - \mu}{\sigma} \right)^3 \right| \tag{3.3}$$

  Noise-free ECG signals typically have moderate skewness values close to zero. [65].

- Relative Power of QRS Complex (pSQI): Signal quality can also be assessed using metrics derived from the frequency-domain characteristics of the ECG. pSQI is a frequency-domain signal quality index that measures the proportion of ECG signal energy concentrated within the QRS frequency band [11, 67, 82].

$$pSQI = \frac{\int_{f=5\,Hz}^{f=15\,Hz} P(f)df}{\int_{f=5\,Hz}^{f=40\,Hz} P(f)df} \tag{3.4}$$

  where $P(f)$ is the power spectral density of the ECG signal. The numerator is the energy of the QRS complex, whose dominant frequency content lies within 5–15 Hz. The denominator integrates over 5–40 Hz, which corresponds to the clinically relevant ECG bandwidth after excluding low-frequency baseline wander (0–5 Hz) and high-frequency noise above 40 Hz.

  Noise-free ECG signals concentrate most of their power in the QRS band, producing higher pSQI values. When noise is present energy spreads into other frequency regions, causing pSQI to decrease [11].

- Relative Power in the Baseline (basSQI): This SQI quantifies baseline wander by measuring the proportion of the ECG's energy that lies in the very-low-frequency band. Following Zhao et al. [11], basSQI is defined as:

$$\text{basSQI} = 1 - \frac{\int_{f=0\,Hz}^{f=1\,Hz} P(f)\,df}{\int_{f=0\,Hz}^{f=40\,Hz} P(f)\,df} \tag{3.5}$$

  where $P(f)$ denotes the power spectral density. The numerator captures baseline wander (< 1 Hz), while the denominator represents the overall energy of the ECG signal (0–40 Hz). Values close to 1 indicate negligible baseline drift, whereas lower values suggest excessive wandering of the ECG baseline due to motion or electrode instability.

These SQIs can be thresholded or combined to form composite indices of ECG quality. In this work, we adopt a template matching approach based on the average correlation coefficient to assess signal quality. The details of these methods are described in the following section.

Template Matching via Average Correlation Coefficient

Template matching offers a technique to assess signal quality. It works by first detecting individual heartbeats (pulses) within the ECG recording. Then, each detected pulse is compared to a predefined waveform template, often representing a typical heartbeat. The Pearson correlation coefficient is computed between each pulse and a waveform template. A high correlation between the pulses and the waveform template signifies minimal noise or distortion in the signal. Therefore, a high correlation coefficient indicates good alignment between the detected pulse and the template, suggesting potentially high-quality ECG data . Conversely, lower correlation values can point towards issues like noise, artifacts, or inaccuracies in pulse detection or the template itself [11, 72, 83].

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(t_i - \bar{t})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(t_i - \bar{t})^2}} \tag{3.6}$$

where x_i represents the samples of the detected pulse, t_i represents the samples of the template waveform, $\bar{x}$ and $\bar{t}$ denote their respective means, and N is the number of samples per beat.

The average correlation coefficient across all detected beats is computed as:

$$\text{Average Correlation} = \frac{1}{M}\sum_{j=1}^{M} r_j$$

where M is the total number of beats in the ECG segment.

Fuzzy Classifier-Based Assessment

Conventional SQI approaches often rely on rigid thresholding schemes that classify signals into binary categories (acceptable or unacceptable). However, this may lead to misclassifications in borderline cases, especially in mobile or ambulatory ECG monitoring environments where noise characteristics are highly variable. To address this limitation, fuzzy comprehensive evaluation frameworks have been proposed as an effective strategy to integrate multiple SQIs and model uncertainty in signal quality assessment [11]. Zhao et al. [11] introduced a fuzzy evaluation system combining simple heuristic fusion and fuzzy logic to assess the quality of single-lead ECG recordings. This system operates in two stages:

1. **SQI Extraction and Fuzzification**: Several SQIs like kSQI, basSQI, pSQI are computed to quantify diverse aspects of ECG quality. Each SQI is fuzzified (e.g., "low," "medium," "high"), and fuzzy rules combine them. For example:
    - If pSQI is high and kurtosis is high, the segment is likely high quality.
    - If pSQI is low, the segment is likely low quality regardless of kurtosis.
2. **Fuzzy Comprehensive Evaluation**: In the second stage, the method builds a fuzzy evaluation matrix by defining membership functions for each SQI relative to three quality levels: excellent, barely acceptable, and unacceptable.

3.1.3 Signal Quality Assessment of Respiration Signal

Respiration rate monitoring provides critical information about respiratory rate (RR), tidal volume, and breathing regularity. However, respiration signals are prone to various artifacts such as motion, sensor displacement, talking, coughing, or electrical interference, all of which can compromise the reliability of estimated respiratory parameters [84,85]. Changes in respiration are known to be closely linked with emotional experiences. Research has shown that different emotions can influence breathing rate, depth, and rhythm. For example, fear and anger are often associated with faster, shallower breathing, while sadness may lead to slower, deeper breaths [38]. However, if the respiration signal is noisy or distorted, these subtle emotion-related patterns may be lost or misinterpreted. Therefore, SQA is an essential preprocessing step to ensure that analyses are based only on physiologically valid data segments.

A typical respiration waveform shows a smooth, sinusoidal pattern corresponding to the inhalation and exhalation cycle. Clean respiration signals exhibit regular peaks and troughs, with relatively stable amplitude and frequency within physiological ranges (typically 0.1–0.5 Hz in adults at rest) [86]. A common way to check if respiration signals are good enough to use is to create SQIs. Many researchers have suggested methods that look at several aspects of the signal together such as its shape, how steady it is over time, and whether the breathing patterns are plausible [84, 85, 87–89]. In this project, we adopt three-step process to assess respiration signal quality inspired by the work of Charlton et al. [84]:

1. Breath Identification: finding where each breath starts and ends.
2. Plausibility Check: making sure the detected breaths have realistic durations.

3. Template Matching: comparing each breath to an average breathing pattern to see if it looks normal.

In this work, a respiration belt was used to record the respiration signals. A respiration belt measures the deflection of the chest as a subject breathes.

3.2 Dataset

This study uses a custom-collected multimodal dataset recorded in the Inspire Laboratory, Department of Psychology, University of Ottawa. The aim was to investigate human emotional responses using physiological and speech signals. The dataset includes recordings from wearable sensors, audio recordings, and self-reported emotional annotations. The experimental protocol was developed in collaboration with the Department of Psychology to ensure realistic emotional responses while minimizing potential confounding factors.

For this study, six common emotion categories were targeted: sadness, fear, joy, disgust, surprise, and anger. For each emotion category, a participant watched six 1-3 minute movie segments intended to evoke that emotion. The study included 99 student participants aged 18 to 27 (M = 19.9, SD = 2.5). The majority were women (81 women, 16 men, and 1 non-binary), and most participants were native English speakers (65 spoke English as their first language, 12 spoke French, and 20 spoke other languages). The sample was ethnically diverse, comprising 16 Asian, 20 Black/African, 7 Hispanic/Latino, 1 Indigenous, 15 Mixed/Multiple Ethnicities, 33 White/Caucasian, and 5 participants identifying as Other.

3.2.1 Hardware Setup and Physiological Signals

The hardware setup involved a PC with two monitors — one used for presenting the video clips to participants and the other for monitoring the physiological signals in real-time. Following informed consent, physiological sensors were attached to the participant's body in accordance with the standardized procedure. The Mindware system was used to record ECG, galvanic skin conductance (GSC, also referred to as electrodermal activity or EDA), and respiration signals. ECG was captured using a three-electrode configuration, with electrodes placed at the right clavicle and the lower left and right ribs. GSC electrodes were positioned on the palm, specifically over the thenar and hypothenar eminences. Respiration was measured using a belt wrapped around the torso. All the three signals were sampled at

1000 Hz. The physiological data acquired by the Mindware hardware were visualized and monitored in real time using BioLab, a dedicated data acquisition and processing software developed by Mindware.

In addition to the Mindware setup that was used as the reference device, wearable devices were integrated into the experiment to capture physiological data in a less obtrusive manner. The Empatica EmbracePlus wristwatch was used to record electrodermal activity, skin temperature, blood volume pulse, and accelerometer signals. The VitalTracer chest patch was used to record ECG, photoplethysmogram (PPG), and accelerometer (ACC) signals.

3.2.2 Experiment Protocol

Each participant completed a structured session to record their emotional responses through physiological signals and speech. The session began with the PANAS questionnaire to assess their mood before starting the experiment. After this, participants recorded a neutral audio recording, reading out the text shown on the screen. Then, they were then shown a series of twelve videos—six emotional, each corresponding to six different emotions (anger, disgust, fear, joy, sad, surprise), and six neutral—presented in random order. A neutral image was shown before each video to bring their emotional state back to baseline. After watching each video, participants rated their emotional experience in terms of valence (pleasant to unpleasant) and arousal (low to high intensity), and ranked how strongly they felt six specific emotions anger, disgust, joy, fear, sadness on a 5-point Likert scale going from *not at all* to *strongly*. This setup allowed us to collect synchronized physiological data, V-A ratings (valence - arousal ratings), categorical self-reports for each emotional stimulus. The entire protocol is illustrated in Figure 3.1.

3.2.3 Video Stimuli

The twelve movie clips used in the study and their meta information is listed in Table A.1. The movie sources are cited in Movie Clip column. Out of the twelve movie clips, seven were used in other psychological research. For the five other clips, we did a pilot study consisting of 25 participants and validated the effectiveness of the clips before the final data collection. The psychology team was responsible for curating the video stimuli and the experiment.

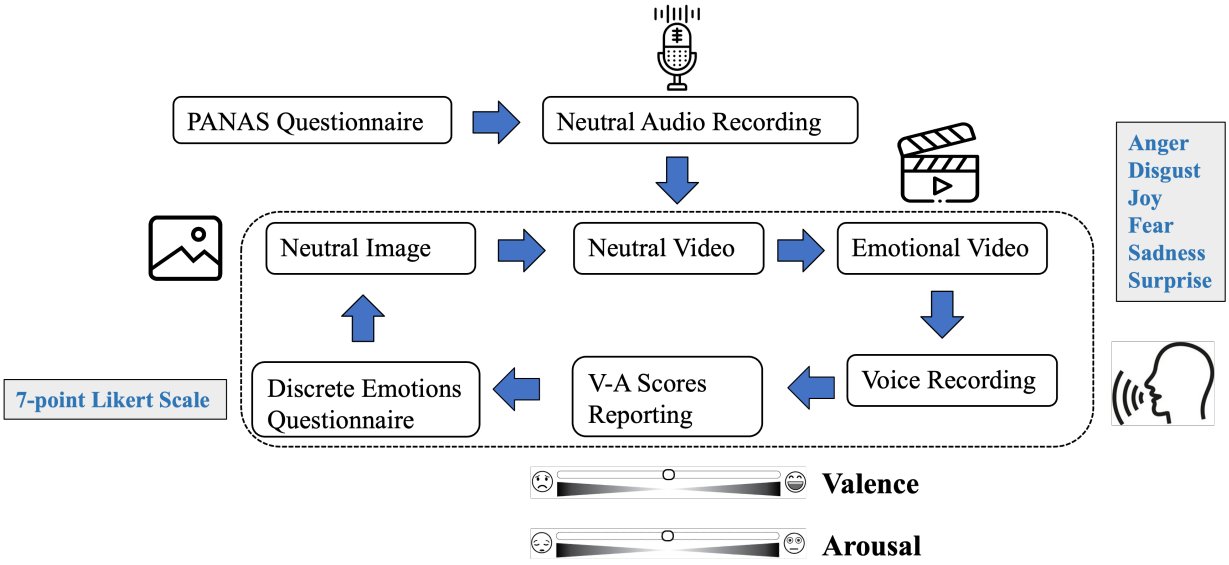


Figure 3.1: Overview of the experimental protocol used for data collection. The session began with the PANAS questionnaire followed by a neutral audio recording to establish baseline physiological signals. Participants were then shown a series of neutral and emotional videos, each preceded by a neutral image. After each video, participants provided a voice recording describing their feelings, rated their experience using valence-arousal sliders, and completed a discrete emotions questionnaire. The setup allowed for synchronized collection of physiological data, self-reported annotations, and speech responses.

3.3 Materials and Methods

3.3.1 Dataset

The dataset used in this thesis consists of synchronized multimodal time-series data collected during the experiment. Each row in the dataset represents a single timestamped data point, sampled at 1000 Hz, and includes physiological signals, metadata, and corresponding emotional annotations.

The dataset contains the following 13 columns:

- **time**: Timestamp of each recorded sample.
- **ecg**: Electrocardiogram signal representing cardiac electrical activity.

- **gsc**: Galvanic Skin Conductance (also known as Electrodermal Activity or EDA).
- **rsp**: Respiration signal representing breathing patterns.
- **video**: The name of the video being presented at that moment (e.g., `Neutral5.mp4`, `1disgust.mp4`).
- **valence**: Self-reported valence score indicating how pleasant or unpleasant the participant felt.
- **arousal**: Self-reported arousal score indicating the intensity of the emotional experience.
- **sadness**, **disgust**, **joy**, **fear**, **surprise**, **anger**: Discrete emotion annotations. Participants rated each emotion on a Likert scale (typically 0 to 5) for each video.

During the neutral video phases, emotion-related annotations (discrete emotions rankings) are absent (marked as `NaN`), since participants did not provide subjective input during passive baseline viewing. These fields are only populated during or after emotional video segments, when participants provided self-reports.

The dataset contains recordings from 99 participants. Figure 3.2 shows a small excerpt of the dataset corresponding to participant number 10, illustrating a few rows of signal recordings along with video labels and emotional annotations.

3.3.2 Procedure

EDA

The EDA signal quality assessment is based on two main criteria: (1) a zero-line check to detect flat or inactive segments, and (2) a rate of amplitude change (RAC) check to detect motion artifacts. The results of both tests are combined using a logical AND operation to classify each time window as either good or bad quality.

The signal is divided into windows of length 1 second. And the checks are applied on these windows.

	time	ecg	gsc	rsp	video	valence	arousal	sadness	disgust	joy	fear	surprise	anger
0	2024-02-08 13:32:30.827	0.043182	0.389404	-0.074463	Neutral5.mp4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	2024-02-08 13:32:30.828	0.041809	0.389709	-0.073395	Neutral5.mp4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	2024-02-08 13:32:30.829	0.042572	0.389252	-0.073242	Neutral5.mp4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	2024-02-08 13:32:30.830	0.046234	0.389404	-0.073242	Neutral5.mp4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	2024-02-08 13:32:30.831	0.049286	0.389557	-0.072327	Neutral5.mp4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1397916	2024-02-08 14:14:30.118	0.024109	0.154419	0.252075	1disgust.mp4	0.214664	0.991308	0.0	6.0	0.0	0.0	0.0	0.0
1397917	2024-02-08 14:14:30.119	0.027466	0.154266	0.252075	1disgust.mp4	0.214664	0.991308	0.0	6.0	0.0	0.0	0.0	0.0
1397918	2024-02-08 14:14:30.120	0.030975	0.154419	0.252228	1disgust.mp4	0.214664	0.991308	0.0	6.0	0.0	0.0	0.0	0.0
1397919	2024-02-08 14:14:30.121	0.031433	0.154877	0.252838	1disgust.mp4	0.214664	0.991308	0.0	6.0	0.0	0.0	0.0	0.0
1397920	2024-02-08 14:14:30.122	0.030823	0.154877	0.253143	1disgust.mp4	0.214664	0.991308	0.0	6.0	0.0	0.0	0.0	0.0

1397921 rows × 13 columns

Figure 3.2: Sample rows from the dataset of Participant 10, showing recorded physiological signals: **ecg** (electrocardiogram in volts), **gsc** (galvanic skin conductance in microsiemens, μS, also known as electrodermal activity or EDA), and **rsp** (respiration signal in volts). The **video** column indicates the name of the video being shown. The **valence** and **arousal** columns reflect self-reported scores (in the range 0–1), provided after the video ends, and are therefore constant for the entire duration of that video. The last six columns (**sadness, disgust, joy, fear, surprise**, and **anger**) represent discrete emotion ratings on a Likert scale (0–5), also reported after each emotional video. The physiological signals are sampled at 1000 Hz. Valence, arousal, and emotion labels are populated across the entire duration of emotional video segments and serve as supervised labels, while the physiological signals act as input features for model training.

1. EDA is out of Range

This step checks for segments where the EDA signal remains below a defined threshold throughout the 1-second window. Such segments may indicate poor skin contact or sensor failure.

Let $s(t)$ be the EDA signal and θ be the minimum acceptable threshold (0.05 µS [36]). For a window w_i, the quality metric is defined as:

$$\text{Metric}_1(w_i) = \begin{cases} 1, & \text{if } \forall t \in w_i,\ s(t) > \theta \\ 0, & \text{otherwise} \end{cases}$$

Only segments where all values exceed the threshold are considered valid.

2. Amplitude Spike Index (ASI)

The EDA signal is highly susceptible to motion-induced artifacts that appear as sudden spikes in amplitude. To detect these artifacts, we compute the **Amplitude Spike Index (ASI)**, which quantifies the magnitude of abrupt amplitude excursions within a window w_i. The index measures the relative difference between the maximum and minimum values in the window, normalized by the larger magnitude of the two values:

$$\mathrm{ASI}(w_i) = \begin{cases} \frac{\max(w_i) - \min(w_i)}{|\min(w_i)|}, & \text{if } \arg\min(w_i) < \arg\max(w_i) \\ \frac{\min(w_i) - \max(w_i)}{|\max(w_i)|}, & \text{if } \arg\min(w_i) > \arg\max(w_i) \end{cases} \tag{3.7}$$

Higher ASI values indicate abrupt spikes, often caused by motion artifacts or abrupt changes in skin conductance unrelated to physiological responses.

Although the literature commonly uses a threshold of $|\mathrm{ASI}| \leq 0.2$ for EDA quality assessment [90, 91], repeated checks and visual inspections of the signals in this study indicated that a more stringent threshold of $|\mathrm{ASI}| \leq 0.1$ was better suited for the characteristics of the current dataset.

A window w_i is accepted only if:

$$-0.1 \leq \mathrm{ASI}(w_i) \leq 0.1$$

The binary metric used to mark acceptance or rejection is:

$$\mathrm{Metric}_2(w_i) = \begin{cases} 1, & \text{if } |\mathrm{ASI}(w_i)| \leq 0.1 \\ 0, & \text{otherwise} \end{cases} \tag{3.8}$$

3. Signal Flatness (Standard Deviation)

This metric evaluates the variability of the EDA signal using a longer 4-second window. Electrodermal activity (EDA) is a slow-varying physiological signal and does not exhibit rapid fluctuations over short time scales. Therefore, applying the flatness check over a longer window is crucial to avoid falsely flagging valid segments as poor quality.

Using a 4-second window allows the metric to capture at least one full cycle of the underlying physiological response [30] and provides a more robust estimate of signal variability. This helps distinguish truly flat or uninformative segments—such as those caused

by sensor detachment or poor skin conductance—from naturally slow but valid changes in the signal.

The variability is quantified using the standard deviation $\sigma(w_i^{(4s)})$ of each 4-second window. A threshold of $\tau = 0.003$ was empirically selected based on visual inspection of the signals in this dataset to retain segments that remain physiologically useful. The quality metric is defined as:

$$\text{Metric}_3(w_i) = \begin{cases} 1, & \text{if } \sigma(w_i^{(4s)}) \geq 0.003 \\ 0, & \text{otherwise} \end{cases} \tag{3.9}$$

Since the flatness metric is computed over 4-second windows, the resulting values are expanded to match the 1-second evaluation windows used for the final decision.

4. Combined Quality Assessment

The final quality label for each 1-second window is computed by combining the results of all three metrics:

$$\text{Quality}(w_i) = \text{Metric}_1(w_i) \wedge \text{Metric}_2(w_i) \wedge \text{Metric}_3(w_i) \tag{3.10}$$

Only windows that satisfy all three conditions are classified as clean and usable for downstream analysis.

4. Signal Quality Score

An overall quality score is computed as the proportion of good-quality windows over the entire segment:

$$\text{Score} = \frac{\sum_{i=1}^{N} \text{Quality}(w_i)}{N} \tag{3.11}$$

where N is the total number of windows.

ECG

To ensure that only clean and reliable ECG data were used in downstream analysis, signal quality was assessed using two methods: a fuzzy logic-based signal quality index and a template matching method.

The first method is based on the fuzzy logic approach proposed by Zhao et al. [11] and implemented in the NeuroKit2 library [92]. Raw ECG signals were first notch filtered to remove powerline interference and then cleaned using NeuroKit's `ecg_clean()` function. The cleaned ECG was passed to the `ecg_quality()` function using the `‘‘zhao2018"` method with the `‘‘fuzzy"` approach. This algorithm evaluates ECG signal quality based on three key features:

- **pSQI**: Ratio of QRS-band power (5–15 Hz) to total power (5–40 Hz),
- **kSQI**: Kurtosis of the ECG signal,
- **basSQI**: Baseline wander ratio (low-frequency power / total power).

These features are evaluated using the following empirical thresholds:

- **pSQI** > 0.5: Indicates strong QRS energy,
- **kSQI** > 5: Indicates a peaked and noise-free distribution,
- **basSQI** < 0.1: Reflects minimal baseline drift.

Based on these values, a fuzzy inference system assigns each ECG segment to one of the following categories:

- **Excellent**: All three features satisfy ideal thresholds — pSQI > 0.5, kSQI > 5, and basSQI < 0.1.
- **Barely acceptable**: One or more features fall slightly outside the ideal range — e.g., $0.4 <$ pSQI ≤ 0.5, $4 <$ kSQI ≤ 5, or $0.1 \leq$ basSQI < 0.15.
- **Unacceptable**: Two or more features are outside acceptable limits — e.g., pSQI ≤ 0.4, kSQI ≤ 4, or basSQI ≥ 0.15.

Only segments classified as "excellent" or "barely acceptable" were retained; segments marked as "unacceptable" were excluded from further analysis.

In the second method, morphological consistency of the ECG signal was evaluated using template matching. The ECG signal was processed using the BioSPPy module [93] to detect R-peaks and extract QRS complexes. A median QRS template was generated from all beats in a segment, and each beat was compared to this template using the Pearson correlation coefficient. Segments with an average correlation coefficient **below 0.8** were considered morphologically inconsistent and were discarded.

These two complementary approaches ensured that both statistical and morphological characteristics of the ECG signal were taken into account while selecting only good-quality segments for downstream processing and classification.

Respiration

The respiration signal quality assessment used in this study follows a structured three-step approach inspired by Charlton et al. [84], aimed at ensuring the accuracy of respiratory rate estimation and downstream analysis. The steps are described below:

1. Breath Identification

The first step involves detecting individual breaths within the respiration signal. This is achieved using a combination of signal preprocessing and peak detection techniques. Specifically:

- **Bandpass filtering** is applied to remove baseline drift and high-frequency noise, preserving the dominant breathing components typically found between 0.1–0.5 Hz [84].
- **Amplitude thresholding** is used to discard minor fluctuations that may arise from noise. Only prominent peaks corresponding to actual inhalation or exhalation cycles are retained.
- **Cycle segmentation** defines a single breath cycle as the time interval between two successive peaks (or troughs), allowing for clear identification of breath duration and frequency [85].

Reliable breath detection is essential, as incorrect identification can lead to errors in respiratory rate estimation and signal quality classification.

2. Plausibility Check

After breath cycles are detected, their durations are evaluated to ensure they fall within physiologically plausible limits. This step helps in removing artifacts such as spurious peaks caused by coughing or motion. The following checks are applied:

- **Duration range filtering**: Breaths that are too short or too long (e.g., outside 1–10 seconds) are considered invalid.
- **Breath variability constraint**: The normalized standard deviation of breath durations must remain below a threshold (e.g., 0.25).
- **Outlier exclusion**: Breaths deviating more than 50% from the median duration are flagged and excluded if they exceed a defined proportion (e.g., 15% of total breaths).
- **Occupancy check**: Valid breaths must cover at least 60% of the total segment duration, ensuring the signal is not dominated by noise or flat regions.

These rules ensure that only realistic and consistent breathing patterns are retained for further analysis.

3. Template Matching

In the final step, the morphology of each detected breath is evaluated by comparing it to an average breath template derived from the same segment. This assesses how consistent each breath is with typical respiratory patterns.

The Pearson correlation coefficient is used to compare each breath waveform x_i with the template waveform m_i, defined as:

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(m_i - \bar{m})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{N}(m_i - \bar{m})^2}}$$

A high correlation indicates that the breath closely resembles the average template, suggesting good signal quality. Segments with a mean correlation above a threshold (e.g., 0.75) are classified as high-quality. This step ensures that only segments with consistent morphology are used for estimating respiratory rate.

Respiratory Rate Estimation

For segments deemed high quality, the respiratory rate (RR) is computed as:

$$RR = \frac{N_{\text{breaths}}}{T_{\text{seg}} \text{ (in minutes)}}$$

where N_{breaths} is the number of valid detected breaths within the segment, and T_{seg} is the duration of that segment in minutes. This provides a reliable estimate of the respiratory rate for their quality

3.4 Results

3.4.1 ECG Signal Quality

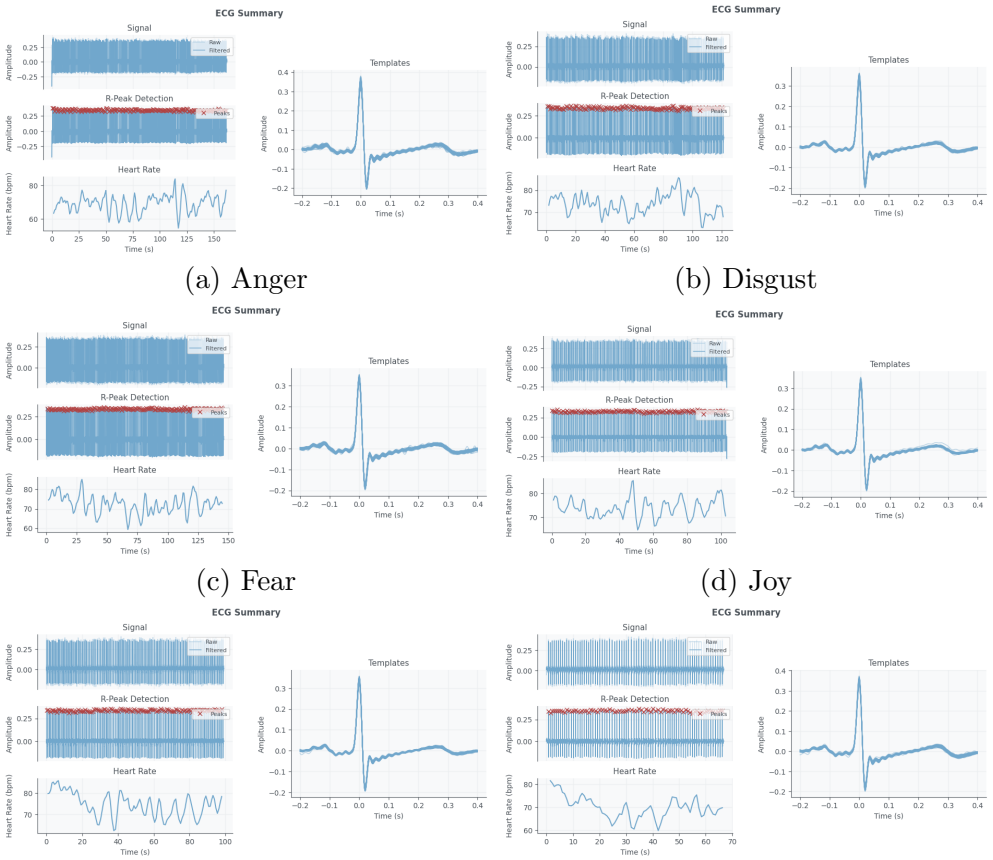
The quality of the ECG signals was assessed using two complementary approaches: a SQI-based fuzzy logic method and a template matching technique, as described in Section 3.3.2. Following the initial signal cleaning and preprocessing, both methods demonstrated consistent performance in identifying high-quality segments. The results from the template matching approach showed good quality for the signals of all participants.

Figure 3.4 illustrates the template matching results for emotional video segments from participant 59. The average correlation between individual QRS complexes and the median template exceeded the quality threshold, indicating clean and reliable ECG morphology. Furthermore, all ECG signals recorded using the Mindware system for this participant were classified as 'Excellent' by the SQI-based fuzzy algorithm.

These results suggest that the ECG data is of sufficient quality for downstream feature extraction and emotion classification tasks.

3.4.2 Respiration Signal Quality

The quality of respiration signals from all 99 participants was evaluated, with each participant contributing six recordings corresponding to six different video sessions. The majority of the recordings demonstrated excellent quality, characterized by smooth and consistent respiratory cycles, minimal baseline drift, and low noise. These signals required minimal correction and were deemed suitable for further analysis.



(a) Anger (b) Disgust

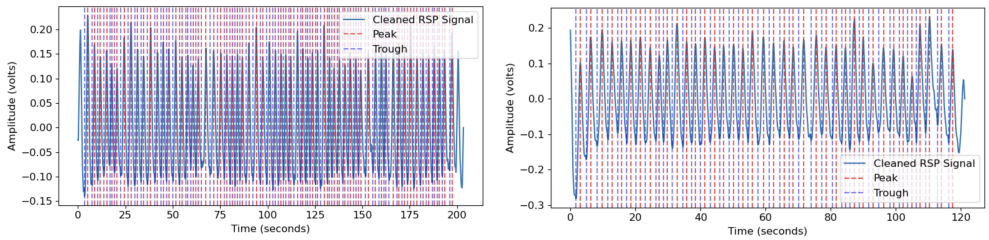
(c) Fear (d) Joy

(e) Sad (f) Surprise

Figure 3.3: Template Matching of ECG signals obtained by Biolab Participant 10 (selected at random).

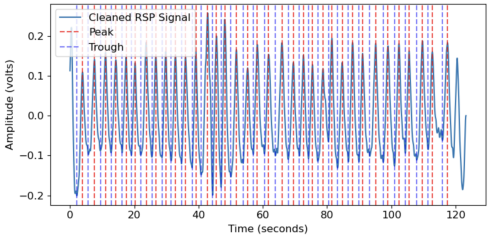
However, a small subset of the recordings initially exhibited suboptimal quality, falling below the predefined threshold due to issues such as motion artifacts or irregular breathing patterns. These low-quality signals were processed using the signal cleaning method proposed by Khodadad et al. [94], as implemented in the NeuroKit2 [94] library. The combination of bandpass filtering and smoothing techniques effectively suppressed noise and gave clean signals.

After applying this preprocessing step across all recordings, even the initially poor-quality signals achieved acceptable levels of clarity and consistency. As a result, all respiration recordings passed the final quality assessment, yielding an overall signal quality score of 100%.

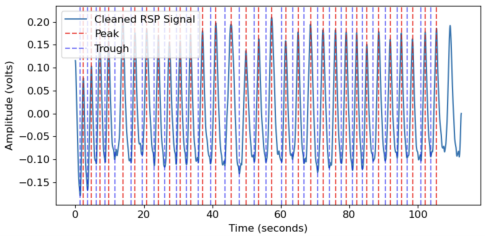


(a) Anger

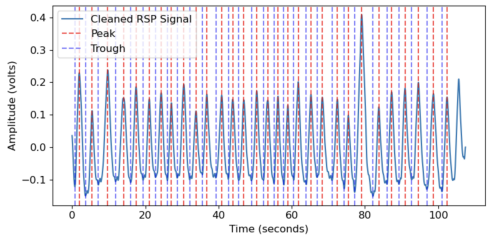
(b) Disgust



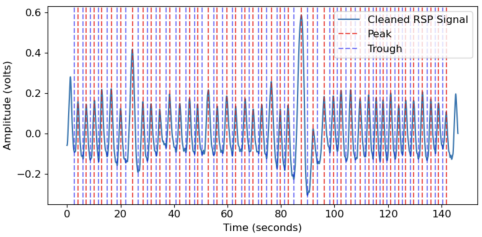
(c) Fear



(d) Joy



(e) Sad



(f) Surprise

Figure 3.4: Template matching of respiration (RSP) signals collected using the Mindware setup for Participant 10 (selected at random), across six emotional video segments: (a) Anger, (b) Disgust, (c) Fear, (d) Joy, (e) Sadness, and (f) Surprise. The blue curve represents the cleaned RSP signal after preprocessing. The dashed red vertical lines indicate detected peaks (points of maximum inhalation), and the dashed violet vertical lines indicate detected troughs (points of maximum exhalation). Each peak-trough pair defines a single breathing cycle, which is used for signal quality assessment and template matching.

3.4.3 EDA Signal Quality

The electrodermal activity (EDA) signals recorded using the Mindware two-electrode setup exhibited several quality-related challenges, including motion-induced artifacts, abrupt baseline shifts, and flat segments with negligible skin conductance variability. These issues were particularly noticeable in participants with excessive hand movements or poor electrode contact during data collection.

To systematically assess signal reliability, each recording was segmented into 1-second non-overlapping windows and evaluated using a set of SQIs, as described in Section 3.3.2. Each window was labeled as either good or bad, and an overall signal quality score was computed for each video as the percentage of good windows.

Video segments with a quality score below 60% were flagged for re-evaluation. Participants whose recordings consistently displayed poor-quality windows across multiple videos—often seen as predominantly red in quality maps—were excluded from further analysis. In contrast, signals with only a few isolated bad windows were corrected using signal cleaning techniques such as median smoothing. This two-stage quality control process ensured that only physiologically meaningful and reliable data were retained for downstream emotion classification.

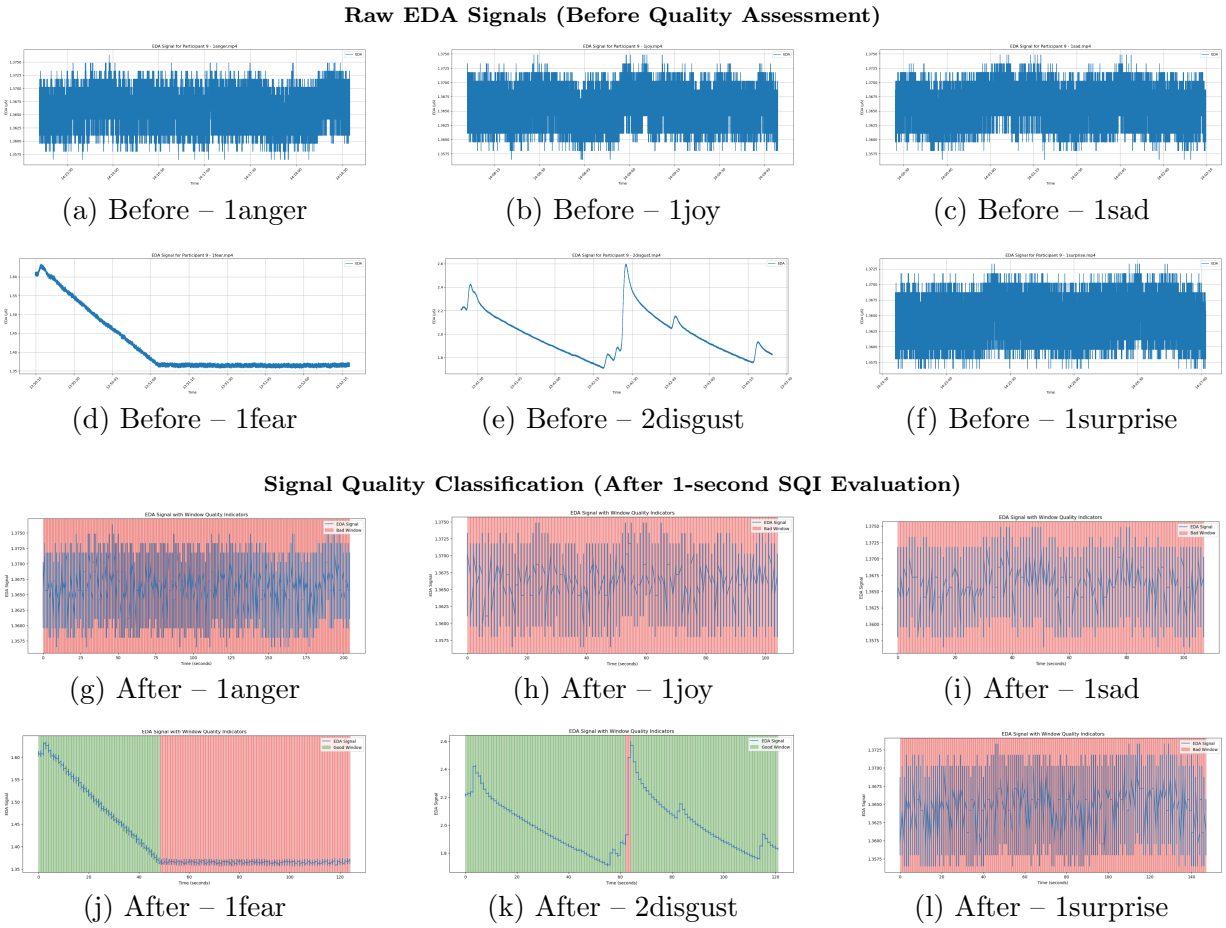
Figure 3.5 illustrates the signal quality evaluation process for EDA signals recorded from Participant 9 across six emotional video segments. The plots display raw EDA signals followed by their corresponding quality assessments using one-second non-overlapping windows. The upper row reveals raw signal characteristics, showing distinct artifacts such as high-frequency noise (e.g., 1anger and 1joy), flatlining or low variability (e.g., 1fear), and inconsistent baseline drifts (e.g., 2disgust).

In the bottom row, signal segments are color-coded based on quality classification derived from three signal quality indices (SQIs): zero-line threshold, rate of amplitude change (RAC), and standard deviation (STD) flatness. Green windows denote segments that passed all quality checks, while red windows indicate poor-quality segments.

For the 1fear video (Figure 3.5j), the EDA signal satisfied the zero-line and RAC metrics throughout the recording, confirming valid amplitude and physiologically plausible changes. However, it failed the STD flatness metric from window 45 onward, indicating a prolonged period of low signal variability. Consequently, only the initial 39% of the signal is marked as high quality, while the remaining 61%—visually shaded red—reflects poor signal due to near-flatline activity. This results in a low overall quality score of 0.39, suggesting the signal is unsuitable for further analysis due to potential sensor detachment or reduced skin conductance response.

In contrast, the 2disgust video (Figure 3.5k) demonstrates consistently high-quality signal, with all windows passing the zero-line and flatness metrics. Only two consecutive windows (around 61–62 seconds) failed the RAC metric. These are visibly marked in red but represent a negligible portion of the signal. The resulting quality score is 0.983, indicating overall high signal reliability. Because the deviation is brief and isolated, the signal is retained for further analysis.

In the remaining emotion conditions—1anger, 1joy, 1sad, and 1surprise (Figures 3.5g,



(a) Before – 1anger (b) Before – 1joy (c) Before – 1sad

(d) Before – 1fear (e) Before – 2disgust (f) Before – 1surprise

(g) After – 1anger (h) After – 1joy (i) After – 1sad

(j) After – 1fear (k) After – 2disgust (l) After – 1surprise

Figure 3.5: Raw electrodermal activity (EDA) signals from Participant 9 during six emotional video segments are in the top two rows. The signals exhibit various artifacts including extreme noise, flatlining, and abrupt discontinuities, which can negatively impact feature extraction and emotion classification. The bottom two rows represent the signal quality classification of the 1 second segments based on the SQIs. Red regions indicate low-quality signal segments, while green regions denote acceptable signal quality. In this case, almost the entire signal is of bad quality, and thus, this participant's data is rejected.

3.5h, 3.5i, 3.5l), most of the signal windows are marked in red. This is primarily due to consistent failure of the flatness metric, indicating minimal physiological variability. Despite passing the amplitude and RAC thresholds, the lack of variation renders these signals unreliable.

Table 3.1: List of participants with bad quality EDA signals and the associated videos and scores

Participant	**Video(s)**	**Score(s)**
7	2fear.mp4	0.521
9	2disgust.mp4, 1fear.mp4, 1sad.mp4, 1joy.mp4, 1surprise.mp4, 1anger.mp4	0.983, 0.39, 0.000, 0.000, 0.000, 0.000
23	1fear.mp4, 1joy.mp4, 2surprise.mp4, 1disgust.mp4, 1anger.mp4	0.000, 0.612, 0.394, 0.488, 0.241
26	2surprise.mp4, 1anger.mp4, 2fear.mp4, 2disgust.mp4, 2joy.mp4, 2sad.mp4	0.000, 0.000, 0.000, 0.000, 0.000, 0.000
31	1fear.mp4	0.000
35	2sad.mp4, 2joy.mp4	0.351, 0.221
41	1anger.mp4, 2sad.mp4, 1disgust.mp4, 2disgust.mp4	0.284, 0.000, 0.000, 0.000
45	2disgust.mp4	0.000
46	1disgust.mp4, 1sad.mp4, 2fear.mp4	0.000, 0.000, 0.000
51	1disgust.mp4, 2fear.mp4, 1fear.mp4	0.000, 0.000, 0.000
53	2joy.mp4	0.232
54	1joy.mp4, 1anger.mp4	0.116, 0.260
59	1fear.mp4, 1anger.mp4	0.000, 0.126
60	1sad.mp4, 2joy.mp4, 1disgust.mp4	0.000, 0.191, 0.000
62	1joy.mp4	0.000
92	1anger.mp4	0.000

Given the continous signal degradation across most emotion conditions for Participant 9, this participant's EDA data was excluded from downstream analysis and model training for emotion classification.

Table 3.1 provides a summary of the EDA signal quality assessment results across participants, based on the evaluation of 1-second non-overlapping windows using signal quality indices (SQIs). Only those participants whose signals failed to meet the minimum quality threshold (i.e., signal quality score below 60%) for at least one video segment are included in the table.

The "Participant" column identifies the subject number. The "Video(s)" column lists

the emotional video segments for which the participant's EDA signal was deemed of poor quality. The corresponding "Score(s)" column reports the computed signal quality score for each of those segments. A score of 0.000 indicates that none of the 1-second windows in the segment passed the quality check, implying complete signal degradation.

For example, Participant 9 shows extremely low scores (including multiple zeros) across six different video segments, clearly indicating unreliability and justifying exclusion from downstream analysis. In contrast, Participant 35 has only two segments with moderately low scores (0.351 and 0.221), suggesting that their data might still be salvageable after applying signal cleaning techniques. Participants 9, 26, 31, 45, 46, 51, 62, and 92 recorded some (or all) segments with a score of 0.000, indicating that the entire EDA signal was unusable due to artifacts, flatlining, or extreme noise. These participants will be excluded from downstream signal quality assessment tasks and classification models to ensure that only high-integrity signals contribute to the emotion recognition pipeline.

In addition to automated evaluation using SQIs, visual inspection of EDA signals was also conducted to validate the assessment results. Given the characteristic nature of artifacts such as flatlining, high-frequency noise, and abrupt shifts, many signal quality issues were clearly identifiable through manual observation. This visual verification served as a supplementary check to ensure that the SQI-based labels were consistent with observable signal patterns, thereby enhancing the robustness of the proposed EDA quality assessment process.

The EDA SQA ensured that only reliable EDA recordings were retained for analysis. By identifying and excluding participants with severely corrupted signals and applying targeted cleaning techniques to borderline cases, the integrity of the dataset was significantly enhanced.

3.4.4 Final Results

To ensure the integrity and consistency of the physiological dataset used for emotion recognition, a rigorous multi-stage filtering and validation process was employed. This included signal quality assessment, manual inspection of emotion labels and audio transcripts for annotation verification, and screening for medical or procedural concerns. Based on this comprehensive evaluation, a number of participants were excluded from the final classification experiments.

Participants 9, 41, 45, 46, 59, 60, 62, and 92 exhibited severely corrupted physiological signals, including undetectable GSC values, ECG baseline drift (visual inspection), that failed the signal quality thresholds or could not be reliably recovered. Participant 54

Table 3.2: Participants excluded from downstream classification tasks, categorized by exclusion criteria.

Reason for Exclusion	Participants and Explanation
Severely corrupted physiological signals	**Participants 9, 41, 45, 46, 59, 60, 62, 92**: Exhibited severely corrupted physiological signals, including undetectable GSC values, ECG baseline drift, or inconsistent respiration waveforms, that failed quality thresholds or could not be reliably recovered. **Participant 54**: Showed persistent motion artifacts or sensor dropout in multiple modalities. **Participant 31**: Suffered data interruptions or signal loss during recording, leading to incomplete physiological profiles.
Faulty or incomplete annotations	**Participants 9, 13, 35, 51, 53, 59**: Displayed annotation issues such as mismatched video-emotion labels, incomplete affect ratings, or verbal responses inconsistent with stimulus timelines.
Medical or procedural anomalies	**Participant 18**: Reported a medical condition (e.g., cardiac or metabolic) that could impacted physiological measurements. **Participant 23**: Encountered setup issues resulting in unreliable baseline signals. **Participant 26**: Did not meet physiological or demographic inclusion criteria. **Participant 44**: Displayed atypical physiological patterns likely caused by sensor misplacement or other confounding factors.

demonstrated persistent motion artifacts or sensor dropout in multiple modalities, while Participant 31 experienced data interruptions leading to signal loss during recording, leading to incomplete physiological profiles. Although no explicit artifacts were detected in the signals of Participant 18, the recordings exhibited considerable noise, and the extracted features deviated significantly from the normal patterns observed among the rest.

Several participants were excluded due to faulty or incomplete annotations. Partic-

ipants 9, 13, 35, 51, 53, and 59 showed annotation discrepancies such as mismatched video-emotion labels, missing affect ratings, or verbal summaries that contradicted the stimulus timeline, undermining the reliability of emotional ground truth.

Additional exclusions were based on medical, demographic, or procedural concerns. Participant 23 encountered procedural issues during sensor setup, resulting in unreliable baseline measurements. Participant 26 did not meet the physiological or demographic inclusion criteria, while Participant 44 displayed physiological patterns inconsistent with the general cohort, likely due to sensor misplacement or other confounding factors.

After consolidating all exclusion criteria, the following participants were removed from downstream classification tasks: **9, 13, 18, 23, 26, 31, 35, 41, 44, 45, 46, 51, 53, 54, 59, 60, 62, 92**. A detailed summary is provided in Table 3.2. **In total, 81 out of the original 99 participants were retained** for downstream classification tasks. This final filtered dataset ensures high data quality, and physiologically meaningful signals required for building the models for emotion recognition.

Chapter 4

Classification of Emotions based on Handcrafted Features

This chapter presents the modeling approaches used in this study, incorporating both classical machine learning and deep learning methods applied to the cleaned and preprocessed physiological signal dataset. The chapter first describes the extraction and selection of handcrafted features and then evaluates multiple classifiers based on these features. Several experimental setups were explored to address different formulations of the emotion recognition task, including: (i) binary classification of arousal and valence, and (ii) multiclass classification of discrete emotions. All models were trained and validated using stratified group cross-validation to ensure participant-independent evaluation and to assess the generalizability of the classification performance.

4.1 Aim and Focus

This chapter investigates the use of handcrafted features derived from physiological signals for emotion classification. The objective is to explore the performance of classical machine learning models in recognizing emotional states based on interpretable signal characteristics. One of the aims is to assess how well these models generalize across participants, particularly in predicting emotions in previously unseen data. In addition, we examine the discriminative capacity of individual physiological signals in distinguishing between different emotional states. This chapter also investigates the benefits of multimodal fusion by incorporating features from physiological signals and speech data. We compare classifier

performance using a range of metrics and provide a detailed analysis of confusion matrices and participant-wise results.

4.2 Methodology

4.2.1 Pre-processing of Physiological Signals

In this study, physiological signals including electrocardiogram (ECG), electrodermal activity (EDA), and respiration (RSP) were recorded using the MindWare acquisition system. The dataset consists of recordings from 99 participants.

Prior to analysis, the quality of all physiological signals was evaluated using signal quality metrics described in Chapter 3. Artifacts in the signals were mitigated through preprocessing techniques such as rolling filter and median smoothing for EDA, the Pan-Tompkins algorithm for ECG [95, 96]. Participants with consistently poor signal quality across modalities were excluded from the study. Specifically, data from the following participants were discarded due to extensive signal artifacts: 9, 13, 18, 23, 26, 31, 35, 41, 44, 45, 46, 51, 53, 54, 59, 60, 62, and 92.

Two segmentation strategies were used and are compared in the later sections:

- **Fixed-length segmentation:** The signals were segmented into overlapping windows of 30 seconds with a 10-second overlap. This approach allows consistent comparison across time windows and has been commonly used by several studies [97–99] for training machine learning models. Prior work shows that ultra-short HRV recordings (10–30 s) retain strong agreement with standard 5-min HRV measurements and are suitable for tracking autonomic changes [100]. A 30-second window is also long enough to capture complete EDA phasic responses (typically 1–4 s rise, $\sim$ 10 s recovery) [30] and multiple respiration cycles ($\approx$ 12–20 breaths/min) [101]. The 10-second overlap preserves continuity and reduces the risk of cutting emotional transitions between windows.
- **Emotion-triggered segmentation:** Each video was reviewed by expert annotators to identify the approximate start point of emotion elicitation. Based on these annotations (as listed in Table 4.1), the physiological signals were segmented from the identified trigger point to the end of the corresponding video. Each video thus corresponds to one segment, with durations varying depending on the length of the emotional clip.

Label	Video Clip	Duration	Trigger	Key Moment/Quote
1Fear	Blairwitch Project Finale (Final scene)	2:03	0:10	"Oh Jesus"
2Fear	The Conjuring (Bedroom scene)	2:26	0:21	"Cindy?", Girl banging head on cupboard
1Anger	American X (Savage attack on car stealers)	3:24	0:58	Opens the door with brutality and gun
2Anger	Platoon (Burning village and rape scene)	2:42	0:38	Villagers pushed around in burning village
1Joy	Baby laughing hysterically at ripping paper	1:44	0:09	Baby starts laughing
2Joy	Cats and Dog playing together on a bed	1:53	0:02	Joyful music with cat walking in front of dog
1Surprise	One Day (Emma does not make it home)	2:26	0:36	"I am sorry"
2Surprise	Neighbors (Airbag prank scene)	1:07	0:37	Sits on exploding airbag
1Disgust	TrainSpotting (Worst Toilet in Scotland scene)	1:23	0:10	Man enters toilet cabinet
2Disgust	Planet Terror (Hospital horrible infections scene)	2:01	0:11	Man shows infected arm
1Sad	Bambi with dead mother	1:44	1:00	Bambi is sitting by dead mother's body with sad music notes
2Sad	My Girl (Funeral scene)	1:39	0:22	Crying girl speaks to dead boy in casket

Table 4.1: Start point of emotions elicitation in the video stimuli used in the experiment. Each row corresponds to a specific video clip used to elicit a target emotion. The **Label** indicates the emotion type and the clip index. **Video Clip** describes the video used. **Duration** refers to the full length of the video clip. **Trigger** marks the timestamp (in minutes:seconds) where the emotion-inducing moment begins. **Key Moment/Quote** highlights the scene or dialogue intended to trigger the emotional response.

4.2.2 Feature Engineering

We extracted a comprehensive set of 44 handcrafted features from all the modalities ECG, EDA, and respiration. For ECG signals, time-domain and frequency-domain features were extracted using the NeuroKit2 [92] and HeartPy [102] libraries. The preprocessing included peak detection (R-peaks) and artifact correction. From the resulting inter-beat intervals (IBIs), we computed: Heart rate (HR) and heart rate variability (HRV) metrics including: Mean HR, SDNN, RMSSD, frequency-domain components like LF (low frequency), HF (high frequency), LF/HF ratio, and nonlinear Poincaré features (SD1, SD2).

The EDA signal was first decomposed into phasic, and tonic components using convex optimization technique introduced by Greco et al. [103], and then features like skin conductance level (SCL), number of skin conductance responses, mean amplitude of SCRs, rise time, and recovery time were extracted. The Neurokit2 library was used in extracting both the EDA, and Respiration signal features. The extracted features were validated against the features that are extracted by the Biolab data processing software by Mindware.

The complete set of features have been listed in the Table 4.2. These features have been widely used in emotion recognition research and shown to be effective in capturing autonomic nervous system responses associated with emotional arousal and valence.

Feature Name	**Description**
ECG Features	
ECG-BPM	Beats Per Minute - Reflects the average heart rate.
ECG-IBI	Interbeat Interval - The time interval between consecutive heartbeats.
ECG-SDNN	Standard Deviation of NN Intervals - Measures variability in beat intervals.
ECG-SDSD	Standard Deviation of Successive Differences - Reflects short-term HRV.
ECG-RMSSD	Root Mean Square of Successive Differences - A measure of short-term HRV.
ECG-PNN20	Percentage of NN intervals differing by >20 ms - HRV indicator.
ECG-PNN50	Percentage of NN intervals differing by >50 ms - HRV indicator.

Feature Name	Description
ECG-HR-MAD	Median Absolute Deviation of Heart Rate.
ECG-SD1	Short-term HRV via Poincaré plot.
ECG-SD2	Long-term HRV via Poincaré plot.
ECG-S	Area of ellipse in Poincaré plot - reflects both short/long-term HRV.
ECG-SD1/SD2	Ratio of SD1 to SD2 - balance between HRV types.
ECG-Breathing Rate	Derived breathing rate via respiratory sinus arrhythmia.
RSP Features	
RSP_Rate_Mean	Mean respiratory rate post stimulus.
RSP_Intervals	Intervals between respiratory peaks.
RSP_Std	Standard deviation of respiration signal.
RSP_Mean	Mean of Respiratory Amplitude Variability.
RSP_Phase_Duration_Ratio	Inspiratory-to-expiratory time ratio (I/E ratio).
RSP_RVT	Respiratory Volume per Time.
RSP_STE	Short Time Energy of respiration.
RSP_ZCR	Zero Crossing Rate of respiration signal.
RSP_Centroid	Spectral centroid of respiration signal.
RSP_Flux	Spectral flux of respiration signal.
RSP_Spread	Spectral spread of respiration signal.
EDA Features	
scr_count	Number of GSR amplitude peaks.
scr_meanpeakAmp	Mean peak amplitude of GSR.
scr_scl	Skin Conductance Level.
scr_minSpectPow	Minimum spectral power of GSR.
scr_maxSpectPow	Maximum spectral power of GSR.
scr_mean	Mean of GSR signal.
scr_median	Median of GSR signal.
scr_std_dev	Standard deviation of GSR signal.
scr_mfcc_mean	Mean of Mel-frequency Cepstral Coefficients (MFCC) of GSR.
scr_mfcc_median	Median MFCC coefficient of GSR.
scr_mfcc_std_dev	Std. deviation of MFCC coefficients of GSR.
scr_DWT_mean	Mean of DWT coefficients of GSR.
scr_DWT_median	Median of DWT coefficients of GSR.

Feature Name	Description
scr_DWT_std_dev	Std. deviation of DWT coefficients of GSR.
scr_ste	Short Time Energy of GSR.
scr_zcr	Zero Crossing Rate of GSR.
scr_rms	Root Mean Square of GSR signal.
scr_centroid	Spectral centroid of GSR signal.
scr_flux	Spectral flux of GSR signal.
scr_spread	Spectral spread of GSR signal.
scr_slope	Slope of GSR signal.
scr_rolloff	Spectral rolloff of GSR signal.
scr_decrease	Spectral decrease of GSR signal.
scr_flatness	Spectral flatness of GSR signal.
scr_kurtosis	Kurtosis of GSR signal.
scr_skewness	Skewness of GSR signal.
scr_crest_factor	Crest factor of GSR signal.

Table 4.2: Features and their Descriptions for ECG, RSP, and SCR Signals

All features were normalized per modality to ensure consistent scaling before fusion or classification.

Feature Selection

Feature selection is a crucial step in building efficient and interpretable machine learning models. The following systematic approach is adopted to select the most relevant features:

Feature Selection

Feature selection is a critical step for improving model interpretability, reducing redundancy, and preventing overfitting. In this study, feature selection was performed using a **Recursive Feature Elimination (RFE)** approach. This method iteratively removes less informative features based on their importance values, allowing the model to retain only those that contribute most to classification accuracy.

The process followed in this work can be summarized as follows:

1. **Model training:** A classifier is trained on the full handcrafted feature set. Multiple models were used in the RFE process, including both tree-based (Random Forest, Decision Tree, XGBoost) and non-tree-based (Logistic Regression, Perceptron) classifiers.

2. **Compute feature importance:** Feature importance values are obtained differently depending on the classifier type:

   - For **tree-based models**, feature importance is derived from the model's internal impurity reduction criterion (`feature_importances`), which quantifies each feature's contribution to improving node purity during decision splitting.
   - For **linear models**, feature importance is calculated as the absolute magnitude of the model coefficients, where larger coefficients indicate stronger influence on the decision boundary.

3. **Feature elimination:** The feature with the lowest importance score is removed from the dataset.

4. **Retraining and evaluation:** The model is retrained on the reduced feature subset, and its classification performance (accuracy or F1-score) is re-evaluated.

5. **Recursive refinement:** Steps 2 to 4 are repeated iteratively until further feature removal leads to a decline in performance. The remaining features at this stage represent the optimal subset for emotion classification.

This iterative feature elimination framework allows the model to balance complexity and generalization, ensuring that only the most discriminative physiological features are retained for subsequent classification tasks.

4.2.3 Classification Methods

We evaluated the classification capacity of handcrafted features, we evaluated the following models :-

Random forest

Random Forest is an ensemble learning algorithm that builds multiple decision trees during training and outputs the mode of their predictions. It introduces randomness by selecting

random subsets of features for each tree and uses bootstrap aggregation (bagging) to ensure diversity among trees. This results in a model that is robust to overfitting and capable of handling high-dimensional data with complex interactions. One of its strengths lies in the ability to provide feature importance scores, which helps interpret the relevance of physiological features in emotion recognition. Due to its effectiveness in modeling non-linear decision boundaries and reducing variance, Random Forest is widely adopted in biosignal classification tasks [16, 104].

Decision Tree

A Decision Tree is a supervised learning algorithm that splits the feature space into regions by creating a flowchart-like structure of decisions based on input variables. Each internal node represents a decision rule on a feature, and each leaf node represents an outcome. Decision Trees are easy to visualize and interpret, making them particularly useful when interpretability is essential. However, they are prone to overfitting, especially when grown to full depth. In this study, they serve as a baseline to evaluate the performance of a single-tree model against more complex ensemble methods for classifying emotional states from physiological signals [105].

Gradient Boosting

Gradient Boosting is a machine learning technique that builds models sequentially, with each new model aiming to correct the errors made by the previous ones. Unlike Random Forests, which focus on reducing variance, Gradient Boosting reduces bias by optimizing a loss function using gradient descent. It assigns more focus to difficult cases, leading to strong predictive performance. In this study, Gradient Boosting is used to identify complex patterns in physiological signals associated with different emotional states [106].

Logistic Regression

Logistic Regression is a widely used method for binary and multiclass classification. It estimates the probability of a class by applying a sigmoid function to a linear combination of input features. While simple, it is effective when the relationship between input features and the outcome is roughly linear. In this work, Logistic Regression serves as a strong baseline to evaluate how well handcrafted features can classify emotional states [107].

Perceptron

The Perceptron is one of the earliest neural models, used for binary classification tasks. It calculates a weighted sum of input features and applies a threshold function to make decisions. Though it only works well with linearly separable data, it remains useful for understanding neural-based classifiers. Here, the Perceptron is employed to assess how a basic neural model performs in classifying emotions from physiological signals [108].

Gated Linear Units (GLU)

Gated Linear Units (GLU) are a type of neural architecture that builds upon standard feedforward networks by introducing a learnable gating mechanism at each layer. In their simplest form, GLUs apply two parallel linear transformations to the input: one path passes through a sigmoid activation to produce a gate, while the other generates the raw linear output. These two outputs are then multiplied element-wise. This enables the model to learn which features to retain and which to suppress, offering more control over information flow. Mathematically, for an input vector $\mathbf{x}$, the output of a GLU layer is given by:

$$\text{GLU}(\mathbf{x}) = (\mathbf{x} \cdot \mathbf{W}_1 + \mathbf{b}_1) \odot \sigma(\mathbf{x} \cdot \mathbf{W}_2 + \mathbf{b}_2) \tag{4.1}$$

where $\odot$ denotes element-wise multiplication, σ is the sigmoid activation function, and $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{b}_1$, $\mathbf{b}_2$ are trainable parameters. This structure resembles a traditional multi-layer perceptron (MLP) but includes a dynamic filtering mechanism that adjusts the influence of features during training [109].

In this study, we implemented a fully connected GLU-based neural network composed of multiple hidden layers, interleaved with batch normalization and dropout for regularization. The architecture can be seen as an extension of conventional MLPs, enhanced by the gating mechanism to better capture feature interactions—particularly in high-dimensional tabular settings such as physiological feature data. The model was trained using the Adam optimizer with categorical cross-entropy loss for multiclass emotion classification. Due to its gating structure, GLUs are often more expressive than standard MLPs while maintaining a simple and interpretable form. Their ability to learn which features to emphasize at each layer makes them particularly suitable for emotion recognition from structured physiological signals.

TabNet

TabNet is a deep learning architecture specifically designed for learning from tabular data, a domain where traditional deep neural networks often underperform compared to ensemble methods like Random Forests and Gradient Boosting Machines. Unlike conventional models that process all input features at once, TabNet introduces a **novel sequential attention mechanism** that allows the network to selectively attend to different subsets of features at each decision step. This feature-wise attention is learned dynamically, enabling the model to focus on the most informative inputs while ignoring irrelevant or redundant features. As a result, TabNet achieves a balance between performance and interpretability, making it a strong candidate for physiological signal analysis where certain features may dominate depending on the emotional state being classified [110].

One of the key advantages of TabNet is its inherent ability to provide feature importance scores using sparse feature selection masks. During training, the model applies a sparsemax activation function, which encourages sparsity in the learned feature masks helping to reduce overfitting and enhancing generalizability [111]. The architecture also supports decision aggregation across multiple steps, where partial decisions from each step are combined to produce the final output, mimicking an ensemble of weak learners within a single model. In this study, TabNet is applied to handcrafted features extracted from ECG, EDA, and RSP signals. The model is trained using categorical cross-entropy loss with early stopping, and its performance is evaluated using stratified cross-validation. Due to its structured attention and interpretability, TabNet is particularly well-suited for understanding which physiological patterns contribute most to emotion discrimination.

4.3 Experimental Tasks

This section presents a series of experiments designed to evaluate the effectiveness of handcrafted features in classifying emotions using physiological signals. The experiments explore three primary classification tasks:

- **Binary Arousal Classification:** Classifies emotional responses as either high or low arousal, based on participants' self-reported arousal scores.
- **Binary Valence Classification:** Categorizes emotions into high or low valence classes, according to the participants' reported valence values.

- **Multiclass Emotion Classification:** Assigns one of six discrete emotional categories—*anger*, *disgust*, *fear*, *joy*, *sadness*, and *surprise*, based on the intended emotion for each video stimulus.

Additionally, we test performance under different experimental settings, including baseline classification, personalization improvements through baseline reduction, and multimodal fusion with speech features. A consistent evaluation framework was followed using participant-independent 10-fold cross-validation, with performance assessed through standard metrics accuracy, F1-score, and confusion matrix.

4.3.1 Experiment 1: Emotions Classification

4.3.1 Experiment 1: Emotions Classification

The first experiment aims to assess the discriminative ability of handcrafted features derived from ECG, EDA, and Respiration signals for emotion classification. During preprocessing, the data were segmented in two ways: (1) fixed-length 30-second windows with a 10-second overlap, and (2) variable-length segments based on manually annotated emotion trigger timestamps.

For the binary classification tasks, these segmentation strategies were evaluated separately to determine their influence on model performance for **binary arousal (high/low)** and **binary valence (high/low)** prediction.

For the multiclass emotion classification task, two types of discrete labels were used:

- **Participant-Reported Labels:** self-reported emotional states recorded after each stimulus, reflecting each participant's subjective perception.
- **Intended Video Emotion Labels:** target emotions defined by the experimental design, corresponding to the annotated emotion of each video stimulus.

These two labeling schemes allow a comparative analysis between model performance on participant responses and experiment-intended emotional categories.

To ensure robust evaluation and prevent subject leakage, a strictly subject-independent protocol was employed. Participants with IDs greater than 80 were reserved exclusively for the test set, while participants up to 80 were used for training and validation. During training, a Stratified Group K-Fold cross-validation strategy was applied, ensuring that all

samples from a single participant remained within the same fold. This setup maintains the balance of emotion classes across folds while preserving subject independence.

After labeling and partitioning, the following preprocessing steps were applied to prepare the data for model training:

- **Label Encoding:** The categorical emotion labels were converted into numerical format using label encoding to make them compatible with the machine learning models.
- **Feature Scaling:** The features were standardized using the `StandardScaler` from the `scikit-learn` library, ensuring that all features were on the same scale.

Model hyperparameters were tuned using `GridSearchCV` from the `scikit-learn` library, which performs an exhaustive search over predefined parameter combinations within each cross-validation fold to identify the optimal configuration for each classifier. For each segmentation and labeling strategy, models were trained using the extracted handcrafted features, and multiple classifiers were evaluated. Model performance was assessed using standard evaluation metrics, including accuracy, F1-score, and confusion matrices.

4.3.2 Experiment 2: Binary (One v/s Rest) Emotions Classification

This experiment is designed to classify each of the six target emotions—sadness, disgust, joy, fear, surprise, and anger—as either present or absent in a given physiological signal segment. Each classification task is structured as a binary classification problem, with samples labeled either as containing the target emotion or not. Since this formulation treats each emotion independently and separates the data accordingly, it can be viewed as a form of **one-class classification applied in a binary classification setup**.

The experiment follows a pipeline to ensure class balance, participant independence, and fair evaluation across multiple classifiers:

1. Data Preparation and Emotion Labeling

The features corresponding to trigger-based segments are used in this experiment. In the dataset (Section 3.3.1), each sample includes continuous self-reported emotion scores for the six target emotions. For each target emotion E, a binary label is assigned to each segment based on the participant's self-reported emotion scores:

- Samples where the emotion E score is high (typically $\geq$ threshold ($= 4$ in this case) are labeled as **1** (emotion present).
- Samples where the emotion E score is low (i.e., below the threshold) are labeled as **0** (emotion absent).

This binary labeling strategy is applied independently for each of the six target emotions: sadness, disgust, joy, fear, surprise, and anger. As a result, six binary classification tasks are created—one for each emotion.

Due to the relatively fewer samples where any single emotion score exceeds the threshold, this labeling process naturally results in class imbalance, with a larger number of samples labeled as 0 (emotion absent) than 1 (emotion present).

2. Handling Class Imbalance with Subsampling

In most cases, the number of emotion-absent (label 0) samples is significantly larger than emotion-present (label 1) samples, leading to class imbalance. To address this, all positive class samples are retained. The negative class (label 0) samples are randomly divided into multiple equal-sized subsets, each **matching** the number of positive samples. Each subset is paired with the full positive class to form a **balanced binary dataset**, resulting in multiple **outer fold groups** for each emotion.

3. Cross-Validation with Participant Grouping

For each outer fold group (one positive class + one negative subset), we perform a **5-fold Stratified GroupKFold cross-validation**. Grouping is based on participant IDs to prevent data from the same participant appearing in both training and testing sets, ensuring no participant leakage. This split is stratified on the binary label to maintain label balance across folds.

4. Classifier Training and Evaluation

Within each fold of the cross-validation procedure, features are standardized using `StandardScaler` to ensure consistent scaling across participants. A range of classification algorithms are then trained and evaluated to compare their performance on the binary emotion recognition tasks. These include Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Logistic Regression (LR), Decision Tree (DT). In addition to

these classical methods, a Deep Neural Network (DNN) is implemented using the Keras library and wrapped with `SciKeras` to integrate seamlessly into the scikit-learn pipeline. The DNN consists of three hidden layers with dropout regularization and employs early stopping based on validation loss to prevent overfitting.

For each classifier, performance is evaluated using standard classification metrics including accuracy, precision, recall, and F1-score. A confusion matrix and detailed classification report are also generated for each fold to provide insight into the classifier's strengths and weaknesses across the two classes.

5. Subset Selection and Best Model Identification

Given that multiple outer folds are created for each emotion through balanced sampling of the negative class, we identify the best-performing subset by comparing average F1-scores across the five inner folds. For each classifier-emotion pair, the outer fold that yields the highest average F1-score is selected as the most effective configuration. The confusion matrix from this best-performing outer fold is averaged across the inner folds and visualized as a heatmap, offering a clear depiction of classification performance.

4.3.3 Experiment 3: Baseline Reduction to Improve Personalization

Baseline reduction accounts for individual variability by removing the influence of personal baseline levels, which can differ significantly across subjects. By normalizing data relative to each individual's baseline, models can focus on the changes caused by emotional stimuli, enhancing their personalization and improving accuracy in emotion recognition tasks. Wirawan et al. [112] have used different baseline reduction methods in order to tackle individual differences while classifying emotions. From their study we use the following three methods and test their feasibility on our dataset 3.2:

1. **Difference Method:** This method computes the difference between the DE (Differential Entropy) feature values of the emotional video and the corresponding baseline values obtained from the neutral video at the same time point:

$$Final_{v_j}(x) = Emo_{v_j}(x) - Neut_{v_j}(x) \tag{4.2}$$

where $Emo_{v_j}(x)$ represents the DE feature value from the emotional video at time x, and $Neut_{v_j}(x)$ is the corresponding DE feature value from the neutral video. By

subtracting the baseline, this technique isolates the emotional responses, effectively reducing inter-individual variability and emphasizing the dynamic changes elicited by the stimulus.

2. **Relative Difference Method:** In this method, the DE feature value from the emotional video is normalized by dividing it by the corresponding DE feature value from the neutral video:

$$Final_{v_j}(x) = \frac{Emo_{v_j}(x)}{Neut_{v_j}(x)} \tag{4.3}$$

   Here, $Emo_{v_j}(x)$ is the DE feature value from the emotional video at time x, and $Neut_{v_j}(x)$ is the baseline value from the neutral video. This technique captures the relative change in feature values, making it robust to variations in individual baseline magnitudes while highlighting proportional changes due to emotional stimuli.

3. **Fractional Difference method**: This method combines the Difference and Relative Difference methods. The DE feature in the emotional video is subtracted from the corresponding DE feature in the neutral video at the same time, and the result is then divided by the value of the DE feature in the neutral video:

$$Final_{v_j}(x) = \frac{Emo_{v_j}(x) - Neut_{v_j}(x)}{Neut_{v_j}(x)} \tag{4.4}$$

   where $Emo_{v_j}(x)$ is the value of the DE feature from the emotional video and $Neut_{v_j}(x)$ is the corresponding value from the neutral video.

Importance of Baseline Reduction

Baseline reduction plays a crucial role in personalizing machine learning models, particularly for emotion recognition. Individual differences in physiological signals, such as heart rate or skin conductance, can mask the subtle changes triggered by emotional stimuli [113]. By removing these baseline variations, the model can focus on the meaningful fluctuations that reflect emotional states, resulting in:

- Improved generalization across diverse subjects.
- Enhanced sensitivity to subtle emotional responses.
- Reduction of inter-subject variability, making models more robust and interpretable.

- A foundation for developing personalized emotion recognition systems tailored to individual physiological profiles.

This ensures that the features used for training and evaluation reflect true emotional responses rather than static individual differences, thus enhancing the accuracy and personalization of the emotion recognition model.

Model Training

The process of training the model begins with feature extraction from the data frames. In experiments involving baseline reduction, the features corresponding to the emotional videos are normalized using the ratio, difference, or fractional difference methods relative to the features from the neutral videos. This step ensures that the final dataset reflects meaningful changes associated with emotional stimuli, minimizing the influence of individual baseline variability.

Once the final feature dataset is prepared, we conduct three types of experiments based on the labeling scheme used, binary arousal, binary valence, and multiclass classification.

4.3.4 Experiment 4: Fusion of Physiological Signals with Speech Features

This experiment evaluates the effectiveness of multimodal fusion by integrating handcrafted physiological features with speech-derived embeddings for emotion classification. The speech data were collected during the emotional recall phase and processed through a structured transcription and embedding pipeline.

Audio responses were first transcribed using the `openai/whisper-large-v3` [114, 115] automatic speech recognition model. The resulting text transcripts were then encoded into 384-dimensional sentence embeddings using the `sentence-transformers/all-MiniLM-L6-v2` [116] model, which captures both semantic and contextual information. To ensure compact and discriminative representations, dimensionality reduction was performed using Uniform Manifold Approximation and Projection (UMAP) configured with `n_neighbors = 15`, `min_dist = 0.1`, `n_components = 10`, and `metric = cosine`. The UMAP transformation was fit exclusively on the training participants within each fold and later applied to the validation and test participants to prevent representation leakage.

The reduced 10-dimensional speech embeddings were concatenated with handcrafted physiological features derived from ECG, EDA, and respiration signals to form a unified

multimodal feature vector. These fused representations were used to train a Random Forest classifier for three classification tasks: binary arousal, binary valence, and multiclass emotion recognition.

A subject-independent evaluation protocol was strictly followed. Participants were split into disjoint training and test sets (80%/20%), ensuring no subject overlap. On the training set, `StratifiedGroupKFold` cross-validation was employed to maintain class balance while isolating participants across folds. Hyperparameter optimization was conducted within cross-validation using `GridSearchCV` with weighted F1-score as the selection criterion. The optimized parameters were then fixed and used for final evaluation on the held-out test set.

Model performance was evaluated using accuracy, precision, recall, and F1-score, along with confusion matrices and fold-wise performance plots. This fusion approach combines complementary information from speech and physiological signals to improve the overall accuracy and consistency of emotion recognition across participants.

4.3.5 Experiment 5: Subject-wise Analysis

This experiment evaluates model performance at the participant level using multiclass emotion classification. The analysis is conducted using three classifiers: Random Forest, GLU, and TabNet.

All participant feature files were combined into a single dataset, excluding neutral videos. The dominant emotion for each video was identified using the video filename. The `StratifiedGroupKFold` method with five splits was employed to ensure that data from any single participant appeared in only one fold, preventing data leakage and enabling subject-independent evaluation.

For each fold, the model was trained on data from a subset of participants and tested on unseen participants. Predictions were generated for each video segment of a participant, and the accuracy and macro F1 score were computed by comparing predicted labels to the intended video emotion. Additional metrics, including true positives, false negatives, and the most frequently predicted label per video, were also recorded. These results were aggregated into a summary table to compare classifier performance on a per-participant basis.

Additionally, a video-wise misclassification table is generated to identify patterns in incorrect predictions, revealing which emotions tend to be confused with others across models.

4.4 Results and Discussion

4.4.1 Evaluation Metrics

To assess the performance of our emotion classification framework, we relied on standard evaluation metrics widely used in affective computing research, including **accuracy** and **F1 score** [96, 117–119]. These metrics were computed based on the number of *true positives (TP)*, *true negatives (TN)*, *false positives (FP)*, and *false negatives (FN)* obtained during classification.

In our setup, the high arousal class was treated as the positive class, while the low arousal class was considered negative. The **accuracy** metric indicates the proportion of correctly classified samples out of the total samples and is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.5}$$

To capture both the precision and the recall of the model, we use the **F1 score**, which is the harmonic mean of precision and recall. These intermediate metrics are computed as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.6}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.7}$$

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.8}$$

These metrics were calculated for each fold in the k-fold group cross-validation procedure and averaged to provide a robust evaluation of the model's generalization performance.

4.4.2 Performance Evaluation

Experiment 1: Emotions Classification using Machine Learning Classifiers

In this experiment, the performance of various classical machine learning models on classifying emotional states—binary valence, binary arousal, and multiclass emotions—using

both 30s overlapping segments and trigger-based segments. Seven classifiers were evaluated: Decision Trees (DT), Gradient Boosting (GB), Perceptron, Logistic Regression, Random Forest (RF), GLU, and TabNet. The performance was assessed using accuracy and F1-score, averaged over all folds of participant-independent cross-validation.

The classification performance for all models and emotion types across both segmentation strategies is summarized in the Table 4.3. Overall, the results indicate that ensemble methods and deep learning models significantly outperform simpler classifiers. Specifically, Random Forest, TabNet, and GLU consistently emerged as the top performers across most tasks. In the binary valence classification using trigger-based segments, Random Forest achieved the highest accuracy and F1-score (0.69 and 0.67, respectively). Following closely, TabNet (F1 score - 0.64) and Gradient Boosting (F1 score - 0.63) also showed strong, competitive results. In the case of arousal classification on trigger segments, the top accuracy (0.60) was shared by both TabNet and GLU, followed closely by other ensemble models. Most classifiers exhibited similar F1 scores, ranging between 0.58 and 0.61, indicating relatively balanced precision-recall trade-offs.

The 30s segments consistently yielded lower accuracy and F1 scores across all models for both arousal and valence classification, as shown in Table 4.3 and Figure 4.1. This discrepancy highlights the importance of using manually annotated trigger points, which indicate the precise onset of emotion elicitation during the video stimuli. Unlike fixed-length 30-second segments, which may include portions of the signal that are not emotionally salient or are contaminated with noise (e.g., before the emotion is actually evoked), trigger-based segments focus specifically on the time windows most representative of the emotional response. As a result, these segments provide cleaner and more meaningful input to the classifiers, leading to improved performance and more reliable emotion recognition.

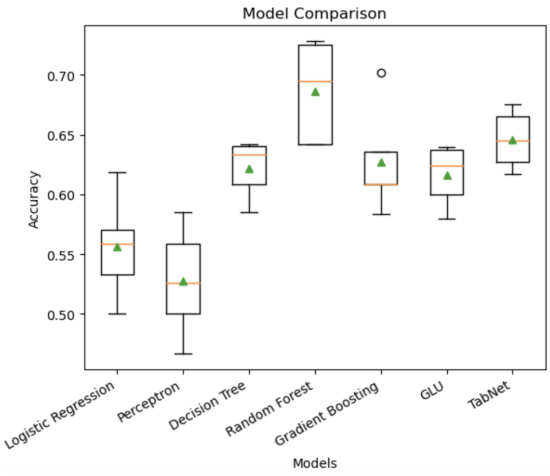
Figure 4.2 presents the confusion matrices and fold-wise performance scores (accuracy, precision, recall, and F1-score) for the Random Forest classifier across three classification tasks: binary valence, binary arousal, and multiclass emotion classification using trigger-based segments. In the binary valence classification (Column 1), the model achieved relatively high accuracy, with 219 low-valence and 187 high-valence samples correctly classified. The confusion matrix shows a balanced distribution of predictions, and the fold-wise performance metrics remain consistent with accuracy and F1-scores generally above 0.60.

For binary arousal classification (Column 2), the model correctly predicted 173 low-arousal and 181 high-arousal instances. Although the confusion matrix reveals slightly higher confusion between classes compared to valence, the overall performance remained stable with comparable fold-wise accuracy and F1 scores.

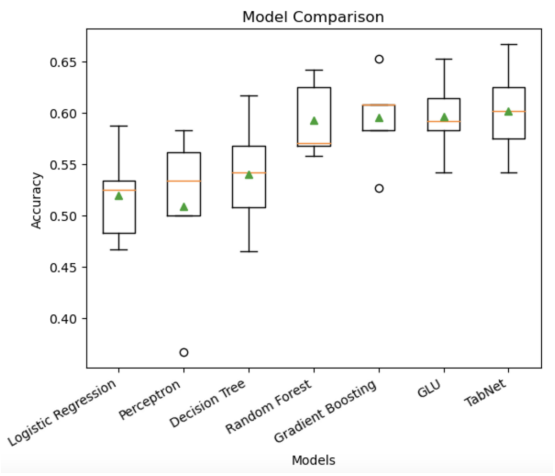
For the **multiclass emotion classification task**, the performance of the model was

Table 4.3: Performance Comparison for Different Classifiers Across Segment Types and Emotion Labels

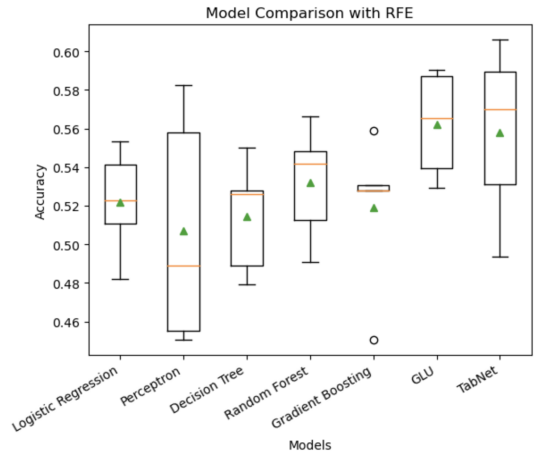
Classifier	**Trigger Segments**				**30s Segments**			
	Valence		**Arousal**		**Valence**		**Arousal**	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Decision Trees	0.60	0.60	0.54	0.54	0.51	0.53	0.52	0.52
Gradient Boosting	0.64	0.63	0.60	0.60	0.52	0.52	0.51	0.50
Perceptron	0.53	0.54	0.51	0.48	0.51	0.52	0.49	0.48
Logistic Regression	0.56	0.55	0.52	0.51	0.52	0.52	0.53	0.47
Random Forest	**0.69**	**0.67**	0.59	0.59	0.53	0.53	0.52	0.50
GLU	0.62	0.60	0.60	0.58	**0.56**	**0.55**	0.54	0.53
TabNet	0.65	0.64	**0.60**	**0.61**	0.56	0.54	**0.54**	**0.50**



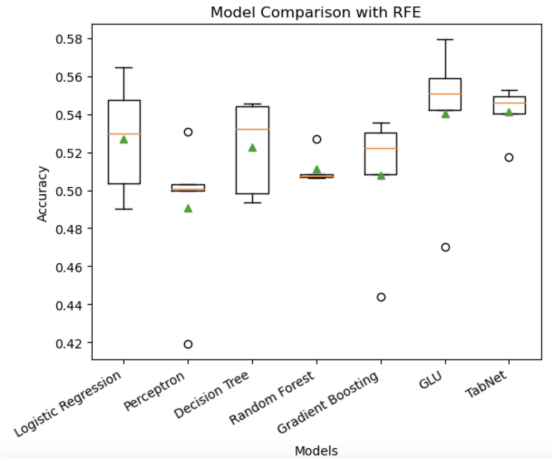
(a) Trigger Segments - Binary Valence



(b) Trigger Segments - Binary Arousal

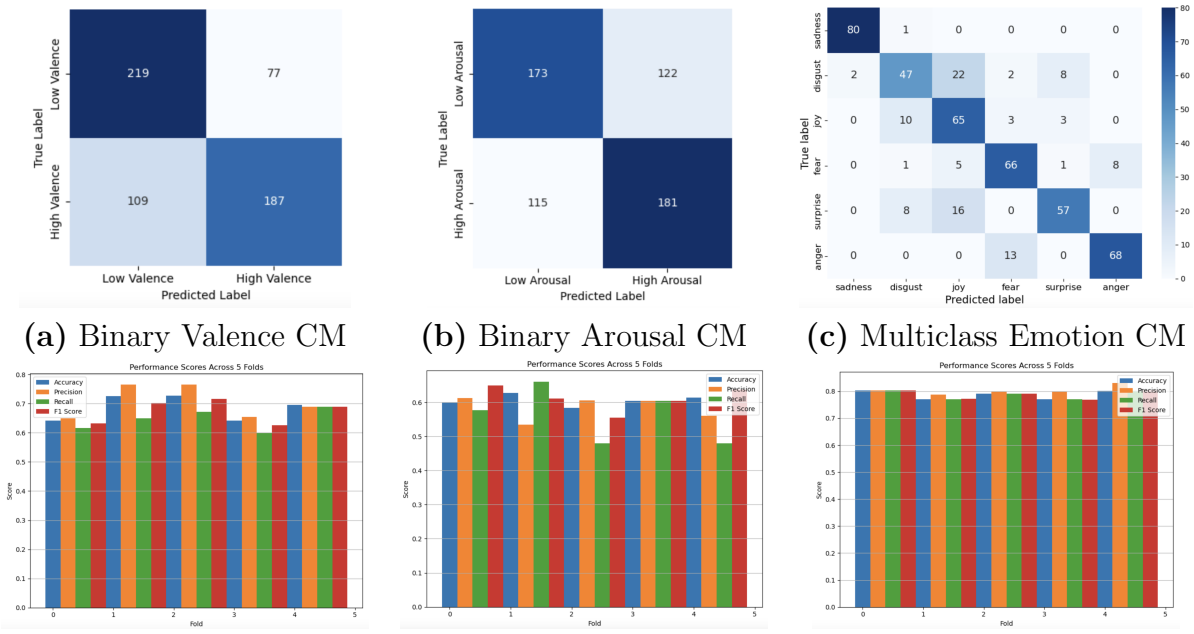


(c) 30s Segments - Binary Valence



(d) 30s Segments - Binary Arousal

Figure 4.1: Model-wise accuracy comparison for binary valence and arousal classification using different input segment types.



(a) Binary Valence CM **(b)** Binary Arousal CM **(c)** Multiclass Emotion CM

(d) Binary Valence **(e)** Binary Arousal **(f)** Multiclass Emotion

Figure 4.2: Confusion matrices (top row) and fold-wise performance metrics (bottom row) for binary valence, binary arousal, and multiclass emotion classification using Random Forest Classifier on trigger segments.

Table 4.4: Comparison of multiclass emotion classification performance using different labelling strategies.

Labelling Strategy	**Metrics**	
	Accuracy	**F1 Score**
Participant-Reported Labels	0.58	0.57
Intended Video Emotion Labels	**0.79**	**0.79**

evaluated using two different labeling schemes: (i) the **participant-reported emotion labels**, which represent the emotions self-reported by participants after watching each emotional video, and (ii) the **intended video emotion labels**, which correspond to the target emotion categories defined during the experimental design. A comparative summary of results for these two labeling schemes is presented in Table 4.4.

The classifier achieved higher accuracy and F1-score when trained and evaluated using the intended video emotion labels (accuracy = 0.79, F1 = 0.79) compared to the

participant-reported labels (accuracy = 0.58, F1 = 0.57). The superior performance with intended labels indicates that the video stimuli were effective in eliciting distinct and consistent emotional responses across participants, resulting in clearer physiological patterns for classification.

In contrast, the lower performance observed with participant-reported labels highlights potential inconsistencies in self-reported emotional states. These discrepancies may stem from individual differences in emotional perception, self-awareness, or interpretation of the stimuli. These findings suggest that participant-reported labels may not accurately reflect the emotions actually experienced by participants.

The confusion matrix in Figure 4.2c for the multiclass emotion classification corresponds to the intended video emotion labels. The overall performance of the multiclass emotion classifier, evaluated on trigger-based segments using the intended video emotions as labels, yielded an average **accuracy of 78.8%, precision of 80.4%, recall of 78.8%, and F1 score of 78.8%**. Hyperparameter optimization for the Random Forest classifier was conducted using `GridSearchCV`, yielding the best configuration with $n\_estimators = 100$, $max\_depth = Unlimited$, $min\_samples\_split = 2$, and $min\_samples\_leaf = 2$.

Among all the emotions, *sadness* and *anger* were the most accurately classified, with 80 and 68 correct predictions, respectively. These results suggest that these emotions exhibit more distinct physiological patterns, which are effectively captured by the model. In contrast, *disgust* showed more confusion, with frequent misclassifications as *joy* and *surprise*, possibly due to overlapping signal characteristics or subtle expression differences.

Moreover, *surprise* was often misclassified as *joy* (16 cases) and *disgust* (8 cases), indicating difficulties in distinguishing between high-arousal emotional states that may share similar physiological responses.

In summary, while the classifier shows strong performance in recognizing emotions such as sadness, anger, and fear, it encounters greater difficulty distinguishing emotions like disgust and surprise. These misclassifications highlight the overlap in physiological responses among certain emotional states and underscore the challenges of achieving consistent emotion recognition across all emotions.

Experiment 2: Binary Emotions Classification

This binary classification experiment assessed whether physiological features from ECG, EDA, and respiration could distinguish between the presence or absence of individual emotional states. Binary labels were assigned based on a threshold (≥ 4) applied to

Table 4.5: Binary Emotion Classification Results across Classifiers: Accuracy and F1 Score

Emotion	DNN		LDA		Random Forest		SVM		kNN	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Anger	0.73	0.72	0.70	0.69	**0.81**	**0.80**	0.74	0.72	0.70	0.68
Disgust	0.62	0.62	0.64	0.61	**0.77**	**0.77**	0.57	0.53	0.61	0.61
Fear	0.63	0.59	0.62	0.54	**0.69**	**0.63**	0.67	0.55	0.62	0.53
Joy	0.68	0.63	0.67	0.63	**0.79**	**0.77**	0.64	0.47	0.65	0.62
Sadness	0.61	0.59	0.61	0.61	**0.69**	**0.68**	0.58	0.53	0.62	0.56
Surprise	0.60	0.59	0.61	0.59	**0.65**	**0.61**	0.59	0.43	0.59	0.52

continuous emotion ratings for six target emotions: anger, disgust, fear, joy, sadness, and surprise.

Table 4.5 presents the performance of several classifiers: Random Forest (RF), Deep Neural Network (DNN), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), and k-Nearest Neighbors (kNN). Among these, RF consistently delivered the highest accuracy and F1 scores across all emotions. For instance, RF achieved an F1 score of 0.80 for anger, 0.77 for both joy and disgust, and maintained solid performance on sadness and surprise—emotions that are typically harder to classify due to subtle physiological expressions.

DNN and LDA demonstrated moderate performance, particularly for anger and joy. However, their performance varied considerably across emotion categories. DNN exhibited lower discriminative ability for sadness (F1-score = 0.59), indicating limited sensitivity to low-arousal emotional states. SVM and kNN generally showed lower overall performance, especially for surprise and fear, where F1-scores dropped below 0.55, reflecting their sensitivity to class imbalance and reduced robustness across participants. Figure 4.3 and Table 4.5 illustrate these classifier-specific differences, where confusion matrices reveal strong separability for anger, but notable overlap between emotion-present and emotion-absent samples for surprise and fear, even under binary labeling. High false-negative rates observed for fear further suggest weaker generalization for low-arousal or ambiguous emotions. These results are consistent with the trends observed in the multiclass classification analysis (Figure 4.2).

The bottom row of the figures presents metric-wise comparisons across classifiers. RF consistently balances precision, recall, and F1, while other models—especially SVM and LDA—often exhibit trade-offs. For example, SVM achieves high precision but suffers in

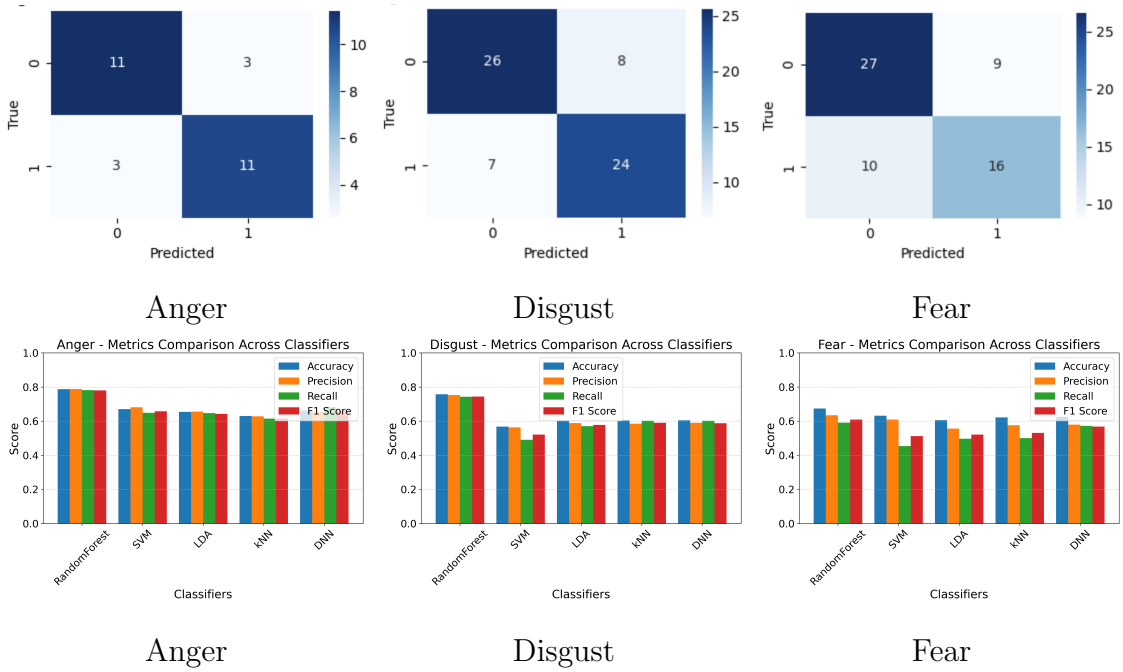


Figure 4.3: Binary classification performance for emotions: Anger, Disgust, and Fear. Top: Confusion matrices from the best-performing classifier (Random Forest). Bottom: Metric-wise classifier comparison, in the order RF, SVM, LDA, KNN, DNN.

recall, reflecting conservative classification thresholds.

Overall, binary classification offers a promising alternative to multiclass setups for physiological emotion recognition. It improves the detection of emotions like disgust but also offers clearer interpretability through class-focused training. These results support the integration of emotion-specific binary models into broader emotion recognition systems.

Experiment 3: Baseline Reduction Methods

The baseline reduction techniques described in Section 4.3.3 were applied to all three classification tasks using the Random Forest classifier. The evaluation was performed on physiological signals segmented into overlapping 30-second windows, with a 10-second overlap 4.7, and when the segments were divided on the basis of emotional triggers 4.6.

Among the three methods-Difference, Relative Difference, and Fractional Difference—the Fractional Difference method consistently provided the highest accuracies for all classifica-

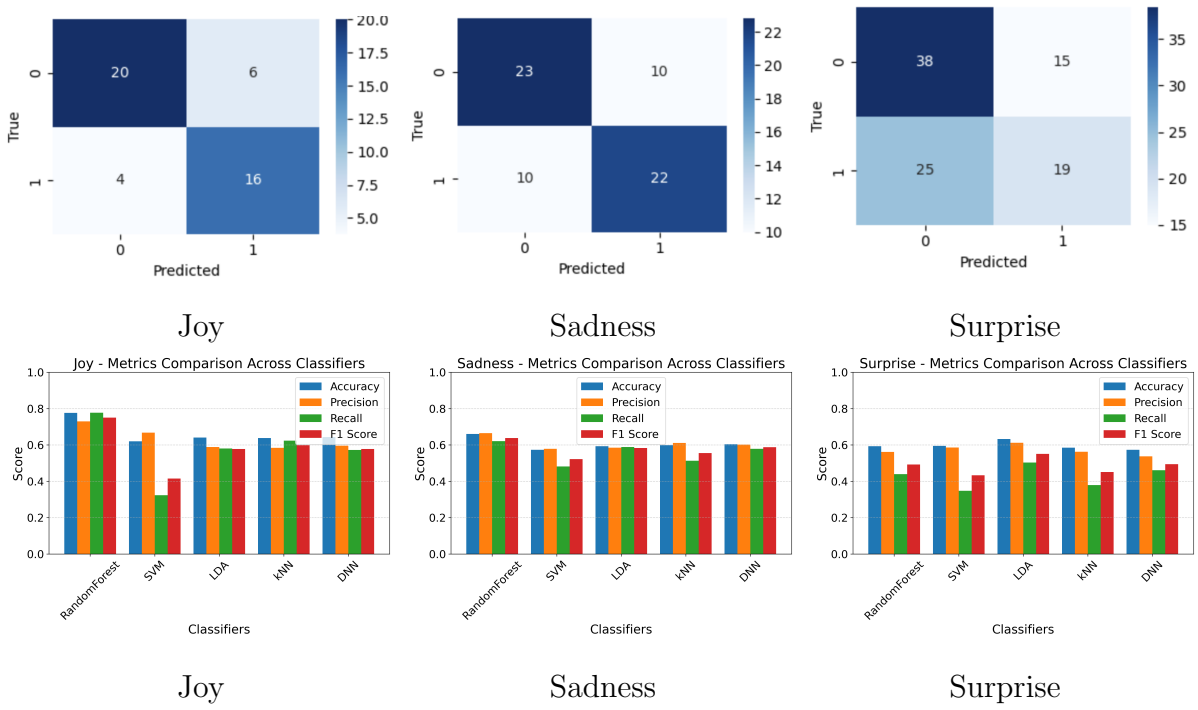


Figure 4.4: Binary classification performance for emotions: Joy, Sadness, and Surprise. Top: Confusion matrices from the best-performing classifier (Random Forest). Bottom: Metric-wise classifier comparison, in the order RF, SVM, LDA, KNN, DNN.

Table 4.6: Comparison of accuracies generated using Random Forest model, across different methods of baseline reduction when the data is selected according to the emotional triggers.

	Binary Arousal	**Binary Valence**	**Multi-class Classification**
Difference	52.86	54.15	25.96
Relative Difference	50.54	52.95	24.73
Fractional Difference	53.05	57.93	32.67

tion tasks when emotional triggers were used as starting points. This method achieved an accuracy of 57.93% for binary valence and 32.67% for multi-class classification. However, applying baseline correction using the three difference methods resulted in a consistent decrease in classification accuracy across all tasks. These results indicate that the implemented baseline correction techniques may not effectively remove the baseline physiological patterns, and might even discard emotionally informative components of the signal. Rather

Table 4.7: Comparison of accuracies generated using Random Forest model, across different methods of baseline reduction when the data is segmented into windows of 30 seconds with 10 seconds overlap for each emotional video.

	Binary Arousal	**Binary Valence**	**Multi-class Classification**
Difference	48.10	47.41	22.76
Relative Difference	48.75	49.79	22.00
Fractional Difference	48.47	50.94	21.22

than improving model performance, they seem to diminish the discriminative power of the features. This suggests that the current baseline removal methods may be insufficient or overly simplistic for physiological emotion recognition, and further research is required to develop more effective and signal-specific baseline correction strategies.

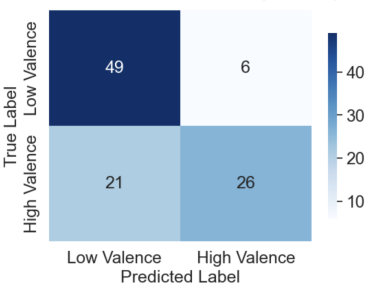
Experiment 4: Fusion of Physiological Signals with Speech Features

The speech data is recorded after every emotional video. The speech is then transcribed into text and UMAP (Uniform Manifold Approximation and Projection) embeddings are generated for the text. The UMAP embeddings extracted for the speech data are concatenated at the end of each frame of the corresponding emotional video to carry out feature-level fusion.

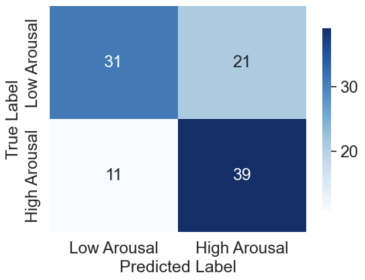
Table 4.8: Performance (Accuracy and F1 Score) of Random Forest classifier on the Test set using early fusion of speech and physiological features.

Segmentation	**Valence**		**Arousal**		**Multiclass**	
	Acc.	F1	Acc.	F1	Acc.	F1
Trigger Segments	0.73	0.68	0.64	0.67	0.97	0.97
30s Segments	0.65	0.63	0.63	0.62	0.92	0.92

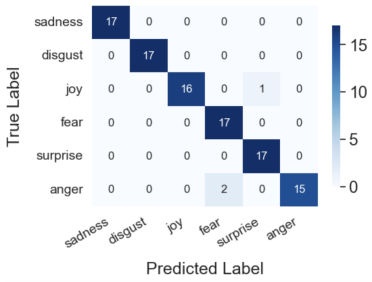
The results in Table 4.8 show the performance of the Random Forest classifier when speech and physiological features are combined using early fusion. The models were trained and evaluated for binary classification (valence and arousal) and multiclass emotion classification using two segmentation strategies: emotionally triggered segments and fixed 30-second segments. The early fusion of speech features (UMAP embeddings) with physiological signals leads to a notable uplift in classification performance across all tasks.



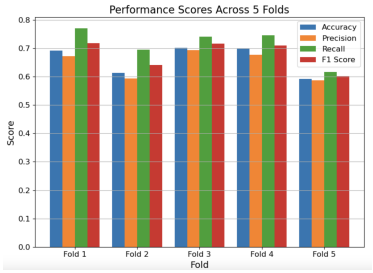
(a) Binary Valence CM



(b) Binary Arousal CM

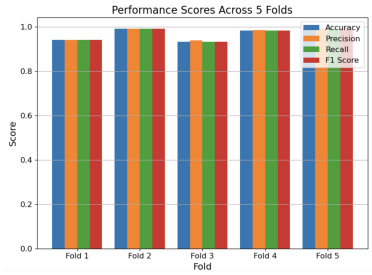


(c) Multiclass Emotion CM



(d) Binary Valence Scores

(e) Binary Arousal Scores



(f) Multiclass Emotion Scores

Figure 4.5: Confusion matrices (top row) and fold-wise performance metrics (bottom row) for binary valence, binary arousal, and multiclass emotion classification using Random Forest Classifier with early fusion of speech and physiological features on trigger segments.

Combining UMAP embeddings of speech features with physiological signals significantly improved classification accuracy, especially for multi-class classification, which achieved 97% accuracy for emotional trigger based segments.

This improvement in accuracy can be attributed to the superior performance of speech features compared to physiological features. By combining the two, the influence of the speech features enhanced the overall fusion accuracy.

Since trigger-based segments yielded better overall performance across all tasks, we analyze the confusion matrices and k-fold cross-validation results for this segmentation in Figure 4.5.

- Arousal and Valence: The 5-fold bar graph shows consistency with F1 scores peaking at 0.74, confirming the model's generalizability. It seems that arousal is inherently harder to capture from physiological signals, yet the early fusion approach helps in slightly elevating the accuracy. On the other hand, valence has been more predictable from physiological signals even in the precious experiments. In this case, low valence

instances (TP = 212) are better classified compared to high valence (TP = 179).

- Multiclass: The performance of multi-class classification improved significantly with the fusion of speech and physiological features, indicating that discrete emotional states may have become more distinguishable due to the UMAP word embeddings compared to dimensional labels in the given dataset.

Experiment 5: Subject-wise Results Analysis

While previous experiments reported average classification metrics across the dataset, this experiment focuses on evaluating model performance at the individual participant level. Such analysis offers deeper insights into how well each model generalizes to unseen subjects and highlights consistency across participant-specific physiological patterns. For this purpose, we conducted participant-level inference across a 5-fold stratified group cross-validation setup. In each fold, the model was trained on a subset of participants and tested on unseen participants. The results of three classifiers—TabNet, GLU, and Random Forest—are evaluated. This procedure ensures that the predictions for every participant are out-of-fold and unbiased, i.e., made by a model that has never seen their data during training. The final combined results consist of 486 such "clean" predictions—each corresponding to a test sample made with no data leakage from the same participant.

The detailed participant-wise prediction tables are provided in Appendix B, showing the predicted labels per emotion category (anger, fear, disgust, joy, sadness, surprise) for each participant, and highlighting the misclassified cases in 'red'. In order to summarize those results, we have created the confusion matrices (see Figure 4.6) by combining these out-of-fold prediction. Each represents a true prediction on a held-out participant. These matrices reflect true subject-level generalization without any re-evaluation or aggregation over repeated samples.

From the results in the Appendix B, the overall mean accuracy for each model is calculated by computing the proportion of correctly predicted emotions across all participants. **Random Forest achieved the highest mean accuracy (0.788), followed by GLU (0.586) and TabNet (0.537)**.

Figure 4.6 shows the confusion matrices for the three models TabNet, GLU, and Random Forest based on these participant-wise predictions. These models were chosen based on their results in the Experiment 1. Among the three models, Random Forest exhibits the clearest diagonal structure, indicating stronger discriminative performance across most emotion classes.

(a) TabNet (b) GLU (c) Random Forest

Figure 4.6: Participant-wise confusion matrices for emotion classification using TabNet, GLU, and Random Forest. Each matrix is built using out-of-fold predictions for participants unseen during training.

Table 4.9: Video-wise Emotion Misclassifications by Each Classifier

Video	**TabNet**	**GLU**	**Random Forest**
1anger.mp4	[joy, fear, surprise]	[fear]	[fear]
1disgust.mp4	[sadness, fear, joy]	[sadness, joy]	[sadness]
1fear.mp4	[disgust, joy, anger, sadness, surprise]	[surprise, disgust, joy, anger]	[disgust, joy, anger, surprise]
1joy.mp4	[disgust, fear, sadness, anger]	[disgust, sadness]	[]
1sad.mp4	[disgust, surprise]	[surprise]	[]
1surprise.mp4	[joy, fear, anger, disgust, sadness]	[fear, joy, disgust, anger]	[joy, disgust]
2anger.mp4	[fear, sadness, joy]	[fear]	[fear]
2disgust.mp4	[joy, fear, surprise, anger, sadness]	[joy, fear, surprise]	[joy, surprise, fear]
2fear.mp4	[anger, joy, sadness, surprise]	[anger, disgust, surprise]	[anger]
2joy.mp4	[disgust, fear, sadness, anger, surprise]	[disgust, surprise, fear, anger]	[disgust, fear, surprise]
2sad.mp4	[disgust, joy]	[disgust, joy]	[disgust]
2surprise.mp4	[sadness]	[sadness]	[]

In addition to the aggregated results in the confusion matrices, Table 4.9 presents a finer-grained video-wise analysis of misclassified emotions. This table lists the specific emotions each model predicted incorrectly for each of the 12 emotional videos. Along with this, the Table 4.10 showcases the accuracy results for each emotion category as predicted by the three models.

Following are the observations drawn from these results:-

- **Disgust** remains the most error-prone emotion across models and videos. In `1disgust.mp4`, all models misclassified the emotion, commonly predicting *sadness*

Table 4.10: Emotion-wise Mean Accuracy for Each Classifier

Emotion	TabNet	GLU	Random Forest
Anger	0.728	0.765	**0.840**
Disgust	0.235	0.272	**0.580**
Fear	0.605	0.556	**0.815**
Joy	0.469	0.519	**0.802**
Sadness	0.778	0.802	**0.988**
Surprise	0.407	0.605	**0.704**

and *joy*. The video eliciting disgust shows consistent stimuli from beginning to end. Similarly, emotions like sadness and joy are also characterized by relatively low variability. It is possible that the physiological pattern of disgust overlaps with these low-variability emotions, leading to confusion. This observation aligns with its poor performance in the confusion matrices (Figure 4.6), where disgust was frequently confused with high-valence or low-arousal emotions such as *joy* and *sadness*.

- In `1surprise.mp4`, the surprise event occurs only towards the end of the video (approximately 7 seconds before it ends). Prior to this moment, there are no notable surprise-inducing stimuli, which may explain the absence of corresponding physiological responses. Consequently, the signals from this video often result in misclassification. In contrast, `2surprise.mp4` contains multiple high-arousal surprise events throughout its duration, resulting in zero misclassifications with Random Forest model. This suggests that the relatively poor classification accuracy for the *surprise* emotion is primarily attributable to `1surprise.mp4`.

- The videos targeting *joy* achieve a respectable accuracy of approximately 80%, yet this is still lower compared to the near-perfect accuracy for *sadness* (99%). Interestingly, *joy* is most frequently misclassified as *disgust* (22 times), indicating possible confusion between the physiological signals of these two emotions.

- Videos designed to elicit **fear** (e.g., `1fear.mp4` and `2fear.mp4`) exhibit widespread misclassification across all three models. For instance, `1fear.mp4` is frequently misclassified as *disgust*, *joy*, *anger*, and *surprise*. As a high-arousal emotion, fear shares physiological characteristics with other high-arousal emotions like anger and surprise, which likely contributes to the observed confusion. This is also evident from the confusion matrices, where *fear* is most often misclassified as *anger*.

- *Anger* is predominantly misclassified as *fear* in both corresponding videos. This likely results from the similarity in physiological responses associated with these two high-intensity emotions. Nevertheless, anger still achieves the second-highest prediction accuracy (84%) with the Random Forest classifier. In contrast, *sadness*, characterized by low arousal and low valence, demonstrates minimal physiological variation, making it the most reliably classified emotion (with an accuracy of 99%) across all models.

- The **Random Forest** model exhibits the most consistent and focused predictions. It shows fewer misclassifications per video and even achieves perfect predictions (i.e., no errors) for certain cases such as `1joy.mp4`, `1sad.mp4`, and `2surprise.mp4`. These results are consistent with its superior performance observed in both the confusion matrices and the mean accuracy metrics.

- In contrast, the **TabNet** model tends to exhibit broader misclassification across a wider range of emotion categories. For example, in `2joy.mp4` and `1fear.mp4`, TabNet misclassifies five different emotion categories, indicating weaker class separability and lower discriminative power in its predictions.

4.5 Observations

This chapter investigated several experiments to classify human emotions using handcrafted features derived from ECG, EDA, and RSP signals. The core methodology involved data preprocessing, the extraction of 44 features across time, frequency, and non-linear domains, and a systematic evaluation of several classical machine learning classifiers, along with deep learning classifiers. The research was designed to answer several key questions through different experiments, including the best way to segment the data, the effectiveness of different classification tasks, whether adding speech data could improve results, and subject-independent model classification performance. The findings provide critical insights into the capabilities and limitations of using physiological signals for emotion recognition.

Analysis of Experimental Findings

Experiment 1: General Emotion Classification and the Importance of Segmentation Strategy

A key finding from the first experiment was that the segmentation strategy plays a crucial role in determining model performance. The analysis demonstrated that using emotion-triggered segments—data windows aligned with the onset of emotion-inducing moments in the video—produces more accurate and reliable results than using fixed-length overlapping 30-second windows. For instance, the Random Forest classifier achieved notably higher accuracy in valence classification when trained on these trigger-based segments. This finding highlights that focusing on the most emotionally relevant portions of the physiological signals improves the model's ability to detect emotional states.

In addition, when comparing multiclass classification using *intended video emotion* labels versus *participant-reported emotion* labels, the model performed substantially better with the intended labels. This outcome indicates that physiological responses are more consistent with the emotions intended by the stimuli, suggesting that the emotional videos effectively elicited distinct and measurable physiological reactions in participants. In contrast, participant-reported labels introduced greater variability, likely due to inconsistencies in how individuals interpreted or reported their emotional experiences.

Threats to Validation

From the results obtained in Experiment 1, discrepancies were observed in the participant-reported emotion labels. The discrete self-reported labels were collected using a 7-point Likert scale ranging from "strongly disagree" to "strongly agree". Such subjective ratings can introduce several forms of bias or uncertainty. Participants may have struggled to accurately identify or articulate their emotional states, leading to inconsistent or ambiguous responses. Additionally, some participants may have misunderstood the questions or provided socially desirable answers, aligning their responses with what they believed the experimenters expected.

A core reason for these inconsistencies is that emotion formation itself is a complex cognitive appraisal process, shaped by interpretation, context, and prior experience [120]. Classical appraisal theories such as Schachter et al.'s two-factor model [121] and Mandler's extensions [122] emphasize that emotional experience arises from interpreting physiological arousal through cognitive evaluation. Because this appraisal process is subjective and varies across individuals, two people with similar physiological responses may self-report different

emotional states. Yang et al. further emphasize this point, noting that self-assessed labels are vulnerable to individual biases related to personal experience, cognitive ability, and situational factors [120]. Their analysis also shows that emotional reactions can occur at an unconscious level and therefore remain absent from self-reports despite being clearly observable in physiology [123].

Moreover, self-reports are influenced by limited introspective access, as individuals may struggle to accurately interpret or describe their internal emotional states, leading to imprecise or overly generalized ratings [124]. Social-desirability pressures can further shape these responses, with participants adjusting their answers to match perceived expectations or norms rather than reporting their genuine emotions [125].

These factors suggest that self-reported emotional data may not always provide a fully reliable measure of experienced emotion. This limitation is not unique to the current dataset; it is a common challenge in psychological and affective computing studies that rely on introspective reporting [126]. Such inconsistencies represent potential threats to the validation of emotional responses and must be considered when interpreting results.

Experiment 2: Binary (One-vs-Rest) Emotion Classification

By reframing the problem from a single multiclass task to six independent binary tasks, the study revealed which emotions have the most distinct physiological signatures. This approach proved highly effective for emotions that were otherwise poorly classified. For instance, **anger** achieved a remarkable 81% accuracy in its binary classification task using a Random Forest model. This suggests anger has a unique physiological pattern that is easily masked by or confused with other high-arousal emotions like fear in a competitive multiclass scenario.

Experiment 3: The Challenge of Baseline Reduction

Interestingly, the attempt to personalize the data by removing each individual's baseline physiological state did not work as expected. All three baseline correction methods—Difference, Relative Difference, and Fractional Difference—that were tested led to a **decrease in classification accuracy** across all tasks. This finding suggests that the implemented methods, while theoretically sound for reducing inter-subject variability, were likely too simplistic. They might have accidentally removed important emotional information from the signals while trying to filter out noise. It shows that personalizing models is a complex challenge that requires more advanced techniques.

Experiment 4: Success Through Multimodal Fusion with Speech

The final experiment clearly showed the benefit of using more than one type of data. When features from the physiological signals were combined with numerical embeddings created from participants' spoken responses, the performance of the classifiers improved dramatically. The most impressive result was in the multiclass emotion task, where accuracy jumped from 78% to 97%. This strongly indicates that the sentence embeddings from speech data contain rich information complementing the emotion discrimination that helps resolve the ambiguities inherent in physiological signals alone, leading to a more accurate assessment of a person's emotional state.

Experiment 5: Subject-Wise Analysis Reveals Model Nuances

The participant-level analysis provided the most granular insights into model behavior and emotion separability. By evaluating predictions on unseen subjects, it was possible to assess true generalization.

Classifier Performance. The **Random Forest** model consistently provided the most reliable results, achieving the highest mean accuracy of 78.8% and showing the fewest misclassifications per video. It even achieved perfect classification accuracy for some videos, underscoring its robustness. In contrast, **TabNet** exhibited broader misclassification patterns, indicating weaker class separability in this experiment.

Highly Separable Emotions. **Sadness** was the most reliably classified emotion, with the Random Forest model achieving near-perfect accuracy (99%). This high performance is likely due to its distinct physiological signature characterized by low arousal and low valence. **Anger** and **fear** also demonstrated strong separability, with the Random Forest model achieving classification accuracies of 84% and 81.5%, respectively.

Poorly Separable Emotions. **Disgust** proved to be the most challenging emotion to classify, frequently being confused with *joy* and *sadness* across all classifiers. The analysis suggests that disgust shares low-variability physiological patterns with other emotions, which may contribute to its high misclassification rate. The classification of **surprise** was found to be highly dependent on the nature of the stimulus video. A video containing only a single, delayed surprise event (`1surprise.mp4`) led to multiple misclassifications, while

another video featuring multiple surprise-inducing moments throughout (`2surprise.mp4`) was classified perfectly by the Random Forest model.

Final Conclusions

Based on the comprehensive experiments, clear patterns emerge regarding the separability of emotions and the effectiveness of different classifiers.

- **Emotion Separability:** Emotions such as *sadness*, *anger*, and *fear* displayed distinctive physiological patterns, resulting in higher classification accuracy. In contrast, emotions like *disgust* and *surprise* were more difficult to distinguish, likely due to overlapping physiological responses or variations in the timing and intensity of emotional cues.
- **Classifier Robustness in Multiclass Setting:** The **Random Forest** model consistently demonstrated the highest accuracy and the most stable performance in the multiclass classification experiments. It showed fewer misclassifications and achieved near-perfect accuracy for certain emotions, making it the most reliable choice for fine-grained emotion classification using physiological signals.
- **Binary Classification Performance:** In the case of binary arousal and valence classification tasks, both **TabNet** and **GLU** models achieved strong and comparable performance. Their ability to model complex nonlinear boundaries proved effective in distinguishing high vs. low arousal and positive vs. negative valence categories.
- **Impact of Stimuli Design:** The classification accuracy, particularly for emotions like *surprise*, was found to be highly dependent on the temporal design of the emotional stimuli. Videos with delayed or weak emotional triggers resulted in lower prediction accuracy, emphasizing the need for well-timed and clearly expressive stimuli in experimental setups.
- **Participant-Level Generalization:** This study adopted a strict stratified group cross-validation approach to ensure no participant's data was used during both training and testing. This provided an unbiased, realistic evaluation of how well each model generalizes to unseen individuals—an essential consideration for practical deployment.

Chapter 5

Conclusion and Future Works

5.1 Conclusion

This thesis presented a framework for multimodal emotion recognition using physiological signals, specifically Electrocardiogram (ECG), Electrodermal Activity (EDA), and Respiration (RSP). The primary objective of this work was to support the development and technical validation of a large-scale physiological dataset by establishing baseline models and experimental protocols for emotion classification. This was achieved through the creation of standardized processing pipelines and baseline benchmarks encompassing three major components: (i) a comprehensive signal quality assessment (SQA) and artifact removal procedure, (ii) systematic data organization and cleaning for reproducibility, and (iii) in-depth performance analyses across multiple modeling strategies and labeling approaches.

A major component of this work focused on implementing a structured and reproducible signal quality assessment (SQA) and artifact removal pipeline to ensure the reliability of physiological recordings. The proposed framework consolidates established quality metrics for ECG and respiration from the literature and applies them consistently across modalities. For the EDA signal, multiple quality indicators—such as range validation, the Amplitude Spike Index (ASI), and a measure of signal sharpness—were introduced to identify abrupt spikes, baseline drifts, and non-physiological patterns. This integrated approach enabled the systematic exclusion of participants with low-quality data, resulting in a cleaner and more trustworthy dataset that forms the foundation for all subsequent analyses and classification experiments.

The study systematically evaluated the performance of classical machine learning models, including Random Forest, Decision Trees, Gradient Boosting, and Logistic Regression, and deep learning models Multi Layer Perceptron, GLU, TabNet, on handcrafted features extracted from the cleaned physiological signals. A critical finding emerged from the comparison of two segmentation strategies: fixed-length 30-second overlapping windows versus emotion-triggered temporal segments. The results demonstrated that aligning signal segments with the precise onset of emotionally salient events, as identified by expert annotators, led to markedly superior classification performance. This highlights that focusing on the most emotionally relevant periods of physiological response provides cleaner and more discriminative input for classifiers, leading to improved accuracy in binary valence, binary arousal, and multiclass emotion classification tasks. Among the classical models, Random Forest consistently exhibited strong and reliable performance across multiclass classification settings. In the binary classification scenarios, TabNet, and GLU performed really well. In multiclass scenario, the comparison between participant-reported and intended video emotion labels revealed that models trained on intended labels achieved more consistent performance, suggesting greater alignment between the stimuli and the corresponding physiological responses. In contrast, self-reported labels introduced variability that affected the overall reliability of classification outcomes. These inconsistencies highlight a broader challenge in emotion research—ensuring that emotional annotations accurately represent the underlying affective experience. Such discrepancies, whether due to reporting bias, interpretation errors, or variations in emotional awareness, emphasize the need for caution when validating emotion datasets and interpreting model performance based on self-reported data.

While physiological signals alone demonstrated a reasonable capacity for distinguishing emotions, the detailed subject-wise analysis revealed important nuances. Emotions with distinct physiological signatures, such as the low-arousal state of *sadness* (99% accuracy) and the high-arousal states of *anger* (84% accuracy) and *fear* (81.5% accuracy), were classified most effectively. Conversely, *disgust* proved to be the most challenging emotion, frequently confused with *joy* and *sadness*, suggesting a significant overlap in their physiological manifestations. Furthermore, the analysis highlighted the critical impact of stimulus design, as the classification of surprise was highly dependent on the timing and intensity of the eliciting events in the video clips.

A dedicated binary emotion classification task was also introduced to assess the ability of physiological signals to distinguish the presence or absence of each of the six target emotions (anger, disgust, fear, joy, sadness, and surprise). The results from this experiment highlighted that certain emotions such as anger gave the best results 81% F1 score in the binary classification task, suggesting the unique physiological patterns that are easily

masked by or confused with other high-arousal emotions like fear in a competitive multiclass scenario. These results were consistent with what we received from the other experiments.

Another experiment involved fusing speech embeddings derived from participants' verbal responses, with physiological features. By concatenating UMAP embeddings of speech features with the handcrafted physiological features, we observed a significant and consistent uplift in classification performance across all tasks. Most notably, the multiclass emotion classification accuracy surged from 78% to 97% when speech and physiological features were combined using trigger-based segments. This compelling result strongly suggests that embeddings from speech carry complementary information to emotions discrimination, and their early fusion with physiological signals provides a more holistic and accurate representation of an individual's emotional state. The superior performance of speech features, when combined with physiological data, effectively compensated for the inherent ambiguities or limitations of physiological signals alone in differentiating subtle emotional nuances.

Conversely, the exploration of baseline reduction techniques (Difference, Relative Difference, and Fractional Difference methods) yielded unexpected results, leading to a decrease in classification accuracy. This suggests that the implemented baseline correction strategies, while intended to reduce inter-subject variability, may have inadvertently removed emotionally informative components or were overly simplistic for the complex nature of physiological responses to emotional stimuli.

5.2 Future Works

The findings and limitations of this thesis open several promising avenues for future research in multimodal emotion recognition. Addressing these directions will be crucial for advancing the field towards more accurate, robust, and deployable affective computing systems.

5.2.1 Validation and Extension of SQA

The SQA framework introduced in this work, particularly the EDA quality metrics, presents several directions for further validation and improvement:

- **Cross-Dataset Validation:** Verify the effectiveness of the SQA metrics and thresholds using external physiological emotion datasets to assess their generalizability across different populations, acquisition environments, and sensor configurations.

- **Comparative Evaluation:** Compare the performance of the proposed SQA pipeline with other established quality assessment frameworks to evaluate consistency and identify areas of improvement.

5.2.2 Model Validation and Ablation Studies Across Datasets

Future studies should aim to extend and validate the modeling framework proposed in this thesis by applying it to additional multimodal emotion datasets. Key directions include:

- **Cross-Dataset Modelling:** Reimplement the proposed preprocessing and classification pipelines on other publicly available datasets to examine the transferability of the developed methods and to validate the robustness of the curated dataset used in this study.
- **Ablation Studies:** Perform controlled ablation experiments to isolate the contribution of each pipeline component, such as segmentation strategy, baseline correction, or feature subset, to overall classification performance.

5.2.3 Enhancing Baseline Correction and Personalization

The observed decrease in classification accuracy with the applied baseline reduction techniques highlights a significant area for improvement. Future work should focus on developing more sophisticated and adaptive baseline correction strategies. This could involve:

- **Dynamic Baseline Tracking:** Instead of a single static baseline, explore methods that continuously track and adapt to a participant's changing physiological baseline throughout an experiment, accounting for factors like fatigue, environmental shifts, or habituation.
- **Personalized Normalization:** Investigate personalized normalization techniques that are tailored to individual physiological profiles, rather than applying a universal method. This might involve learning individual-specific transformation functions.
- **Contextual Baseline Adjustment:** Incorporate contextual information (e.g., time of day, prior activity, self-reported mood before the experiment) into baseline adjustment models to better account for non-emotional physiological fluctuations.

5.2.4 Improving Discriminability of Confused Emotions

The subject-wise analysis revealed persistent confusion between certain emotions, particularly disgust, joy, and sadness. Future research could focus on:

- **Targeted Feature Engineering:** Explore advanced or domain-specific features designed to capture the subtle differences between physiologically similar emotional states.
- **Attention Mechanisms:** Implement attention mechanisms in deep learning models to allow them to dynamically focus on the most discriminative features or signal segments for each specific emotion.
- **Hierarchical Classification:** Develop a hierarchical classification system that first distinguishes between broad categories (e.g., high vs. low arousal) before attempting to classify specific emotions within those categories.

5.2.5 Exploring Advanced Feature Engineering and Deep Learning Approaches

While handcrafted features provided valuable insights, the field is rapidly moving towards more automated feature learning. Future research should leverage the power of deep learning to potentially capture more nuanced and complex patterns from raw or minimally processed physiological signals:

- **End-to-End Deep Learning:** Implement deep learning architectures, such as Convolutional Neural Networks (CNNs) for spatial feature extraction and Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTMs) networks for capturing temporal dependencies, directly on raw physiological waveforms. This would eliminate the reliance on manual feature engineering, potentially discovering more optimal representations.
- **Attention Mechanisms in Multimodal Fusion:** Implement more advanced multimodal fusion techniques, such as attention mechanisms, to dynamically weigh the importance of different modalities (ECG, EDA, RSP, speech) based on the emotional context. This could allow the model to focus on the most salient cues for a given emotion.

References

[1] Y. Cai, X. Li, and J. Li, "Emotion Recognition Using Different Sensors, Emotion Models, Methods and Datasets: A Comprehensive Review," *Sensors*, vol. 23, no. 5, p. 2455, Jan. 2023, number: 5 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/23/5/2455

[2] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 Audio-Visual Emotion Database," Jan. 2006. [Online]. Available: https://openreview.net/forum?id=fOL5GjEBeO

[3] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169–200, May 1992, publisher: Routledge _eprint: https://doi.org/10.1080/02699939208411068. [Online]. Available: https://doi.org/10.1080/02699939208411068

[4] W. Szwoch, "Emotion Recognition Using Physiological Signals," in *Proceedings of the Mulitimedia, Interaction, Design and Innnovation*, ser. MIDI '15. New York, NY, USA: Association for Computing Machinery, Jun. 2015, pp. 1–8. [Online]. Available: https://doi.org/10.1145/2814464.2814479

[5] R. W. Picard, *Affective Computing*. The MIT Press, Sep. 1997. [Online]. Available: https://doi.org/10.7551/mitpress/1140.001.0001

[6] R. Plutchik, *Emotion, a Psychoevolutionary Synthesis*. Harper & Row, 1980, google-Books-ID: G5t9AAAAMAAJ.

[7] M. Gjoreski, B. Mitrevski, M. Lustrek, and M. Gams, "An Inter-domain Study for Arousal Recognition from Physiological Signals," *Informatica (Slovenia)*, Jan. 2018. [Online]. Available: https://openreview.net/forum?id=cyoRt39SOd&referrer=%5Bthe%20profile%20of%20Blagoj%20Mitrevski%5D(%2Fprofile%3Fid%3D~Blagoj_Mitrevski1)

[8] X. Shen, X. Liu, X. Hu, D. Zhang, and S. Song, "Contrastive Learning of Subject-Invariant EEG Representations for Cross-Subject Emotion Recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2496–2511, Jul. 2023, arXiv:2109.09559 [cs]. [Online]. Available: http://arxiv.org/abs/2109.09559

[9] S. Taylor, N. Jaques, Weixuan Chen, S. Fedor, A. Sano, and R. Picard, "Automatic identification of artifacts in electrodermal activity data," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Milan: IEEE, Aug. 2015, pp. 1934–1937. [Online]. Available: http://ieeexplore.ieee.org/document/7318762/

[10] M. Monteiro, M. Dias, and H. Gamboa, "Assessing Electrocardiogram Quality: A Deep Learning Framework For Noise Detection And Classification," Jul. 2025, pp. 793–804. [Online]. Available: https://www.scitepress.org/Link.aspx?doi=10.5220/0013313800003911

[11] Z. Zhao and Y. Zhang, "SQI Quality Evaluation Mechanism of Single-Lead ECG Signal Based on Simple Heuristic Fusion and Fuzzy Comprehensive Evaluation," *Frontiers in Physiology*, vol. 9, Jun. 2018, publisher: Frontiers. [Online]. Available: https://www.frontiersin.org/journals/physiology/articles/10.3389/fphys.2018.00727/full

[12] M. Martiskainen and B. K. Sovacool, "Mixed feelings: A review and research agenda for emotions in sustainability transitions," *Environmental Innovation and Societal Transitions*, vol. 40, pp. 609–624, Sep. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2210422421000988

[13] C. E. Izard, "The Many Meanings/Aspects of Emotion: Definitions, Functions, Activation, and Regulation," *Emotion Review*, vol. 2, no. 4, pp. 363–370, Oct. 2010, publisher: SAGE Publications. [Online]. Available: https://doi.org/10.1177/1754073910374661

[14] G. Wadley, V. Kostakos, P. Koval, W. Smith, S. Webber, A. Cox, J. J. Gross, K. Höök, R. Mandryk, and P. Slovak, "The Future of Emotion in Human-Computer Interaction," in *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. New Orleans LA USA: ACM, Apr. 2022, pp. 1–6. [Online]. Available: https://dl.acm.org/doi/10.1145/3491101.3503729

[15] J. Shu, M. Chiu, and P. Hui, "Emotion Sensing for Mobile Computing," *IEEE Communications Magazine*, vol. 57, no. 11, pp. 84–90, Nov. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8908557/

[16] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang, "A Review of Emotion Recognition Using Physiological Signals," *Sensors (Basel, Switzerland)*, vol. 18, no. 7, p. 2074, Jun. 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6069143/

[17] D. C. Rubin and J. M. Talarico, "A Comparison of Dimensional Models of Emotion: Evidence from Emotions, Prototypical Events, Autobiographical Memories, and Words," *Memory (Hove, England)*, vol. 17, no. 8, pp. 802–808, Nov. 2009. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2784275/

[18] S. D'mello and A. Graesser, "Mind and Body: Dialogue and Posture for Affect Detection in Learning Environments," in *Proceedings of the 2007 conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*. NLD: IOS Press, Jun. 2007, pp. 161–168.

[19] S. Gu, W. Wang, F. Wang, and J. H. Huang, "Neuromodulator and Emotion Biomarker for Stress Induced Mental Disorders," *Neural Plasticity*, vol. 2016, p. 2609128, 2016.

[20] R. E. Jack, O. G. B. Garrod, and P. G. Schyns, "Dynamic Facial Expressions of Emotion Transmit an Evolving Hierarchy of Signals over Time," *Current Biology*, vol. 24, no. 2, pp. 187–192, Jan. 2014, publisher: Elsevier. [Online]. Available: https://www.cell.com/current-biology/abstract/S0960-9822(13)01519-4

[21] S. PS and G. Mahalakshmi, "Emotion models: a review," *International Journal of Control Theory and Applications*, vol. 10, no. 8, pp. 651–657, 2017.

[22] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980, place: US Publisher: American Psychological Association.

[23] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament," *Current Psychology*, vol. 14, no. 4, pp. 261–292, Dec. 1996. [Online]. Available: https://doi.org/10.1007/BF02686918

[24] R. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: analysis of affective physiological state," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, pp. 1175–1191, Oct. 2001. [Online]. Available: https://ieeexplore.ieee.org/document/954607

[25] F. Agrafioti, D. Hatzinakos, and A. K. Anderson, "ECG Pattern Analysis for Emotion Detection," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 102–115, Jan. 2012. [Online]. Available: https://ieeexplore.ieee.org/document/5999653

[26] E. A. Ashley and J. Niebauer, "Conquering the ECG," in *Cardiology Explained*. Remedica, 2004. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK2214/

[27] L. Wang, J. Hao, and T. H. Zhou, "ECG Multi-Emotion Recognition Based on Heart Rate Variability Signal Features Mining," *Sensors*, vol. 23, no. 20, p. 8636, Jan. 2023, number: 20 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/23/20/8636

[28] M. A. Hasnul, N. A. A. Aziz, S. Alelyani, M. Mohana, and A. A. Aziz, "Electrocardiogram-Based Emotion Recognition Systems and Their Applications in Healthcare—A Review," *Sensors (Basel, Switzerland)*, vol. 21, no. 15, p. 5015, Jul. 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8348698/

[29] A. K. Singh and S. Krishnan, "ECG signal feature extraction trends in methods and applications," *BioMedical Engineering OnLine*, vol. 22, no. 1, p. 22, Mar. 2023. [Online]. Available: https://doi.org/10.1186/s12938-023-01075-1

[30] W. Boucsein, *Electrodermal Activity*. Springer Science & Business Media, Feb. 2012, google-Books-ID: 6N6rnOEZEEoC.

[31] H. D. Critchley, "Electrodermal responses: what happens in the brain," *The Neuroscientist: A Review Journal Bringing Neurobiology, Neurology and Psychiatry*, vol. 8, no. 2, pp. 132–142, Apr. 2002.

[32] D. J. J. Braithwaite, "A Guide for Analysing Electrodermal Activity (EDA) & Skin Conductance Responses (SCRs) for Psychological Experiments."

[33] Realtime Technologies Ltd (Shimmer Sensing), *GSR User Guide, Rev 1.13*, Shimmer Sensing, Dublin, Ireland, Aug. 2018, version 1.13. [Online]. Available: https://shimmersensing.com/wp-content/docs/support/documentation/GSR_User_Guide_rev1.13.pdf

[34] R. Kocielnik, N. Sidorova, F. M. Maggi, M. Ouwerkerk, and J. H. D. M. Westerink, "Smart technologies for long-term stress monitoring at work," in

Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems. Porto, Portugal: IEEE, Jun. 2013, pp. 53–58. [Online]. Available: http://ieeexplore.ieee.org/document/6627764/

[35] M. E. Dawson, A. M. Schell, and D. L. Filion, "The Electrodermal System," in *Handbook of Psychophysiology*, 4th ed., ser. Cambridge Handbooks in Psychology, G. G. Berntson, J. T. Cacioppo, and L. G. Tassinary, Eds. Cambridge: Cambridge University Press, 2016, pp. 217–243. [Online]. Available: https://www.cambridge.org/core/books/handbook-of-psychophysiology/electrodermal-system/90AB2EBAA435385B2FC2BB6C05D4B880

[36] I. R. Kleckner, R. M. Jones, O. Wilder-Smith, J. B. Wormwood, M. Akcakaya, K. S. Quigley, C. Lord, and M. S. Goodwin, "Simple, Transparent, and Flexible Automated Quality Assessment Procedures for Ambulatory Electrodermal Activity Data," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 7, pp. 1460–1467, Jul. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8054675/

[37] M. Bolic, "Chapter 9 - Continuous monitoring of breathing," in *Pervasive Cardiovascular and Respiratory Monitoring Devices*, M. Bolic, Ed. Academic Press, Jan. 2023, pp. 327–368. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128209479000039

[38] F. A. Boiten, N. H. Frijda, and C. J. E. Wientjes, "Emotions and respiratory patterns: review and critical analysis," *International Journal of Psychophysiology*, vol. 17, no. 2, pp. 103–128, Jul. 1994. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0167876094900272

[39] R. A. Hameed, M. K. Sabir, M. A. Fadhel, O. Al-Shamma, and L. Alzubaidi, "Human emotion classification based on respiration signal," in *Proceedings of the International Conference on Information and Communication Technology*, ser. ICICT '19. New York, NY, USA: Association for Computing Machinery, Apr. 2019, pp. 239–245. [Online]. Available: https://dl.acm.org/doi/10.1145/3321289.3321315

[40] BIOPAC Systems, Inc., "Respiration recording," https://www.biopac.com/knowledge-base/respiration-recording/, accessed: 2025-05-10.

[41] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 479–493, Apr. 2021.

[42] S. Koelstra, C. Muhl, M. Soleymani, Jong-Seok Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "DEAP: A Database for Emotion Analysis ;Using Physiological Signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, Jan. 2012, publisher: Institute of Electrical and Electronics Engineers (IEEE). [Online]. Available: http://ieeexplore.ieee.org/document/5871728/

[43] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A Multimodal Database for Affect Recognition and Implicit Tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, Jan. 2012. [Online]. Available: https://ieeexplore.ieee.org/document/5975141

[44] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe, "ASCERTAIN: Emotion and Personality Recognition Using Commercial Sensors," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 147–160, Apr. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/7736040/

[45] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, Apr. 2013, pp. 1–8. [Online]. Available: https://ieeexplore.ieee.org/document/6553805

[46] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ser. ICMI '18. New York, NY, USA: Association for Computing Machinery, Oct. 2018, pp. 400–408. [Online]. Available: https://doi.org/10.1145/3242969.3242985

[47] K. Sharma, C. Castellini, E. L. van den Broek, A. Albu-Schaeffer, and F. Schwenker, "A dataset of continuous affect annotations and physiological signals for emotion analysis," *Scientific Data*, vol. 6, no. 1, p. 196, Oct. 2019, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41597-019-0209-0

[48] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *Journal of Personality*, vol. 60, no. 2, pp. 175–215, Jun. 1992.

[49] Y.-L. Hsu, J.-S. Wang, W.-C. Chiang, and C.-H. Hung, "Automatic ECG-Based Emotion Recognition in Music Listening," *IEEE Transactions on Affective*

Computing, vol. 11, no. 1, pp. 85–99, Jan. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8219396

[50] A. Anderson, T. Hsiao, and V. Metsis, "Classification of Emotional Arousal During Multimedia Exposure," in *Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments*, ser. PETRA '17. New York, NY, USA: Association for Computing Machinery, Jun. 2017, pp. 181–184. [Online]. Available: https://dl.acm.org/doi/10.1145/3056540.3064956

[51] H. Feng, H. M. Golshan, and M. H. Mahoor, "A wavelet-based approach to emotion classification using EDA signals," *Expert Systems with Applications*, vol. 112, pp. 77–86, Dec. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417418303609

[52] P. Kawde and G. K. Verma, "Deep belief network based affect recognition from physiological signals," in *2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON)*, Oct. 2017, pp. 587–592. [Online]. Available: https://ieeexplore.ieee.org/document/8251115

[53] E. Morgan, "All About EDA Part 2: Components of Skin Conductance – MindWare Technologies Support," Apr. 2018. [Online]. Available: https://support.mindwaretech.com/2018/04/all-about-eda-part-2-components-of-skin-conductance/

[54] J. Schumm, Ed., *Quality assessment of physiological signals during ambulatory measurements*. Zürich: ETH, 2010.

[55] S. Schmidt and H. Walach, "Electrodermal Activity (EDA) - State-of-the-art measurement and techniques for parapsychological purposes," *The Journal of Parapsychology*, vol. 64, no. 2, pp. 139–139, Jun. 2000, publisher: Parapsychology Press. [Online]. Available: https://go.gale.com/ps/i.do?p=AONE&sw=w&issn=00223387&v=2.1&it=r&id=GALE%7CA67718670&sid=googleScholar&linkaccess=abs

[56] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tröster, and U. Ehlert, "Discriminating Stress From Cognitive Load Using a Wearable EDA Device," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 410–417, Mar. 2010. [Online]. Available: https://ieeexplore.ieee.org/document/5325784

[57] K. M. Sibley, G. Mochizuki, and W. E. McIlroy, "Perturbation-evoked electrodermal activity responds to instability, not just motor or sensory drives," *Clinical*

Neurophysiology, vol. 120, no. 3, pp. 619–625, Mar. 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1388245708012765

[58] M. Strauss, C. Reynolds, S. Hughes, K. Park, G. McDarby, and R. W. Picard, "The HandWave Bluetooth Skin Conductance Sensor," in *Affective Computing and Intelligent Interaction*. Springer, Berlin, Heidelberg, 2005, pp. 699–706, iSSN: 1611-3349. [Online]. Available: https://link.springer.com/chapter/10.1007/11573548_90

[59] R. Schneider, S. Schmidt, M. Binder, F. Schäfer, and H. Walach, "Respiration-Related Artifacts in EDA Recordings: Introducing a Standardized Method to Overcome Multiple Interpretations," *Psychological Reports*, vol. 93, no. 3, pp. 907–920, Dec. 2003, publisher: SAGE Publications Inc. [Online]. Available: https://doi.org/10.2466/pr0.2003.93.3.907

[60] W. Boucsein, D. C. Fowles, S. Grimnes, G. Ben-Shakhar, W. T. roth, M. E. Dawson, D. L. Filion, and Society for Psychophysiological Research Ad Hoc Committee on Electrodermal Measures, "Publication recommendations for electrodermal measurements," *Psychophysiology*, vol. 49, no. 8, pp. 1017–1034, Aug. 2012.

[61] U. Satija, B. Ramkumar, and M. S. Manikandan, "A Review of Signal Processing Techniques for Electrocardiogram Signal Quality Assessment," *IEEE reviews in biomedical engineering*, vol. 11, pp. 36–52, 2018.

[62] M. Sugadev, M. Kaushik, V. Vijaykumar, and T. Ravi, "Survey on Various noise sources in ECG signal and its Filtering Methods," in *2021 10th International Conference on Internet of Everything, Microwave Engineering, Communication and Networks (IEMECON)*, Dec. 2021, pp. 01–06. [Online]. Available: https://ieeexplore.ieee.org/document/9689132

[63] A. C. V. Maggio, M. P. Bonomini, E. L. Leber, and P. D. Arini, "Quantification of ventricular repolarization dispersion using digital processing of the surface ECG," in *Advances in electrocardiograms*, R. M. Millis, Ed. Rijeka: IntechOpen, 2012, section: 9. [Online]. Available: https://doi.org/10.5772/23050

[64] C. Liu and J. Li, Eds., *Feature Engineering and Computational Intelligence in ECG Monitoring*. Singapore: Springer, 2020. [Online]. Available: https://link.springer.com/10.1007/978-981-15-3824-7

[65] S. Rahman, C. Karmakar, I. Natgunanathan, J. Yearwood, and M. Palaniswami, "Robustness of electrocardiogram signal quality indices," *Journal of The Royal*

Society Interface, vol. 19, no. 189, p. 20220012, Apr. 2022. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rsif.2022.0012

[66] I. Campero Jurado, I. Lorato, J. Morales, L. Fruytier, S. Stuart, P. Panditha, D. M. Janssen, N. Rossetti, N. Uzunbajakava, I. B. Serban, L. Rikken, M. de Kok, J. Vanschoren, and A. Brombacher, "Signal Quality Analysis for Long-Term ECG Monitoring Using a Health Patch in Cardiac Patients," *Sensors*, vol. 23, no. 4, p. 2130, Jan. 2023, number: 4 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/23/4/2130

[67] G. D. Clifford, D. Lopez, Q. Li, and I. Rezek, "Signal quality indices and data fusion for determining acceptability of electrocardiograms collected in noisy ambulatory environments," in *2011 Computing in Cardiology*, Sep. 2011, pp. 285–288, iSSN: 2325-8853. [Online]. Available: https://ieeexplore.ieee.org/document/6164558

[68] C. Daluwatte, L. Johannesen, L. Galeotti, J. Vicente, D. G. Strauss, and C. G. Scully, "Assessing ECG signal quality indices to discriminate ECGs with artefacts from pathologically different arrhythmic ECGs," *Physiological Measurement*, vol. 37, no. 8, pp. 1370–1382, Aug. 2016.

[69] R. He, H. Zhang, K. Wang, Y. Yuan, Q. Li, J. Pan, Z. Sheng, and N. Zhao, "Reducing false arrhythmia alarms in the ICU using novel signal quality indices assessment method," in *2015 Computing in Cardiology Conference (CinC)*, Sep. 2015, pp. 1189–1192, iSSN: 2325-887X. [Online]. Available: https://ieeexplore.ieee.org/document/7411129

[70] D. Janusek, Z. Patłowski, and R. Maniewski, "Evaluation of the T-wave alternans detection methods—simulation study," *Journal of Electrocardiology*, vol. 40, no. 4, Supplement, pp. S36–S37, Jul. 2007. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022073607005626

[71] N. Selvaraj, Y. Mendelson, K. H. Shelley, D. G. Silverman, and K. H. Chon, "Statistical approach for the detection of motion/noise artifacts in Photoplethysmogram," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2011, pp. 4972–4975, 2011.

[72] H. Chan, G. Chen, M. Lin, and S. Fang, "Heartbeat Detection Using Energy Thresholding and Template Match," in *2005 IEEE Engineering in Medicine and*

Biology 27th Annual Conference, Jan. 2005, pp. 6668–6670, iSSN: 1558-4615. [Online]. Available: https://ieeexplore.ieee.org/document/1616032

[73] L. Chen, M. T. Özsu, and V. Oria, "Robust and fast similarity search for moving object trajectories," in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, ser. SIGMOD '05. New York, NY, USA: Association for Computing Machinery, Jun. 2005, pp. 491–502. [Online]. Available: https://doi.org/10.1145/1066157.1066213

[74] V. Krasteva and I. Jekova, "QRS Template Matching for Recognition of Ventricular Ectopic Beats," *Annals of Biomedical Engineering*, vol. 35, no. 12, pp. 2065–2076, Dec. 2007. [Online]. Available: https://doi.org/10.1007/s10439-007-9368-9

[75] K. M. Menon, S. Das, M. Shervey, M. Johnson, B. S. Glicksberg, and M. A. Levin, "Automated electrocardiogram signal quality assessment based on Fourier analysis and template matching," *Journal of Clinical Monitoring and Computing*, vol. 37, no. 3, pp. 829–837, Jun. 2023, company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 3 Publisher: Springer Netherlands. [Online]. Available: https://link.springer.com/article/10.1007/s10877-022-00948-5

[76] Y. Shahriari, R. Fidler, M. M. Pelter, Y. Bai, A. Villaroman, and X. Hu, "Electrocardiogram Signal Quality Assessment Based on Structural Image Similarity Metric," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 4, pp. 745–753, Apr. 2018. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7954742

[77] J. Kužílek, M. Huptych, V. Chudáček, J. Spilka, and L. Lhotská, "Data driven approach to ECG signal quality assessment using multistep SVM classification," in *2011 Computing in Cardiology*, Sep. 2011, pp. 453–455, iSSN: 2325-8853. [Online]. Available: https://ieeexplore-ieee-org.proxy.bib.uottawa.ca/abstract/document/6164600

[78] M. Saad, E.-S. M. El-Rabaie, M. Saber, and W. Raslan, "A Comprehensive Survey of ECG Signal Denoising Techniques, Challenges and Novel Trends," 2024.

[79] J. Behar, J. Oster, Q. Li, and G. D. Clifford, "ECG Signal Quality During Arrhythmia and Its Application to False Alarm Reduction," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 6, pp. 1660–1666, Jun. 2013. [Online]. Available: https://ieeexplore.ieee.org/document/6412778

[80] M. Elgendi, "Optimal Signal Quality Index for Photoplethysmogram Signals," *Bioengineering*, vol. 3, no. 4, p. 21, Dec. 2016, number: 4 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2306-5354/3/4/21

[81] J. H. v. Bemmel, C. Zywietz, and J. A. Kors, "Signal Analysis for ECG Interpretation," *Methods of Information in Medicine*, vol. 29, pp. 317–329, Feb. 2018, publisher: Schattauer GmbH. [Online]. Available: https://www.thieme-connect.de/products/ejournals/abstract/10.1055/s-0038-1634807

[82] Q. Li, R. G. Mark, and G. D. Clifford, "Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman filter," *Physiological Measurement*, vol. 29, no. 1, p. 15, Dec. 2007. [Online]. Available: https://doi.org/10.1088/0967-3334/29/1/002

[83] M. Bolic, "Chapter 4 - Modeling and simulation of biomedical systems," in *Pervasive Cardiovascular and Respiratory Monitoring Devices*, M. Bolic, Ed. Academic Press, Jan. 2023, pp. 125–164. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B978012820947900012X

[84] P. H. Charlton, T. Bonnici, L. Tarassenko, D. A. Clifton, R. Beale, P. J. Watkinson, and J. Alastruey, "An impedance pneumography signal quality index: Design, assessment and application to respiratory rate monitoring," *Biomedical signal processing and control*, vol. 65, p. 102339, Mar. 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7611038/

[85] F. Løberg, V. Goebel, and T. Plagemann, "Quantifying the Signal Quality of Low-cost Respiratory Effort Sensors for Sleep Apnea Monitoring," in *Proceedings of the 3rd International Workshop on Multimedia for Personal Health and Health Care*, ser. HealthMedia'18. New York, NY, USA: Association for Computing Machinery, Oct. 2018, pp. 3–11. [Online]. Available: https://dl.acm.org/doi/10.1145/3264996.3264998

[86] S. Stankoski, I. Kiprijanovska, I. Mavridou, C. Nduka, H. Gjoreski, and M. Gjoreski, "Breathing Rate Estimation from Head-Worn Photoplethysmography Sensor Data Using Machine Learning," *Sensors*, vol. 22, no. 6, p. 2079, Jan. 2022, number: 6 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/22/6/2079

[87] S. Liu, R. X. Gao, D. John, J. Staudenmayer, and P. Freedson, "Tissue Artifact Removal from Respiratory Signals Based on Empirical Mode Decomposition," *Annals of biomedical engineering*, vol. 41, no. 5, pp. 1003–1015, May 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3625471/

[88] C. Orphanidou, T. Bonnici, P. Charlton, D. Clifton, D. Vallance, and L. Tarassenko, "Signal-quality indices for the electrocardiogram and photoplethysmogram: derivation and applications to wireless monitoring," *IEEE journal of biomedical and health informatics*, vol. 19, no. 3, pp. 832–838, May 2015.

[89] D. A. Birrenkott, M. A. Pimentel, P. J. Watkinson, and D. A. Clifton, "Robust estimation of respiratory rate via ECG- and PPG-derived respiratory quality indices," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug. 2016, pp. 676–679, iSSN: 1558-4615. [Online]. Available: https://ieeexplore.ieee.org/document/7590792/

[90] S. Böttcher, S. Vieluf, E. Bruno, B. Joseph, N. Epitashvili, A. Biondi, N. Zabler, M. Glasstetter, M. Dümpelmann, K. Van Laerhoven, M. Nasseri, B. H. Brinkman, M. P. Richardson, A. Schulze-Bonhage, and T. Loddenkemper, "Data quality evaluation in wearable monitoring," *Scientific Reports*, vol. 12, no. 1, p. 21412, Dec. 2022, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41598-022-25949-x

[91] M. Nasseri, E. Nurse, M. Glasstetter, S. Böttcher, N. M. Gregg, A. Laks Nandakumar, B. Joseph, T. Pal Attia, P. F. Viana, E. Bruno, A. Biondi, M. Cook, G. A. Worrell, A. Schulze-Bonhage, M. Dümpelmann, D. R. Freestone, M. P. Richardson, and B. H. Brinkmann, "Signal quality and patient experience with wearable devices for epilepsy management," *Epilepsia*, vol. 61, no. S1, pp. S25–S35, 2020, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/epi.16527. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/epi.16527

[92] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen, "NeuroKit2: A Python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, no. 4, pp. 1689–1696, Aug. 2021, company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 4 Publisher: Springer US. [Online]. Available: https://link.springer.com/article/10.3758/s13428-020-01516-y

[93] C. Carreiras, A. P. Alves, A. Lourenço, F. Canento, H. Silva, A. Fred *et al.*, "BioSPPy: Biosignal processing in Python," 2015–, [Online; accessed ¡today¿]. [Online]. Available: https://github.com/PIA-Group/BioSPPy/

[94] D. Khodadad, S. Nordebo, B. Müller, A. Waldmann, R. Yerworth, T. Becher, I. Frerichs, L. Sophocleous, A. van Kaam, M. Miedema, N. Seifnaraghi, and R. Bayford, "Optimized breath detection algorithm in electrical impedance tomography," *Physiological Measurement*, vol. 39, no. 9, p. 094001, Sep. 2018, publisher: IOP Publishing. [Online]. Available: https://dx.doi.org/10.1088/1361-6579/aad7e6

[95] J. Pan and W. J. Tompkins, "A Real-Time QRS Detection Algorithm," *IEEE Transactions on Biomedical Engineering*, vol. BME-32, no. 3, pp. 230–236, Mar. 1985. [Online]. Available: https://ieeexplore.ieee.org/document/4122029

[96] K. Ross, P. Hungler, and A. Etemad, "Unsupervised Multi-Modal Representation Learning for Affective Computing with Multi-Corpus Wearable Data," Aug. 2020, arXiv:2008.10726 [cs]. [Online]. Available: http://arxiv.org/abs/2008.10726

[97] K. Ross, P. Sarkar, D. Rodenburg, A. Ruberto, P. Hungler, A. Szulewski, D. Howes, and A. Etemad, "Toward Dynamically Adaptive Simulation: Multimodal Classification of User Expertise Using Wearable Devices," *Sensors (Basel, Switzerland)*, vol. 19, no. 19, p. 4270, Oct. 2019. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC6806062/

[98] M. Gnacek, L. Quintero, I. Mavridou, E. Balaguer-Ballester, T. Kostoulas, C. Nduka, and E. Seiss, "AVDOS-VR: Affective Video Database with Physiological Signals and Continuous Ratings Collected Remotely in VR," *Scientific Data*, vol. 11, no. 1, p. 132, Jan. 2024, publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41597-024-02953-6

[99] S. H. Kim, H. W. Son, T. M. Lee, and H. J. Baek, "Drunk Driver Detection Using Multiple Non-Invasive Biosignals," *Sensors*, vol. 25, no. 5, p. 1281, Feb. 2025. [Online]. Available: https://www.mdpi.com/1424-8220/25/5/1281

[100] M. L. Munoz, A. van Roon, H. Riese, C. Thio, E. Oostenbroek, I. Westrik, E. J. C. de Geus, R. Gansevoort, J. Lefrandt, I. M. Nolte, and H. Snieder, "Validity of (Ultra-)Short Recordings for Heart Rate Variability Measurements," *PLoS ONE*, vol. 10, no. 9, p. e0138921, Sep. 2015. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC4586373/

[101] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," *Biological Psychology*, vol. 84, no. 3, pp. 394–421, Jul. 2010. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0301051110000827

[102] P. van Gent, H. Farah, N. van Nes, and B. van Arem, "HeartPy: A novel heart rate algorithm for the analysis of noisy signals," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 66, pp. 368–378, Oct. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1369847818306740

[103] A. Greco, A. Lanata, G. Valenza, E. P. Scilingo, and L. Citi, "Electrodermal activity processing: a convex optimization approach," *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, vol. 2014, pp. 2290–2293, 2014.

[104] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, company: Springer Distributor: Springer Institution: Springer Label: Springer Number: 1 Publisher: Kluwer Academic Publishers. [Online]. Available: https://link.springer.com/article/10.1023/A:1010933404324

[105] I. O. Joudeh, A.-M. Cretu, S. Guimond, and S. Bouchard, "Prediction of Emotional Measures via Electrodermal Activity (EDA) and Electrocardiogram (ECG)," *Engineering Proceedings*, vol. 27, no. 1, p. 47, 2022, number: 1 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/2673-4591/27/1/47

[106] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001, publisher: Institute of Mathematical Statistics. [Online]. Available: http://www.jstor.org/stable/2699986

[107] T. G. Nick and K. M. Campbell, "Logistic Regression," in *Topics in Biostatistics*. Humana Press, 2007, pp. 273–301, iSSN: 1940-6029. [Online]. Available: https://link.springer.com/protocol/10.1007/978-1-59745-530-5_14

[108] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain." *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958, publisher: American Psychological Association (APA). [Online]. Available: https://doi.apa.org/doi/10.1037/h0042519

[109] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language Modeling with Gated Convolutional Networks," Sep. 2017, arXiv:1612.08083 [cs]. [Online]. Available: http://arxiv.org/abs/1612.08083

[110] S. O. Arik and T. Pfister, "TabNet: Attentive Interpretable Tabular Learning," Dec. 2020, arXiv:1908.07442 [cs]. [Online]. Available: http://arxiv.org/abs/1908.07442

[111] D.-K. Kim, "TabNet: A Deep Learning Breakthrough for Tabular Data," Feb. 2025. [Online]. Available: https://medium.com/@kdk199604/tabnet-a-deep-learning-breakthrough-for-tabular-data-bcd39c47a81c

[112] I. M. Agus Wirawan, R. Wardoyo, D. Lelono, S. Kusrohmaniah, and S. Asrori, "Comparison of Baseline Reduction Methods for Emotion Recognition Based On Electroencephalogram Signals," in *2021 Sixth International Conference on Informatics and Computing (ICIC)*, Nov. 2021, pp. 1–7. [Online]. Available: https://ieeexplore.ieee.org/document/9632948

[113] S. Zhao, A. Gholaminejad, G. Ding, Y. Gao, J. Han, and K. Keutzer, "Personalized Emotion Recognition by Personality-Aware High-Order Learning of Physiological Signals," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 15, no. 1s, pp. 14:1–14:18, Jan. 2019. [Online]. Available: https://doi.org/10.1145/3233184

[114] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," Dec. 2022, arXiv:2212.04356 [eess]. [Online]. Available: http://arxiv.org/abs/2212.04356

[115] "openai/whisper," Aug. 2025, original-date: 2022-09-16T20:02:54Z. [Online]. Available: https://github.com/openai/whisper

[116] "sentence-transformers/all-MiniLM-L6-v2 · Hugging Face," Jan. 2024. [Online]. Available: https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[117] H.-C. Yang and C.-C. Lee, "An Attribute-invariant Variational Learning for Emotion Recognition Using Physiology," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 1184–1188, iSSN: 2379-190X. [Online]. Available: https://ieeexplore.ieee.org/document/8683290

[118] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-González, E. Abdulhay, and N. Arunkumar, "Using Deep Convolutional Neural Network for Emotion

Detection on a Physiological Signals Dataset (AMIGOS)," *IEEE Access*, vol. 7, pp. 57–67, 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8543567/

[119] P. Sarkar and A. Etemad, "Self-Supervised Learning for ECG-Based Emotion Recognition," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 3217–3221, iSSN: 2379-190X. [Online]. Available: https://ieeexplore.ieee.org/document/9053985

[120] H.-C. Yang and C.-C. Lee, "Annotation Matters: A Comprehensive Study on Recognizing Intended, Self-reported, and Observed Emotion Labels using Physiology," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, Sep. 2019, pp. 1–7, iSSN: 2156-8111. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8925516

[121] S. Schachter and J. E. Singer, "Cognitive, social, and physiological determinants of emotional state," *Psychological Review*, vol. 69, pp. 379–399, Sep. 1962.

[122] G. Mandler, "Emotion," in *Handbook of Psychology*. John Wiley & Sons, Ltd, 2003, pp. 157–175, section: 8 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0471264385.wei0108. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/0471264385.wei0108

[123] L. Ivonin, H.-M. Chang, W. Chen, and M. Rauterberg, "Unconscious emotions: quantifying and logging something we are not aware of," *Personal and Ubiquitous Computing*, vol. 17, no. 4, pp. 663–673, Apr. 2013. [Online]. Available: https://doi.org/10.1007/s00779-012-0514-5

[124] I. B. Mauss and M. D. Robinson, "Measures of emotion: A review," *Cognition and Emotion*, vol. 23, no. 2, pp. 209–237, Feb. 2009, publisher: Routledge _eprint: https://doi.org/10.1080/02699930802204677. [Online]. Available: https://doi.org/10.1080/02699930802204677

[125] D. L. Paulhus, "Chapter 2 - Measurement and Control of Response Bias," in *Measures of personality and social psychological attitudes*, J. P. Robinson, P. R. Shaver, and L. S. Wrightsman, Eds. Academic Press, 1991, pp. 17–59. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B978012590241050006X

[126] N. Gao, M. Saiedur Rahaman, W. Shao, and F. D. Salim, "Investigating the Reliability of Self-report Data in the Wild: The Quest for Ground Truth," in *Adjunct*

Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers, ser. UbiComp/ISWC '21 Adjunct. New York, NY, USA: Association for Computing Machinery, Sep. 2021, pp. 237–242. [Online]. Available: https://doi.org/10.1145/3460418.3479338

[127] D. Myrick, E. Sánchez, H. Donahue, M. Williams, and J. Leonard, *The Blair witch project*. Artisan Entertainment Santa Monica, CA, 1999.

[128] J. Wan, *The Conjuring [Film]*. Warner Bros. Pictures, 2013.

[129] T. Kaye, "American history x, new line cinema," *Cited on*, p. 99, 1998.

[130] O. Stone, *Platoon [Film]*. Orion Pictures., 1986.

[131] YouTube, "Baby laughing hysterically at ripping paper," https://www.youtube.com/watch?v=RP4abiHdQpc, 2011.

[132] ——, "Mom cat shows baby kittens that golden retriever is safe for them," https://www.youtube.com/watch?v=N2hssAyomdU, 2022.

[133] L. Scherfig, *One Day [Film]*. Focus Features., 2011.

[134] A. J. C. Nicholas Stoller and B. O'Brien, *Neighbors [Film]*. Universal Pictures., 2014.

[135] D. Boyle, "Trainspotting film," *United Kingdom: Polygram Filmed Entertainment*, 1996.

[136] R. Rodriguez, *Planet Terror [Film]*. Dimension Films, 2007.

[137] YouTube, "The saddest video ever captured," https://www.youtube.com/watch?v=Xjwu2Auecko, 2022.

[138] H. Zieff and L. Elehwany, *My Girl [Film]*. Columbia Pictures, 1991.

APPENDICES

Appendix A

Movie Clips Used in the Study

A.1 Movie Information

The twelve movie clips used in the study and their meta information is listed in Table A.1. The movie sources are cited in Movie Clip column. Out of the twelve movie clips, seven were used in other psychological research. For the five other clips, we did a pilot study consisting of 25 participants and validated the effectiveness of the clips before the final data collection.

Movie Clip	Tag	Scene Description	Duration	Validation Source
The Blair Witch Project [127]	Fear	Final scene when screaming intensifies, man standing facing the wall and camera falls.	2:03	schaefer2010assessing
The Conjuring [128]	Fear	Girl gets out of bed at night and bangs her head on a cupboard. Frantic scene.	2:26	iyilikci2024extended
American History X [129]	Anger	Neo-Nazi kills a black man, smashing his head on the curb and then smiles after being arrested.	3:24	schaefer2010assessing
Platoon [130]	Anger	Villagers pushed around in burning village and soldier stops other soldiers from raping a child.	2:42	Author tested in pilot.
Baby laughing at ripping paper [131]	Joy	8-month-old Micah (a boy) laughing hysterically while at-home daddy rips up a job rejection letter.	1:44	Author tested in pilot.
Cats and Dog playing together [132]	Joy	Dog lies peacefully on a large bed with kittens and adult cat moving around. With happy music.	1:53	Author tested in pilot.
One Day [133]	Surprise	Woman rides a bicycle; she gets hit by a truck.	2:26	zupan2020eliciting
Neighbors [134]	Surprise	Woman calls man about missing airbags Man is ejected to an office ceiling.	1:07	Author tested in pilot.
Trainspotting [135]	Disgust	The main character enters "The worst toilet in Scotland" and later dives into a filthy toilet bowl.	1:23	schaefer2010assessing
Planet Terror [136]	Disgust	Scene where man is examined by doctors in a hospital and exposes infected parts of his body.	2:01	michelini2019latemo
Young impala and dead mother [137]	Sadness	Young impala finds adult impala lying down and apparently dead. Then lies by dead animal.	1:44	Author tested in pilot.
My Girl [138]	Sadness	Funeral scene where girl cries and runs away after approaching the casket where a little boy lies.	1:39	gabert2015ratings

Table A.1: Information of 12 movie clips used in this study.

Appendix B

Participant-wise Classification Results

B.1 Participant-wise Classification Results

This appendix provides a detailed, participant-by-participant breakdown of the multiclass emotion classification results from the subject-wise analysis presented in **Experiment 4 (Section 4.4.2)**. The following tables offer a granular view of model performance, allowing for direct comparison of predictions made by the three primary classifiers—**TabNet**, **GLU**, and **Random Forest**—for each of the twelve emotional videos.

Each row corresponds to a single participant's results for a specific set of emotions. To facilitate analysis, any instance where a model's predicted emotion does not match the intended emotion of the video (the ground truth) is highlighted in **red**. Furthermore, the column labeled **"PRE(s)"** lists the emotion(s) that participants self-reported with the highest intensity for each video, allowing comparison among the model's prediction, the intended stimulus emotion, and the participant's subjective experience. These tables are designed to reveal subject-specific patterns, identify consistently misclassified emotions, and provide a transparent, comprehensive view of each model's generalization capabilities to unseen individuals.

Table B.1: Participant-wise comparison of model performance for the emotions **Anger**, **Disgust**, and **Fear**. Each row details a single participant's results across the three emotion-eliciting videos. Misclassifications, where the predicted label does not match the ground truth emotion (the top-level header), are highlighted in red. **PRE(s)**: **Participant Reported Emotion(s)**. This column shows the emotion(s) that was(were) ranked the highest by the participant for that video. **TabNet**: Emotion predicted by the TabNet model. **GLU**: Emotion predicted by GLU model. **RF**: Emotion predicted by the Random Forest model.

Emotion	**Anger**				**Disgust**				**Fear**			
	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF
Participant												
1	sadness, disgust	fear	anger	anger	disgust	joy	disgust	joy	fear, surprise	anger	fear	fear
2	disgust	anger	fear	anger	disgust	joy	disgust	joy	joy	fear	fear	fear
3	anger	fear	fear	fear	disgust	disgust	sadness	disgust	fear	fear	fear	fear
4	disgust	anger	fear	fear	disgust	fear	joy	joy	surprise	anger	fear	fear
5	sadness, disgust, surprise, anger	anger	anger	anger	disgust, surprise	disgust	sadness	disgust	joy, fear, surprise	fear	anger	fear
6	anger	sadness	anger	anger	disgust	fear	joy	joy	fear, surprise	anger	fear	fear
7	disgust	joy	anger	anger	disgust	disgust	joy	disgust	disgust, fear, surprise	joy	fear	anger
8	sadness, disgust, anger	anger	anger	anger	disgust	disgust	joy	disgust	fear	anger	anger	fear
10	sadness, disgust	joy	anger	anger	disgust	joy	joy	surprise	fear, surprise	fear	surprise	fear

Continued on next page

Table B.1 – continued from previous page

Emotion	**Anger**				**Disgust**				**Fear**			
	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF
Participant												
11	sadness, disgust	anger	fear	anger	disgust, surprise	sadness	sadness	disgust	fear	anger	anger	fear
12	disgust	anger	anger	anger	disgust	disgust	fear	disgust	sadness, surprise	fear	fear	fear
14	disgust	anger	anger	anger	disgust	disgust	disgust	disgust	fear	disgust	disgust	fear
15	disgust, anger	anger	anger	anger	disgust	joy	joy	disgust	surprise	fear	fear	fear
16	disgust	fear	anger	anger	surprise	surprise	joy	joy	fear	anger	anger	fear
17	sadness, disgust	anger	anger	anger	disgust	fear	joy	joy	surprise	fear	fear	fear
19	disgust	fear	fear	anger	disgust	fear	joy	joy	fear, surprise	fear	fear	anger
20	sadness	anger	anger	anger	disgust	joy	joy	disgust	fear	joy	joy	disgust
21	disgust	anger	fear	fear	disgust	sadness	sadness	disgust	surprise	anger	anger	fear
22	sadness, disgust, surprise, anger	anger	anger	anger	disgust	sadness	disgust	disgust	fear	anger	fear	fear
24	disgust, surprise, anger	anger	anger	anger	disgust	sadness	sadness	disgust	fear	fear	joy	fear
25	disgust, anger	fear	fear	fear	disgust	fear	joy	disgust	fear	fear	fear	fear
27	disgust	anger	anger	anger	disgust	sadness	sadness	disgust	fear, surprise	fear	fear	fear
28	anger	anger	anger	anger	disgust	joy	disgust	disgust	fear	fear	fear	fear

Continued on next page

Table B.1 – continued from previous page

Emotion	Anger				Disgust				Fear			
	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF
Participant												
29	disgust	fear	anger	anger	surprise	joy	disgust	disgust	fear	fear	disgust	fear
30	sadness, disgust	anger	anger	fear	disgust, surprise	fear	surprise	joy	joy, surprise	fear	anger	fear
32	disgust, surprise	anger	anger	anger	disgust	disgust	sadness	disgust	sadness, disgust	joy	joy	joy
33	sadness	anger	anger	anger	disgust	joy	joy	disgust	fear	fear	fear	fear
34	surprise	anger	anger	anger	disgust	joy	fear	joy	fear	anger	anger	anger
36	sadness	anger	fear	anger	disgust	sadness	disgust	disgust	fear	fear	fear	fear
37	disgust	anger	anger	anger	disgust	fear	sadness	disgust	fear	fear	joy	joy
38	sadness, disgust, anger	anger	anger	anger	disgust	disgust	disgust	disgust	fear	anger	anger	anger
39	disgust, surprise, anger	anger	fear	anger	disgust, surprise	anger	surprise	surprise	fear, surprise	fear	fear	fear
40	disgust	anger	anger	anger	disgust, surprise	sadness	sadness	disgust	fear	sadness	anger	joy
42	sadness, disgust, anger	anger	anger	fear	disgust	disgust	disgust	disgust	fear	disgust	fear	fear
43	sadness, disgust, fear	anger	fear	anger	disgust, surprise	joy	joy	disgust	fear	fear	fear	fear
47	disgust, fear	fear	fear	anger	disgust, surprise	fear	sadness	disgust	fear	fear	fear	fear

Continued on next page

Table B.1 – continued from previous page

Emotion	**Anger**				**Disgust**				**Fear**			
	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF
Participant												
48	disgust, anger	anger	fear	fear	disgust	anger	joy	joy	surprise	joy	surprise	joy
49	sadness, disgust, anger	fear	anger	anger	disgust	anger	joy	disgust	fear	fear	fear	fear
50	sadness, disgust, anger	fear	anger	anger	disgust	disgust	sadness	disgust	fear	fear	joy	joy
52	disgust	fear	anger	anger	joy	anger	disgust	joy	surprise	fear	fear	fear
55	disgust	anger	anger	anger	disgust	fear	joy	disgust	disgust, fear, surprise	anger	fear	fear
56	disgust	anger	anger	anger	disgust	sadness	disgust	disgust	fear	fear	joy	fear
57	sadness, disgust, anger	joy	fear	fear	disgust	joy	surprise	surprise	fear, surprise	fear	fear	anger
58	disgust	joy	anger	anger	disgust	joy	joy	joy	fear	anger	anger	anger
61	anger	anger	anger	anger	joy	disgust	disgust	disgust	surprise	anger	fear	fear
63	disgust, anger	anger	anger	anger	disgust, surprise	fear	fear	fear	fear, surprise	fear	disgust	fear
64	disgust	fear	fear	fear	disgust	sadness	sadness	disgust	fear	fear	fear	fear
65	disgust, surprise, anger	anger	anger	anger	disgust	sadness	sadness	disgust	sadness, fear	fear	joy	fear

Continued on next page

Table B.1 – continued from previous page

Emotion	**Anger**				**Disgust**				**Fear**			
	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF
Participant												
66	sadness, disgust, anger	fear	fear	fear	disgust	sadness	disgust	disgust	surprise	fear	fear	fear
67	sadness	anger	anger	fear	disgust	disgust	joy	joy	fear	disgust	disgust	fear
68	surprise	anger	anger	anger	disgust	fear	joy	surprise	fear	fear	fear	fear
69	disgust	anger	anger	anger	disgust	sadness	sadness	disgust	surprise	fear	fear	fear
70	disgust	anger	anger	anger	disgust	fear	disgust	surprise	surprise	fear	fear	fear
71	sadness, disgust, fear, surprise	anger	anger	anger	disgust	disgust	sadness	disgust	fear	fear	fear	fear
72	disgust, anger	fear	fear	anger	disgust	disgust	sadness	disgust	fear	fear	joy	fear
73	disgust	anger	anger	anger	disgust	sadness	sadness	sadness	surprise	joy	surprise	surprise
74	sadness, disgust	anger	fear	fear	surprise	joy	fear	fear	surprise	anger	fear	fear
75	disgust	anger	anger	anger	disgust	disgust	disgust	disgust	fear	fear	surprise	fear
76	anger	anger	anger	anger	disgust, surprise	fear	fear	joy	surprise	anger	surprise	fear
77	disgust, anger	anger	anger	anger	disgust	joy	fear	surprise	fear	fear	fear	fear
78	sadness, joy	anger	anger	anger	disgust	sadness	sadness	disgust	fear	fear	fear	fear
79	disgust	anger	anger	anger	disgust	surprise	joy	joy	fear, surprise	fear	fear	fear

Continued on next page

Table B.1 – continued from previous page

Emotion	Anger				Disgust				Fear			
	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF
Participant												
80	disgust	anger	anger	anger	surprise	fear	joy	surprise	fear, surprise	fear	fear	fear
81	disgust, fear, anger	surprise	anger	anger	disgust, surprise	joy	disgust	disgust	fear	surprise	fear	fear
82	sadness, disgust, anger	anger	anger	anger	disgust	sadness	disgust	disgust	fear	fear	fear	fear
83	disgust	anger	anger	anger	disgust	surprise	joy	joy	fear, surprise	anger	anger	anger
84	fear, surprise	fear	anger	anger	disgust	anger	disgust	joy	fear	joy	fear	fear
85	surprise	anger	fear	anger	disgust, surprise	fear	joy	joy	surprise	fear	anger	fear
86	surprise	anger	anger	anger	disgust	disgust	joy	disgust	fear	fear	joy	fear
87	disgust, surprise, anger	anger	anger	anger	disgust, surprise	sadness	disgust	disgust	fear	disgust	anger	fear
88	disgust	anger	anger	anger	disgust, surprise	disgust	joy	disgust	sadness, fear	fear	fear	fear
89	disgust, anger	anger	anger	anger	disgust	joy	joy	surprise	surprise	fear	fear	fear
90	sadness, anger	anger	anger	anger	disgust	joy	disgust	joy	fear	anger	anger	fear
91	disgust	anger	anger	anger	disgust	sadness	sadness	sadness	fear	fear	fear	fear

Continued on next page

Table B.1 – continued from previous page

Emotion	**Anger**				**Disgust**				**Fear**			
	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF
Participant												
93	anger	anger	anger	anger	disgust, surprise	disgust	disgust	disgust	joy	fear	fear	fear
94	disgust	anger	anger	anger	disgust	disgust	joy	joy	fear	fear	fear	fear
95	disgust	joy	anger	anger	disgust	surprise	joy	joy	fear, surprise	sadness	anger	anger
96	disgust	fear	fear	fear	disgust, surprise	fear	joy	joy	fear, surprise	surprise	fear	fear
97	anger	anger	anger	anger	disgust	sadness	disgust	disgust	sadness	anger	surprise	fear
98	disgust	anger	anger	anger	disgust	sadness	joy	disgust	sadness, joy, fear	fear	joy	fear
99	disgust	anger	anger	anger	disgust	sadness	sadness	disgust	fear	fear	anger	fear

Table B.2: Participant-wise comparison of model performance for the emotions **Joy**, **Sadness**, and **Surprise**. Each row details a single participant's results across the three emotion-eliciting videos. Misclassifications, where the predicted label does not match the ground truth emotion (the top-level header), are highlighted in red. **PRE(s)**: **Participant Reported Emotion(s)**. This column shows the emotion(s) that was(were) ranked the highest by the participant for that video. **TabNet**: Emotion predicted by the TabNet model. **GLU**: Emotion predicted by GLU model. **RF**: Emotion predicted by the Random Forest model.

Emotion	**Joy**				**Sadness**				**Surprise**			
	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF
Participant												
1	joy	joy	joy	joy	sadness	disgust	disgust	sadness	surprise	joy	fear	joy
2	disgust, fear	joy	disgust	joy	sadness	sadness	sadness	sadness	surprise	surprise	surprise	surprise
3	joy	disgust	disgust	disgust	sadness, surprise	disgust	disgust	sadness	joy, surprise	surprise	surprise	surprise
4	joy	disgust	joy	joy	sadness	sadness	sadness	sadness	sadness, joy	joy	joy	joy
5	anger	fear	disgust	joy	sadness, joy	disgust	sadness	sadness	surprise	sadness	surprise	surprise
6	joy	fear	joy	joy	sadness	sadness	sadness	sadness	surprise	fear	joy	joy
7	joy	disgust	joy	joy	sadness	sadness	joy	sadness	joy, surprise	surprise	surprise	surprise
8	joy	joy	joy	joy	sadness	disgust	sadness	sadness	sadness, fear, surprise	joy	joy	joy
10	joy	sadness	surprise	joy	sadness	sadness	sadness	sadness	surprise	anger	surprise	surprise
11	joy	joy	fear	joy	sadness	sadness	sadness	sadness	joy, surprise	sadness	surprise	surprise

Continued on next page

Table B.2 – continued from previous page

Emotion	Joy				Sadness				Surprise			
	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF
Participant												
12	joy	joy	disgust	joy	sadness	sadness	sadness	sadness	surprise	joy	fear	surprise
14	joy	fear	disgust	joy	sadness	disgust	sadness	sadness	surprise	surprise	surprise	surprise
15	joy	joy	joy	joy	sadness	sadness	disgust	sadness	surprise	joy	joy	joy
16	joy	joy	joy	joy	sadness	sadness	sadness	sadness	surprise	surprise	sadness	surprise
17	joy	joy	disgust	joy	sadness	sadness	sadness	sadness	surprise	surprise	surprise	surprise
19	joy	fear	joy	fear	sadness	sadness	sadness	sadness	surprise	surprise	surprise	surprise
20	joy	joy	joy	joy	sadness	sadness	sadness	sadness	surprise	fear	fear	disgust
21	joy	disgust	joy	joy	sadness	sadness	sadness	sadness	surprise	surprise	surprise	surprise
22	joy	sadness	joy	joy	sadness	sadness	disgust	sadness	joy	surprise	surprise	surprise
24	joy	joy	joy	joy	sadness	sadness	sadness	sadness	surprise	joy	joy	disgust
25	joy	fear	joy	joy	sadness	sadness	surprise	sadness	sadness, fear	fear	surprise	surprise
27	joy	fear	joy	joy	sadness	sadness	sadness	sadness	joy	sadness	surprise	surprise
28	surprise	anger	disgust	disgust	sadness	sadness	sadness	sadness	sadness, fear, surprise, anger	disgust	disgust	disgust
29	joy	surprise	disgust	joy	joy	sadness	sadness	sadness	sadness	fear	disgust	joy
30	joy	anger	joy	joy	sadness	sadness	sadness	sadness	joy, surprise	sadness	surprise	surprise
32	joy	disgust	surprise	disgust	sadness	sadness	sadness	sadness	surprise	disgust	surprise	surprise
33	joy	joy	joy	joy	sadness	sadness	sadness	sadness	surprise	anger	fear	disgust
34	joy	joy	surprise	joy	sadness	sadness	sadness	sadness	surprise	sadness	surprise	surprise
36	joy	disgust	joy	joy	sadness	sadness	sadness	sadness	surprise	surprise	surprise	surprise

Continued on next page

Table B.2 – continued from previous page

Emotion	Joy				Sadness				Surprise			
	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF
Participant												
37	joy	anger	joy	joy	sadness, joy	sadness	sadness	sadness	sadness, disgust, fear, surprise	joy	joy	joy
38	joy	disgust	joy	joy	sadness	disgust	sadness	disgust	fear, surprise	surprise	surprise	surprise
39	joy	anger	fear	disgust	disgust, surprise	joy	sadness	sadness	surprise	fear	surprise	surprise
40	joy	joy	joy	joy	sadness	sadness	disgust	sadness	surprise	joy	anger	surprise
42	joy	joy	joy	joy	sadness	sadness	disgust	sadness	fear, surprise	surprise	surprise	surprise
43	joy	fear	surprise	joy	sadness	disgust	sadness	sadness	sadness	joy	surprise	surprise
47	joy	anger	disgust	joy	sadness	sadness	sadness	sadness	surprise	disgust	surprise	joy
48	joy	anger	fear	joy	sadness, surprise	sadness	sadness	sadness	joy, surprise	surprise	surprise	surprise
49	joy, surprise	joy	joy	joy	sadness, fear	sadness	sadness	sadness	surprise	surprise	surprise	surprise
50	joy	joy	disgust	joy	sadness	sadness	sadness	sadness	fear, surprise	surprise	surprise	surprise
52	surprise	fear	joy	joy	sadness	sadness	sadness	sadness	joy	surprise	surprise	surprise
55	joy	joy	disgust	joy	sadness	surprise	sadness	sadness	sadness, surprise	anger	surprise	surprise
56	joy	fear	fear	joy	sadness	sadness	sadness	sadness	sadness, joy, surprise	joy	fear	surprise

Continued on next page

Table B.2 – continued from previous page

Emotion	Joy				Sadness				Surprise			
	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF
Participant												
57	joy	surprise	surprise	surprise	sadness, surprise	disgust	sadness	sadness	fear, anger	fear	surprise	surprise
58	joy	joy	joy	joy	sadness	sadness	sadness	sadness	fear	joy	disgust	disgust
61	joy	joy	joy	joy	sadness	sadness	sadness	sadness	joy, surprise	surprise	surprise	surprise
63	joy	joy	disgust	disgust	sadness	sadness	disgust	sadness	joy, surprise	surprise	surprise	surprise
64	joy	joy	disgust	joy	sadness	sadness	sadness	sadness	surprise	surprise	surprise	surprise
65	joy	joy	sadness	joy	sadness	sadness	sadness	sadness	sadness, surprise	surprise	joy	surprise
66	joy	joy	surprise	joy	sadness	disgust	sadness	sadness	surprise	surprise	surprise	surprise
67	joy	disgust	surprise	surprise	sadness	sadness	disgust	sadness	surprise	sadness	surprise	surprise
68	joy	surprise	anger	joy	sadness	sadness	sadness	sadness	surprise	joy	surprise	joy
69	joy	joy	disgust	joy	joy	sadness	sadness	sadness	joy, fear, surprise	surprise	surprise	surprise
70	joy	joy	fear	disgust	sadness	disgust	sadness	sadness	surprise	disgust	fear	surprise
71	joy	joy	joy	joy	sadness	sadness	sadness	sadness	surprise	joy	fear	joy
72	joy, surprise	fear	fear	joy	sadness	disgust	disgust	sadness	surprise	surprise	surprise	surprise
73	joy	joy	joy	disgust	sadness	sadness	sadness	sadness	sadness, surprise	fear	surprise	surprise
74	joy	anger	fear	joy	sadness, fear	sadness	sadness	sadness	surprise	surprise	surprise	surprise
75	joy	anger	joy	joy	sadness	sadness	disgust	sadness	surprise	surprise	surprise	surprise
76	joy	joy	disgust	joy	sadness	sadness	sadness	sadness	surprise	anger	surprise	surprise

Continued on next page

Table B.2 – continued from previous page

Emotion	Joy				Sadness				Surprise			
	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF
Participant												
77	joy	fear	joy	fear	sadness	sadness	sadness	sadness	sadness, surprise	fear	joy	surprise
78	joy	joy	joy	joy	sadness, anger	sadness	disgust	sadness	surprise	surprise	surprise	surprise
79	joy	joy	joy	joy	sadness	sadness	sadness	sadness	sadness	joy	fear	surprise
80	joy	anger	joy	surprise	surprise	sadness	sadness	sadness	surprise	surprise	surprise	surprise
81	joy	fear	fear	joy	sadness	surprise	sadness	sadness	fear, surprise	surprise	surprise	surprise
82	joy	fear	fear	joy	sadness, surprise	sadness	sadness	sadness	surprise	joy	joy	joy
83	joy	surprise	anger	joy	sadness	sadness	sadness	sadness	surprise	fear	fear	joy
84	joy	anger	fear	joy	sadness	sadness	sadness	sadness	surprise	sadness	surprise	surprise
85	joy	disgust	surprise	disgust	sadness	disgust	sadness	sadness	surprise	joy	joy	disgust
86	joy	joy	joy	fear	sadness	sadness	surprise	sadness	surprise	joy	fear	joy
87	joy	joy	joy	joy	sadness, surprise	sadness	sadness	sadness	surprise	surprise	surprise	surprise
88	joy	disgust	fear	joy	sadness	disgust	disgust	sadness	sadness, surprise	joy	disgust	surprise
89	joy	joy	disgust	disgust	sadness	disgust	sadness	sadness	sadness	fear	joy	disgust
90	joy	surprise	joy	disgust	sadness	sadness	sadness	sadness	surprise	joy	joy	disgust
91	joy	joy	joy	joy	sadness	sadness	sadness	sadness	surprise	surprise	sadness	surprise
93	joy	joy	joy	joy	sadness	sadness	sadness	sadness	joy, surprise	surprise	surprise	surprise
94	joy	disgust	disgust	joy	sadness	disgust	sadness	sadness	surprise	surprise	surprise	surprise
95	joy	sadness	joy	joy	sadness	sadness	sadness	sadness	surprise	surprise	surprise	surprise
96	joy	joy	joy	joy	sadness	sadness	sadness	sadness	surprise	fear	joy	joy

Continued on next page

Table B.2 – continued from previous page

Emotion	**Joy**				**Sadness**				**Surprise**			
	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF	PRE(s)	TabNet	GLU	RF
Participant												
97	joy	joy	joy	joy	sadness	sadness	sadness	sadness	surprise	joy	fear	joy
98	joy	sadness	joy	joy	disgust	sadness	sadness	sadness	surprise	sadness	surprise	surprise
99	joy	joy	joy	joy	sadness	sadness	surprise	sadness	surprise	joy	fear	joy

Appendix C

Certificate of Ethics Approval

The following pages contain the official Certificate of Ethics Approval issued for this research by the University of Ottawa's Research Ethics Board.

Université d'Ottawa
Bureau d'éthique et d'intégrité de la recherche

University of Ottawa
Office of Research Ethics and Integrity

CERTIFICAT D'APPROBATION ÉTHIQUE | CERTIFICATE OF ETHICS APPROVAL

Numéro du dossier / Ethics File Number	H-10-23-9717
Titre du projet / Project Title	The Emotion Discrimination Data Collection and Analysis project: Towards an AI model of emotion discrimination.
Type de projet / Project Type	Recherche de professeur / Professor's research project
Statut du projet / Project Status	Approuvé / Approved
Date d'approbation (jj/mm/aaaa) / Approval Date (dd/mm/yyyy)	20/11/2023
Date d'expiration (jj/mm/aaaa) / Expiry Date (dd/mm/yyyy)	19/11/2024

Équipe de recherche / Research Team

Chercheur / Researcher	Affiliation	Role
Sylvain GAGNON	École de psychologie / School of Psychology	Chercheur Principal / Principal Investigator
Miodrag BOLIC	École de science informatique et de génie électrique / School of Electrical Engineering and Computer Science	Co-chercheur principal / Co-principal investigator
Vincent FRANCOEUR	École de psychologie / School of Psychology	Assistant de recherche / Research Assistant

Conditions spéciales ou commentaires / Special conditions or comments

550, rue Cumberland, pièce 154
Ottawa (Ontario) K1N 6N5 Canada

550 Cumberland Street, Room 154
Ottawa, Ontario K1N 6N5 Canada

613-562-5387 • 613-562-5338 • ethique@uOttawa.ca / ethics@uOttawa.ca
www.recherche.uottawa.ca/deontologie | www.recherche.uottawa.ca/ethics

Université d'Ottawa
Bureau d'éthique et d'intégrité de la recherche

University of Ottawa
Office of Research Ethics and Integrity

Le Comité d'éthique de la recherche (CÉR) de l'Université d'Ottawa, opérant conformément à l'*Énoncé de politique des Trois conseils* (2014) et toutes autres lois et tous règlements applicables, a examiné et approuvé la demande d'éthique du projet de recherche ci-nommé.

L'approbation est valide pour la durée indiquée plus haut et est sujette aux conditions énumérées dans la section intitulée "Conditions Spéciales ou Commentaires". Le formulaire « Renouvellement ou Fermeture de Projet » doit être complété quatre semaines avant la date d'échéance indiquée ci-haut afin de demander un renouvellement de cette approbation éthique ou afin de fermer le dossier.

Toutes modifications apportées au projet doivent être approuvées par le CÉR avant leur mise en place, sauf si le participant doit être retiré en raison d'un danger immédiat ou s'il s'agit d'un changement ayant trait à des éléments administratifs ou logistiques du projet. Les chercheurs doivent aviser le CÉR dans les plus brefs délais de tout changement pouvant augmenter le niveau de risque aux participants ou pouvant affecter considérablement le déroulement du projet, rapporter tout évènement imprévu ou indésirable et soumettre toute nouvelle information pouvant nuire à la conduite du projet ou à la sécurité des participants.

The University of Ottawa Research Ethics Board, which operates in accordance with the *Tri-Council Policy Statement* (2014) and other applicable laws and regulations, has examined and approved the ethics application for the above-named research project.

Ethics approval is valid for the period indicated above and is subject to the conditions listed in the section entitled "Special Conditions or Comments". The "Renewal/Project Closure" form must be completedfour weeks before the above-referenced expiry date to request a renewal of this ethics approval or closure of the file.

Any changes made to the project must be approved by the REB before being implemented, except when necessary to remove participants from immediate endangerment or when the modification(s) only pertain to administrative or logistical components of the project. Investigators must also promptly alert the REB of any changes that increase the risk to participant(s), any changes that considerably affect the conduct of the project, all unanticipated and harmful events that occur, and new information that may negatively affect the conduct of the project or the safety of the participant(s).

Kim THOMPSON
Responsable d'éthique en recherche / Protocol Officer
Pour/For **Daniel LAGAREC** Président(e) du/ Chair of the **Comité d'éthique de la recherche en sciences de la santé et sciences / Health Sciences and Sciences Research Ethics Board**

550, rue Cumberland, pièce 154
Ottawa (Ontario) K1N 6N5 Canada

550 Cumberland Street, Room 154
Ottawa, Ontario K1N 6N5 Canada

613-562-5387 • 613-562-5338 • ethique@uOttawa.ca / ethics@uOttawa.ca
www.recherche.uottawa.ca/deontologie | www.recherche.uottawa.ca/ethics