Automated Emotion Detection Using Multi-Modal Physiological Signals: A Path Towards Clinical Applications

Zeel Pansara

A thesis submitted for the degree of

Doctor of Philosophy

at the

School of Computer Science and Electronic Engineering

University of Essex

April 2025

Abstract

The ability to accurately detect and understand human emotions is crucial for various applications, from enhancing human-computer interaction to advancing mental health monitoring and neurorehabilitation. This thesis presents a multimodal emotion detection model that integrates Facial Emotion Recognition (FER), pupil size, and Galvanic Skin Response (GSR) to provide a continuous and subtle analysis of emotional states. Unlike conventional emotion detection systems that rely on single-modality approaches, this research overcomes key limitations by combining physiological signals that capture emotions' arousal and valence.

A key contribution of this study is a robust multimodal emotion detection framework that enhances pupil-based emotion prediction by isolating emotional signals from luminosity effects, improving feature reliability. Across 32 emotionally varied video clips shown to 47 participants, our corrected model achieved strong predictive performance (mean correlation of 0.65 $\pm$ 0.12, an $R2 - score$ of 0.43 $\pm$ 0.12, and Normalised Root Mean Square Error ($NRMSE$)) of 0.27 $\pm$ 0.036), significantly outperforming models using uncorrected pupil size. These results highlight the importance of addressing environmental confounds and the model's potential for real-world applications in affective computing.

After obtaining pupil size corrected for luminosity, we also extracted features from Facial Emotion Recognition (FER) and Galvanic Skin Response (GSR). We integrated them at a feature-level fusion. We then trained and evaluated an emotion detection machine learning model on the same 47 participants. The model employs a regression-based approach using the Extreme Gradient Boosting (XGBoost) algorithm, a powerful machine learning technique, to fuse these multimodal features. The model achieves higher accuracy than model trained on single physiological features, with a correlation of 0.91 $\pm$ 0.041, an $R2$ of 0.710 $\pm$ 0.098, and an $NRMSE$ of 0.183 $\pm$ 0.030 for valence and correlation of 0.86 $\pm$ 0.061, an $R2$ of 0.665 $\pm$ 0.359, and an $NRMSE$ of 0.187 $\pm$ 0.070 for arousal, showcasing its ability to predict emotional states continuously.

The model was evaluated on a diverse set of participants, showing robustness to inter-subject variability, and was designed with a lightweight architecture suitable for real-time use on wearable and mobile platforms. By addressing challenges such as signal fusion, temporal misalignment, and computational efficiency, this work advances the deployment of multimodal emotion detection systems. It lays the groundwork for emotion-aware technologies in clinical care, neurorehabilitation, and human–computer interaction, enabling continuous and personalised monitoring of emotional states.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to the “Multimotion Team.” Being part of this incredible group has been an enriching experience. Your collaborative spirit, innovative ideas, and constant support have played a crucial role in shaping my research and personal growth throughout this journey.

I am profoundly thankful to my supervisors, Dr. Caterina Cinel and Dr. Vito De Feo, for their guidance, insight, and unwavering support. Their mentorship has been instrumental at every stage of this work, challenging me to think critically and push the boundaries of my understanding.

To my friends, thank you for keeping me grounded, for the laughter, late-night calls, and all the small (yet meaningful) moments that made the tough times bearable. Your presence has been a much-needed source of strength.

To my parents and brother, your love, encouragement, and belief in me have been the foundation of this journey. Thank you for your endless patience and always being there—your support means the world to me.

This thesis is as much yours as it is mine.

Contents

List of Figures

List of Tables

LIST OF ABBREVIATIONS

ADM - Arousal Detection Model

ANS - Autonomic Nervous System

AI - Artificial Intelligence

ANC - Adaptive Noise Cancellation

ASD - Autism Spectrum Disorder

APA - Average Peak Amplitude

AUC - Area Under Curve

BPD - Borderline Personality Disorder

BVP - Blood Volume Pulse

CNN - Convolutional Neural Network

CCC - Concordance correlation coefficient

CERQ - Cognitive Emotion Regulation Questionnaire

CNS - Central Nervous System

CASE - Continuously Annotated Signals of Emotion

CSEE - Computer Science and Electronic Engineering

DES - Differential Emotions Scale

DT - Decision Tree

DEAP - Dataset for Emotion Analysis using Physiological Signals

DASM - Differential Asymmetry

Emotiw - Emotion detection in the Wild

EDA - Electrodermal Activity

EMG - Electromyogram

EMFAcs - Emotion Facial Action Coding System

ECG - Electrocardiogram

EEG - Electroencephalogram

EDG - Electrodermography

EOG - Electrooculogram

EMD - Empirical Mode Decomposition

ERI - Emotion Detection Impairment

EIIS - Emotion-Induced Interval Study

EII - Emotion-Induced Intervals

FER - Facial Emotion Recognition

FA - Factor Analysis

FrFT - Fractional Fourier Transform

FR - Facial Recognition

FIR - Finite Impulse Response

GSR - Galvanic Skin Response

GAPED - Geneva Affective Picture Database

GAD-7 - Generalized Anxiety Disorder

HCI - Human-Computer Interaction

HR - Heart Rate

HRV - Heart Rate Variability

HMM - Hidden Markov Model

HF - High Frequency

HP - High arousal Positive valence

HN - High arousal Negative valence

ICA - Independent Component Analysis

IIR - Infinite Impulse Response

INDSCAL - Individual Differences Scaling

IBI - Inter-Beat-Interval

IPDE - International Personality Disorder Examination

IMF - Intrinsic Mode Function

JPDA - Joint Probability Domain Adaptation

KNN - K-Nearest Neighbours

LSTM - Long Short-Term Memory

LBP - Local Binary Patterns

LBP-Top - Local Binary Patterns on Three Orthogonal Planes

LMS - Least Mean Squares

LASSO - Least Absolute Shrinkage and Selection Operator

LOPO - Leave-one-participant out

LP - Low arousal Positive valence

LN - Low arousal Negative valence

LF - Low Frequency

LEPM - Luminosity Effect Prediction Model

ML - Machine Learning

MEQ - Multidimensional Emotional Questionnaire

MFCC - Mel-Frequency Cepstral Coefficients

NRMSE - Normalised Root Mean Square Error

NMF - Non-Negative Matrix Factorisation

OCPD - Obsessive Compulsive Personality Disorder

PCA - Principle Component Analysis

PANAS - Positive and Negative Affect Schedule

PNS - Parasympathetic Nervous System

PRV - Pulse Rate Variability

PD - Personality Disorder

PHQ-9 - Patient Health Questionnaire

PTSD - Post-Traumatic Stress Disorder

PSD - Power Spectral Density

PPG - Photoplethysmography

RFE - Recursive Feature Elimination

RMSSD - Root Mean Square of Successive Differences

RSP - Respiration

RASM - Rational Asymmetry

RNN - Recurrent Neural Network

RV - Respiratory Volume

RMS - Root Mean Square

RF - Random Forest

SCR - Skin Conductance Response

SCL - Skin Conductance Level

SDNN - Standard Deviation of Normal-to-Normal intervals

SVR - Support Vector Regressor

SC - Skin Conductance

SKT - Skin Temperature

SMOTE - Synthetic Minority Oversampling Technique

SNR - Signal-to-Noise Ratio

SPSS - Statistical Package for the Social Sciences

STFT - Short-Time Fourier Transform

SVD - Singular Value Decomposition

SP - Shape Parameters

SVM - Support Vector Machine

SHAP - Shapley Additive Explanations

SAM - Self-Assessment Manikin

SD - Semantic Differential

SNS - Sympathetic Nervous System

TBI - Traumatic Brain Injury

TAS-20 - Toronto Alexithymia Scale

VIF - Variance Inflation Factor

VR - Vectorial Representation

WT - Wavelet Transform

WESAD - Wearable Stress and Affect Detection

XGBoost - Extreme Gradient Boosting

Declaration of Authorship

I hereby declare that the paper presented is my own work and that I have not used any sources other than those listed in the bibliography. In addition, I affirm to have clearly marked and acknowledged all quotations or references that have been taken from the works of others. I further declare that I have not submitted this paper to any other chair or institution in order to obtain a grade.

Zeel Mahendrakumar Pansara

Chapter 1

Introduction

In this introductory section, I have outlined the aim and significance of this research, providing relevant background information. Additionally, I have presented the research questions guiding this study, highlighted its contributions, described the thesis structure, and listed related publications.

1.1 Motivation and background

Our study aims to develop a machine-learning model that utilises various biomarkers derived from physiological signals to track and assess emotional states in real-time. By adopting a regression-based approach, we aim to capture subtle emotional changes rather than confining detection to specific categories. This approach fosters a more comprehensive and nuanced understanding of emotional dynamics, enhancing the model's effectiveness across various emotional experiences, particularly in clinical settings.

1.1.1 Emotion

"Emotions help keep us on the right track by ensuring that we are led by more than cognition" This insightful quote by Maurice Elias encapsulates the crucial role emotions play in guiding human behaviour and decision-making. Far from being mere irrational impulses, emotions serve as sophisticated adaptive mechanisms that work in tandem with our cognitive processes [1], [2].

Emotions serve as an evolutionary compass, guiding us through the complexities of our environment and social interactions. They provide rapid, intuitive responses to situations, often before our slower, deliberative cognitive processes can thoroughly analyse the circumstances [3]. This emotional guidance system has been honed over millennia to promote survival and well-being.

The Significance of Self-Awareness and Understanding Others' Emotions

The interplay between emotion and cognition is complex and bidirectional. While cognitive processing is essential for eliciting emotional responses, emotions, in turn, shape and guide our cognitive functions [4]. This dynamic relationship enables us to adapt to our surroundings by integrating rational analysis and intuitive feelings, resulting in well-rounded decision-making. Emotions enhance our cognitive abilities by helping us prioritise information, invest experiences with meaning, and provide valuable insights that pure logic might overlook [5]. Beyond influencing thought processes, emotions play a crucial role in psychological and social functioning, making emotional awareness a vital skill for achieving personal and interpersonal success.

Recognising one's emotions is fundamental to self-awareness, emotional regulation, and well-being. Understanding personal emotions enables individuals to make more informed decisions by striking a balance between logic and intuition [6]. It also plays a crucial role in emotional regulation [7], helping individuals manage stress [8], frustration [9], and impulsive reactions effectively [10]. By being aware of their emotions, individuals gain a deeper understanding of their values, triggers, and motivations, which fosters personal growth and resilience [11]. Furthermore, emotional self-awareness supports goal achievement by maintaining motivation and perseverance, even in challenging situations [12].

Equally important is the ability to recognise emotions in others, which is essential for building strong relationships and effective communication. Understanding the feelings of those around us fosters empathy and compassion, enabling more supportive and meaningful interactions. This skill is particularly valuable in conflict resolution, allowing individuals to navigate sensitive disagreements and find constructive solutions. Recognising emotions enhances teamwork, leadership, and workplace harmony in professional settings by fostering emotional intelligence [13]. Additionally, in clinical and caregiving contexts, the ability to identify emotions plays a vital role in diagnosing and treating emotional or mental health conditions, ensuring appropriate support and intervention [14].

Self-awareness and the ability to perceive others' emotions are crucial for maintaining psychological well-being, achieving social success, and making effective decisions. Developing emotional intelligence enables individuals to foster healthier relationships, promote personal growth, and navigate life's challenges more effectively. By integrating emotional awareness into daily interactions, individuals can enhance their communication, strengthen their resilience, and contribute to a more empathetic and understanding society.

Impact of Emotion Recognition on Individuals with Clinical Conditions

People with mental health conditions like anxiety, depression, personality disorders (PD), and alexithymia [15]–[18] often have trouble recognising emotions in themselves or others. This difficulty is a result of their condition, not its cause. However, having difficulty understanding emotions can make life even more complicated, affecting their relationships, daily life, and overall well-being.

For example, individuals with PDs like borderline or antisocial PD may experience heightened emotional sensitivity or a lack of empathy, which impairs their social interactions and impulse control [19]. Those with anxiety disorders often misinterpret neutral expressions as threatening, exacerbating their distress and avoidance behaviours [20]. People with schizoid personality traits typically struggle with recognising or responding to emotions in others, resulting in social detachment and isolation [21].

In depression, deficits in emotional recognition can lead to negative biases, causing individuals to perceive neutral or positive expressions as unfavourable, reinforcing feelings of worthlessness and hopelessness [20], [22]. Individuals with alexithymia, often linked to brain injuries or trauma, face challenges in identifying and articulating their own emotions, which complicates emotional regulation and interpersonal relationships [15], [19].

Understanding how these conditions affect emotion detection is crucial for developing targeted interventions. Enhancing emotional awareness through therapy [23], training [24], or assistive technologies [25] can improve social functioning, reduce distress, and promote healthier emotional regulation [26], [27]. However, these interventions depend on accurately identifying and categorising emotions [28], a process known as emotion labelling. Performed by individuals, psychologists, researchers, and AI systems, emotion labelling is essential for effective emotion recognition and tailoring appropriate support strategies.

Emotion Labelling

Emotion labelling involves identifying and categorising emotions in text, speech, facial expressions, or physiological signals [29]. It plays a key role in understanding human affective states and is widely applied in psychology, human-computer interaction (HCI), and affective computing. Accurate labelling requires careful interpretation of contextual cues, as emotions are complex and can vary across individuals and cultures [30].

The labelling process typically involves human annotators who assign emotional categories based on predefined classification systems, such as Ekman's six basic emotions—happiness, sadness, fear, anger, surprise, and disgust—or dimensional models like the valence-arousal framework [31], [32]. Multiple annotators often review the

same data to maintain uniformity, and their assessments are validated through inter-rater agreement methods [33].

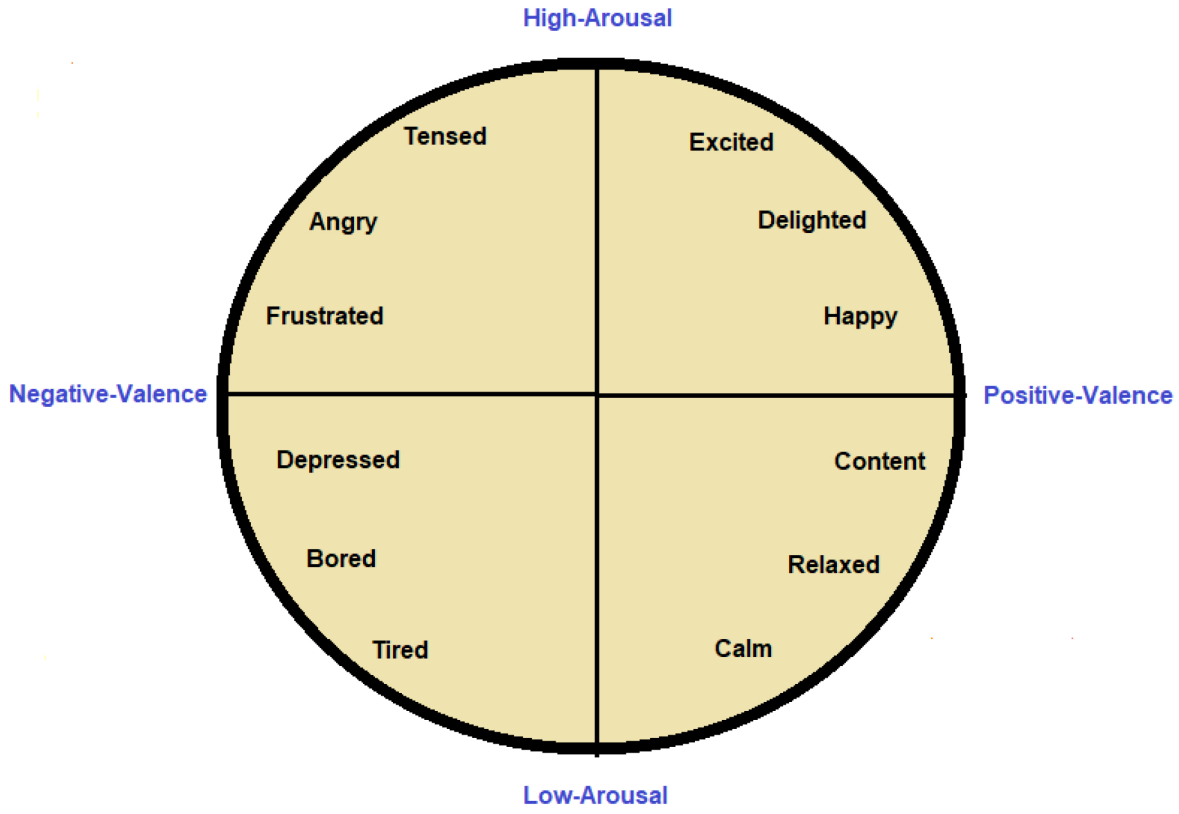


Figure 1.1: Emotion labelling in two-dimensional space [31]

While the valence–arousal model remains widely used due to its simplicity, recent research has questioned its ability to capture the full complexity of human emotions. For example, Cowen and Keltner [34] identified 27 distinct emotion categories that exist on a continuous spectrum, challenging the notion that emotions can be reduced to just two dimensions. Alternative frameworks, such as basic emotion theory or semantic space models, offer more granular representations, though they may be more complex to implement. Choosing the appropriate framework depends on the specific goals of the study and the desired level of emotional nuance.

Emotion labelling helps researchers analyse patterns in human emotional responses, providing insights into behaviour, decision-making, and mental health [35]. In physiological signals studies, labelled emotional data enables the correlation of bodily reactions, such as changes in pupil size or heart rate, with specific affective states.

In this thesis, emotion labelling (see Fig. 1.1) forms the basis for analysing participants' affective states through survey responses. This approach provides deeper insights into emotional dynamics and their impact on cognition and behaviour by systematically mapping emotions. Moreover, precise emotion labelling is crucial for advancing

emotion detection technologies, which are increasingly integrated into various fields, particularly clinical settings.

Building on this foundation, the following section examines how physiological signals can be utilised for emotion detection, offering an objective and continuous method for assessing emotional states beyond self-reported data.

1.1.2 Emotion detection using Physiological signals

Physiological signals are measurable indicators of the body's internal processes, providing valuable insights into health and emotional states [36], [37]. These signals are captured using sensors and devices that track changes in biological parameters [38].

Emotions influence physiological processes through the autonomic nervous system (ANS), which regulates involuntary functions. The ANS consists of the sympathetic nervous system (SNS), responsible for arousal and "fight-or-flight" responses, and the parasympathetic nervous system (PNS), which governs relaxation and "rest-and-digest" states [39]. Emotional states trigger unique physiological changes, detectable through bio-signals such as pupil size, heart rate (HR), respiration, electrodermal activity (EDA) or galvanic skin response (GSR), and Facial Emotion[1].

Pupil Size

Pupil size, or pupillary diameter, refers to the width of the central opening of the iris [40]. Typically measured in millimetres (mm), it fluctuates due to lighting conditions, emotional states, and cognitive load [41]. Manual pupillometers, automated eye trackers, and digital cameras capture baseline and dynamic changes.

The ANS regulates pupil size, with pupils constricting in bright light to protect the retina and dilating in low light to enhance vision [42]. Beyond these reflexive responses, pupil size changes in reaction to emotional and cognitive states [43]. Dilation occurs during heightened arousal, stress, excitement, or increased mental effort, while constriction is associated with relaxation or focused attention [44].

Pupil size measurements are widely applied in psychology, HCI, and medical diagnostics. In psychology, they help assess emotional responses and cognitive load [45], [46]. In HCI, pupil size data informs user engagement and workload evaluation [47]. In medical settings, abnormal pupil responses can indicate neurological disorders or drug effects [48], [49].

Eye trackers are used to precisely measure eye movement and pupil size, for example, through infrared light and high-resolution cameras [50], [51]. They can provide real-time data for research, education, and medical applications [52], [53]. Algorithms

[1]Here we mentioned everywhere facial emotions/expressions as physiological signals, but to be precise, it's a behavioural expression.

analyse the reflections to detect pupil boundaries and calculate the diameter.

Eye tracking technology is widely used in psychological and cognitive research, where pupil dilation is a marker for arousal and mental effort [54]–[56]. In HCI, adaptive interfaces adjust content complexity based on pupil size variations [47]. In medicine, eye trackers assist in diagnosing neurological disorders [57], [58]. In education, they evaluate student engagement and cognitive responses for personalised learning strategies [59], [60].

Galvanic Skin Response

Galvanic Skin Response (GSR), or EDA or skin conductance, measures the skin's electrical conductivity in response to stimuli [61], [62]. The principle behind GSR is that skin conductivity changes with moisture levels, influenced by sweat gland activity [63]. Emotional arousal triggers subtle sweating, increasing electrical conductivity [64]. Unlike voluntary responses like eye movements or facial expressions, sweat gland activity is autonomic, making GSR a reliable indicator of arousal without requiring conscious self-reporting [65], [66].

GSR is measured by applying a small electrical current and recording skin resistance across two electrodes (see Figure 1.2) [63]. The GSR signal has two components:

- Skin Conductance Level (SCL) – slow, long-term changes (tonic response) used for baseline analysis [67].
- Skin Conductance Response (SCR) – rapid, event-driven fluctuations (phasic response) indicating arousal events [68].

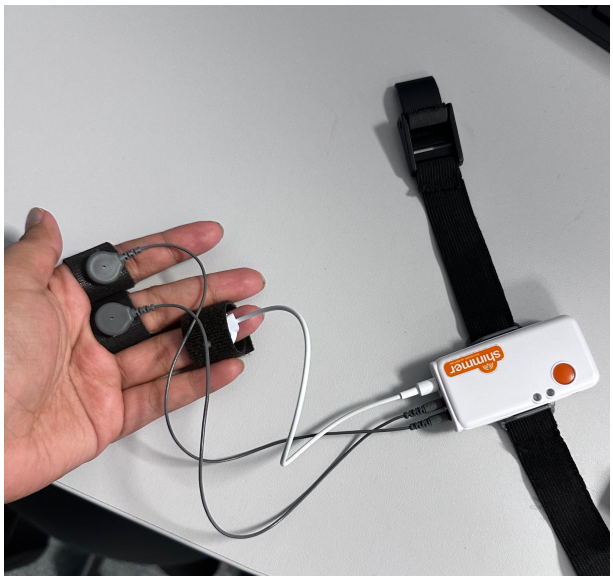


Figure 1.2: Data collection using GSR (Shimmer).

Photoplethysmography-based heart rate

Photoplethysmography (PPG) is a non-invasive optical technique that measures blood volume changes by detecting light absorption or reflection variations [69]. A light source and photodetector placed on the skin (e.g., fingertip, earlobe, or wrist) capture HR and blood oxygen levels (SpO_2) [70].

PPG operates in two modes:

- Transmission mode – light passes through translucent tissue, with a sensor on the opposite side.
- Reflectance mode – light reflects off the skin, with both sensor components on the same side [71].

Widely used in cardiovascular monitoring, pulse oximetry, and wearable devices, PPG is gaining attention in emotion detection, detecting physiological responses to stress, excitement, and relaxation [72], [73]. Unlike self-reports, PPG provides real-time, objective emotional tracking, benefiting mental health, HCI, and affective computing [74]. However, motion distortions and skin pigmentation can affect accuracy, though advancements in signal processing and ML are improving reliability [75].

Facial Emotion Recognition

Facial expressions are one of the most natural and intuitive ways humans communicate their feelings, often reflecting emotional states even when verbal communication is absent. By analysing facial movements, such as eyebrow raises, lip curvature, and eye-widening, researchers can identify key emotional markers associated with happiness, sadness, anger, surprise, fear, and disgust [76], [77]. Like physiological signals, automatic Facial Emotion Recognition (FER) can also detect emotion. It plays a crucial role in affective computing, a field dedicated to processing emotional information through computational means [78]. Facial expressions are a powerful non-verbal communication, often conveying emotions more effectively than words [79]. It enables real-time emotion detection in applications such as virtual assistants, adaptive learning environments, and customer experience analysis [80].

In psychological research, FER offers objective evaluations of emotional reactions, minimising dependence on self-reported data, which could be biased or inaccurate. Additionally, combining facial expression analysis with physiological signals can enhance emotion detection accuracy, offering a more comprehensive understanding of an individual's affective state [36].

Use of Emotion Detection in Clinical Applications

Emotion detection is increasingly recognised as a valuable tool in clinical settings, offering new opportunities for diagnosing, monitoring, and treating various psychological and neurological conditions [81], [82], [83]. Traditional methods of assessing emotional states, such as self-reports and clinician observations, can be biased or inconsistent [84], [85]. Automated emotion detection, primarily through physiological signals, provides an objective and continuous measure of emotional states, improving the accuracy and reliability of emotional assessments [36], [86].

In clinical psychology, emotion detection aids in the early detection of mood disorders such as depression and anxiety by identifying subtle physiological changes linked to emotional dysregulation [87]. For individuals with alexithymia, borderline personality disorder (BPD), or schizophrenia, who may struggle with verbal emotional expression, emotion detection systems that analyse facial expressions, physiological responses, and other biometrics can provide valuable insights [88].

Emotion detection technologies also have applications in personalised mental health interventions, such as biofeedback therapy, where real-time emotional monitoring can help patients develop better emotional regulation strategies [89]. In neurorehabilitation, emotion detection can track emotional responses in individuals recovering from brain injuries, allowing for adjustments to therapeutic approaches [90].

By integrating emotion recognition into clinical practice, healthcare professionals can better understand a patient's emotional well-being, leading to more precise diagnoses and personalised treatment plans [91]. The advancement of Machine learning (ML) models trained on multimodal physiological signals further expands the potential of emotion detection in revolutionising mental health care and HCI.

1.1.3 Machine Learning

ML has emerged as a dominant approach in emotion detection due to its ability to automatically learn patterns from large datasets. Unlike traditional methods that rely on predefined rules, ML algorithms—intense learning models—can identify complex relationships and subtle features within data, such as facial expressions, speech, and physiological signals. ML models can detect emotions through feature extraction by training on labelled datasets, gradually improving their accuracy as they are exposed to more data [92]–[94].

The training process typically begins with collecting diverse datasets containing labelled emotional expressions across various modalities, including facial images, voice recordings, and physiological data. These labelled datasets serve as the ground truth, allowing the model to learn associations between patterns in the data and corresponding emotional states.

Various ML techniques are employed in the training process. Supervised learning is commonly used, where models are trained on labelled data to classify emotional states or predict emotional intensity. Convolutional Neural Networks (CNNs) are often applied for FER, as they excel at identifying spatial patterns in image data. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are well-suited for analysing sequential data like speech, as they capture temporal dependencies in vocal tone and rhythm, critical for emotion detection [95], [96].

After training, the model is evaluated on a separate dataset to assess its generalisability and ensure it can accurately predict emotions in new, unseen data. The evaluation process allows continuous refinement, with the model becoming more precise as it is exposed to a broader range of emotional expressions. Model performance is then evaluated using $accuracy$, $precision$, $recall$, $F1-score$, $R2-Score$, $NRMSE$, and correlation metrics, which measure how well the model predicts emotional states and handles discrepancies between expected and actual labels [97], [98].

In clinical applications, such as emotion-aware therapy or mental health monitoring [93], [94], these trained models have the potential to offer real-time feedback on a patient's emotional state [99]. This, in turn, can enable clinicians to adjust interventions dynamically, enhancing patient engagement and the effectiveness of therapeutic sessions [100], [101]. Furthermore, emotion detection models can be particularly beneficial for individuals with conditions like acquired alexithymia, where patients struggle to recognise or express emotions. By analysing physiological signals or facial expressions, these models can help individuals become more aware of their feelings, promoting emotional regulation and better social functioning [102], [103].

However, using ML for automatic emotion detection presents several challenges. One of the primary difficulties is the variability in emotional expression across individuals and cultures, making it hard to develop models that generalise well across diverse populations [104], [105]. Additionally, emotions are inherently complex and often ambiguous [106], as they can manifest through subtle facial expressions, vocal intonations, physiological signals, or a combination of these modalities [107]. This necessitates robust feature extraction and fusion techniques to effectively capture relevant emotional cues.

Another significant challenge is data quality and availability; emotion detection models require large, labelled datasets, yet accurate and consistent labelling remains subjective and prone to bias [108]. Furthermore, ML models must handle real-world conditions, including variations in lighting [109], background noise [110]–[112], and occlusions in video or audio recordings [113]. Utilising these models in clinical settings introduces additional difficulties, as medical and psychological contexts require high reliability, explainability, and regulatory compliance. Clinical applications must ensure that emotion detection models do not misinterpret psychological states, leading to po-

tential misdiagnoses or inappropriate interventions [114]. Moreover, integrating ML-based emotion detection into healthcare workflows requires careful validation [115], alignment with clinical standards [116], and consideration of patient privacy and ethical concerns [117].

Finally, ensuring understandability and fairness in emotion detection models is crucial, as biased or opaque systems can lead to ethical concerns, particularly in applications involving mental health assessments [118], hiring processes, or HCI [119]. Addressing these challenges requires advancements in data collection, model robustness, and ethical AI practices to improve the reliability and applicability of ML-based emotion detection.

Thus, developing and refining ML algorithms for emotion detection—including large labelled datasets, noise-free signals, and continuous model updates—holds transformative potential in healthcare, education, and beyond. Automatic emotion detection can enable personalised, real-time interventions across various fields, particularly in clinical settings where timely and accurate emotional assessment can enhance patient care and mental health support.

1.2 Problem Statement

Automatic emotion detection through physiological signals, such as pupil size, GSR, PPG-based heart rate, and facial expressions, presents several key challenges. One major issue is the significant variability in physiological responses across individuals, where emotional responses may vary widely, making it difficult to generalise results. These signals are not solely affected by emotional states but are also influenced by genetics, environmental conditions, and cognitive processes, making it difficult to generalise across users.

Pupil size, for instance, is a commonly used indicator of emotional arousal, with dilation often associated with heightened emotions such as excitement or fear. However, significant variability exists due to baseline differences across individuals, which are influenced by age and neurological factors. Additionally, external factors such as changes in ambient luminosity can independently alter pupil size, regardless of emotions. Cognitive load, including concentration and mental effort, can also impact pupil dilation, making it challenging to isolate emotional responses accurately.

Similarly, GSR, which measures changes in skin conductance due to sweat gland activity, is widely used for detecting emotional arousal. However, individual physiological differences mean some people exhibit stronger GSR responses than others, leading to inconsistent readings. Environmental conditions such as temperature and humidity further affect skin conductivity, potentially resulting in misleading interpretations. Additionally, habituation effects can cause a decline in GSR responsiveness over time when

individuals are repeatedly exposed to emotional stimuli, reducing its reliability in continuous emotion detection.

PPG-based heart rate (HR) and heart rate variability (HRV) are also commonly used physiological markers for emotion detection. However, these signals vary significantly among individuals due to differences in baseline heart rate, fitness levels, and age. Non-emotional influences, such as caffeine intake, stress, and physical activity, can also cause fluctuations in heart rate, making it difficult to attribute changes solely to emotional states. Furthermore, heart rate responses to emotions often exhibit a delay, complicating real-time emotion detection applications.

Facial expressions remain one of the most widely studied indicators of emotion, but they, too, present challenges in variability. Cultural differences influence emotions, with some individuals exhibiting more subtle facial expressions than others. Social norms and personal disposition may lead individuals to mask or exaggerate their facial expressions, introducing inconsistencies in emotion detection. Additionally, variations in facial features, including ageing effects, facial hair, and occlusions, can impact the accuracy of facial expression analysis.

Tackling these challenges means creating more substantial and personalised systems for detecting emotions. This research delves further into combining different physiological signals using multimodal fusion techniques to enhance accuracy and address the limitations of individual methods. Plus, adaptive learning models that can adjust to a user's baseline physiological state could significantly enhance the reliability of emotion detection. Additionally, context-aware systems that take into account external factors such as lighting, temperature, and social settings can make emotion recognition models more adaptable in diverse situations. Addressing these hurdles can make emotion detection systems more precise, trustworthy, and valuable in many real-world scenarios.

Traditional emotion detection methods rely on classification models, which may not always effectively handle multimodal data's inherent noise and variability. To overcome these challenges, this thesis employs an ML approach with a regression model, rather than a classification approach, enabling more continuous and nuanced predictions of emotional states. By analysing multimodal physiological signals in conjunction with facial expressions, the model aims to provide more accurate emotional insights.

Furthermore, this research aims to eliminate the impact of luminosity on pupil size data to improve the accuracy of emotion detection. Another major hurdle is the integration of multiple physiological markers; ensuring reliable emotion detection through the combination of these different modalities remains a challenging task.

Overall, this research aims to enhance the reliability and robustness of emotion detection systems, providing more accurate real-time feedback for applications such as neurorehabilitation, emotion-aware therapy, and other fields of affective computing.

1.3 Research questions

1. How can we construct individualised ground truth labels for affective states that take into account the individual differences?
2. How can external factors, such as luminosity changes and environmental noise, be controlled or compensated for to improve the accuracy and reliability of emotion detection from pupil size?
3. How can regression-based models improve emotion detection by handling continuous emotional states and individual variations more effectively than traditional classification approaches?
4. How can multimodal data (e.g., pupil size, GSR, FER) be effectively integrated into machine learning models to improve the accuracy and robustness of emotion detection, and what are this approach's key challenges and benefits?
5. What techniques can optimise ML models for real-time emotion detection using simultaneous physiological signals, and what performance metrics are best suited for evaluating their effectiveness?

1.4 Contributions

This thesis makes several novel contributions to the field of multimodal emotion recognition, which can be summarised as follows:

1. **Novel multimodal configuration:** We propose a multimodal emotion recognition pipeline that integrates pupil size, GSR, and FER derived from 20 action units (AUs). Unlike most prior studies that rely on ECG/EDA or audio–visual signals, this unique combination leverages ocular dynamics, physiological activity, and facial behaviour, representing a multimodal configuration not previously reported for dimensional affect prediction.
2. **Vectorial mapping from FER to Russell's circumplex:** We introduce a vectorial transformation method with the help of students that projects basic emotions into Russell's circumplex space, where emotion categories define angular positions and intensity defines vector magnitude. Unlike traditional approaches that map directly from AUs or categorical FER outputs to valence-arousal, this method produces interpretable, continuous affective coordinates and allows mixtures of emotions to be represented simultaneously.
3. **Temporal aggregation at the vector level:** We propose a temporal averaging strategy that aggregates emotion vectors across time, yielding a stable circumplex trajectory for each stimulus. This approach preserves co-occurring emotions and avoids collapsing data into a dominant label, in contrast with frame-by-frame or regressor smoothing techniques commonly used in FER studies.

4. **Robust participant-level evaluation:** We employ leave-one-participant-out (LOO) cross-validation to explicitly address heterogeneity across participants. This ensures robust generalisation and avoids overfitting to individual-specific patterns, a methodological consideration often overlooked in previous multimodal affect recognition work.
5. **State-of-the-art performance within selected modalities:** Our experiments demonstrate that pupil size-only, GSR-only models, and FER+GSR+pupil fusion achieve concordance correlation coefficients (CCCs) superior to those previously reported for these modalities. While some studies using ECG/EDA with deep networks report higher absolute performance, our results establish a new benchmark within the studied modality set.
6. **Psychological interpretability:** Our framework grounds predictions in Russell's circumplex model, ensuring that outputs are not only accurate but also psychologically meaningful. Each dimension corresponds directly to valence and arousal, providing scientific transparency and practical relevance. This balance of performance and interpretability is rarely addressed in prior multimodal affect recognition research.

To the best of our knowledge, this thesis is the first to combine corrected pupil size, GSR, and FER-based vectorial mapping into Russell's circumplex for continuous emotion recognition. The framework offers a robust and interpretable approach to multimodal affect prediction, advancing emotion detection research in clinical and mental health contexts and supporting the development of more accurate and reliable emotion-aware AI systems.

1.5 Structure of thesis

Here's a brief overview of my thesis structure, summarising the key chapters:

- **Chapter 1: Introduction** This chapter introduces the primary focus of the research, which is the development of a novel machine-learning model for emotion detection using multimodal biomarkers. It outlines the research's motivation, objectives, and significance, highlighting potential applications in mental health and clinical settings. The research questions and the thesis contributions are also presented.
- **Chapter 2: Literature Review** The literature review chapter explores the existing body of research on emotion detection techniques, mainly focusing on physiological signals (such as pupil size, GSR, PPG-based HR) and FER. It covers traditional methods, challenges in multimodal emotion detection, and the application of ML in emotion detection. Additionally, the chapter reviews related work on the clinical use of emotion detection, especially in mental health monitoring.

- **Chapter 3: Development of Methodology for Emotion Detection Model** This chapter details the research methodology, including the techniques and algorithms used to develop the ML model. It explains the process of collecting and preprocessing multimodal data, like the luminosity-isolation technique for pupil size correction, and integrating them with FER and GSR features. The ML model training, including regression-based approaches, is also discussed.
- **Chapter 4: Results** The results chapter presents the experiments' outcomes using the proposed model. It includes performance metrics, comparisons with baseline methods, and evaluations of the model's ability to predict emotions based on multimodal physiological data. The chapter also covers the results from analysing the impact of luminosity isolation on pupil-based emotion detection.
- **Chapter 5: Discussion and Future Work** This chapter discusses the implications of the results, highlighting the strengths and limitations of the developed emotion detection model. It addresses the challenges faced when integrating multiple physiological signals, such as variability across individuals and the influence of external factors. The chapter also compares the model's performance with other existing approaches, discusses the clinical applications in mental health and neurorehabilitation, and works for future research.
- **Chapter 6: Conclusion** The final chapter summarises the key findings of the research and outlines the contributions of the thesis. It reflects on the potential clinical applications, particularly in mental health monitoring and discusses areas for future research. The chapter concludes by emphasising the broader implications of emotion-aware AI systems for improving patient care and therapy outcomes.

This structure provides a clear and logical flow of information, guiding the reader from the background and theoretical foundations to the research's practical contributions and future directions.

1.6 Publications

No.	Title of Publication	Journal/Conference/Year	Status
1	Towards an Accurate Measure of Emotional Pupil Dilation Responses: A Model for Removing the Effect of Luminosity	IEEE Metroxraine, 2024	Published
2	Quantifying Emotional Arousal through Pupillary Response: A Novel Approach for Isolating the Luminosity Effect and Predicting Affective States	IEEE Transactions on Affective Computing, 2025	Submitted

Table 1.2: List of Publications

Chapter 2

Literature Review

In this section, I have provided a comprehensive review of relevant studies related to the key topics addressed in my thesis, including emotion theory and classification, the role of physiological signals in emotion detection, the application of ML techniques for emotion detection, and the use of multi-modal approaches in emotion detection systems.

2.1 Emotions and Emotion Classification

Emotions are fundamental to human experience [120], influencing thoughts [121], behaviours [122], and interactions [122]. They are complex states involving physiological arousal, subjective experiences, and expressive behaviours. The concept of emotion is multifaceted, with diverse interpretations across psychology, neuroscience, and philosophy. Historically, emotions were often viewed as irrational forces needing control [123]. Over time, theories evolved from physiological explanations (James-Lange theory) to cognitive approaches and modern neuroscientific studies [124], [125]. Cultural and social contexts also shape emotional experiences and interpretations [126].

Aristotle defined emotions as involving pain, pleasure, or both. Modern dictionaries describe them as strong feelings like love, anger, or fear [127], [128]. The study of emotions continues to evolve, reflecting the complexity of human experience and the interplay of biological, psychological, and social factors.

Philosophers debate the nature of emotions, questioning their origins and authenticity. Charles Darwin's pioneering work proposed that emotions are universal, evolutionary adaptations crucial for communication and survival [129]. Darwin revealed their innate, shared characteristics by comparing human and animal emotional expressions [129]. His research continues to influence the contemporary understanding of emotions across psychology, anthropology, and neuroscience, highlighting emotions as fundamental biological and social mechanisms.

Schachter-Singer's Two-Factor Theory of Emotion proposes that emotions result from a two-step process: physiological arousal followed by cognitive labelling [130]. When confronted with an emotional stimulus, such as a snarling dog, the body first reacts physiologically (e.g., increased heart rate), and the resulting emotion (e.g., fear) depends on how this arousal is interpreted in context. This theory underscores the dynamic interplay between bodily responses and cognitive processes in shaping emotional experiences.

Emotions remain a continuously evolving topic across scientific [120], cultural [131], technological, and personal domains [132]. Their recognition and understanding are fundamental not only for emotional regulation [133] but also for supporting mental health by reducing stress and anxiety, and enhancing self-awareness [134]. In applied contexts, such as leadership, teamwork [135], and emotion-aware systems [136], accurate emotion detection supports better decision-making by addressing emotional influences and biases [137]. Furthermore, understanding how cognitive appraisal shapes emotions, as illustrated by the Schachter-Singer model, highlights the need for emotion detection frameworks to incorporate both physiological signals and individual cognitive interpretations.

Given the importance of emotion detection, various classification methods and frameworks have been developed to categorise and understand the complexity of human emotions.

Some prominent emotion classification methods include:

Basic Emotion Models. These models attempt to identify a limited set of fundamental and universal emotions biologically ingrained and consistently recognisable across cultures. They categorise emotional experiences into core affective states, often based on evolutionary or behavioural evidence. Some of the best-known models include:

- **Paul Ekman's Six Basic Emotions** [138]: Ekman identified six universal emotions—happiness, sadness, anger, fear, disgust, and surprise—based on facial expressions observed across diverse cultures.
- **Robert Plutchik's Wheel of Emotions** [139]: Plutchik proposed eight primary emotions—fear, sadness, anger, joy, surprise, disgust, anticipation, and trust—arranged in a wheel-like structure to illustrate the relationships between emotions, including their intensities and opposites.

Dimensional Models. In contrast to categorical approaches, dimensional models represent emotions within a continuous affective space. These models are particularly valuable in computational modelling and affective computing, where emotional states are better captured along gradations rather than as discrete labels. Two widely cited dimensional models are:

- **Russell's Circumplex Model** [32]: Emotions are mapped within a two-dimensional circular space defined by:
  - **Valence**: the degree of pleasantness (positive vs. negative emotions).
  - **Arousal**: the level of activation or intensity (high vs. low).

  Conceptually similar emotions appear close together in this space, while opposites are positioned across.

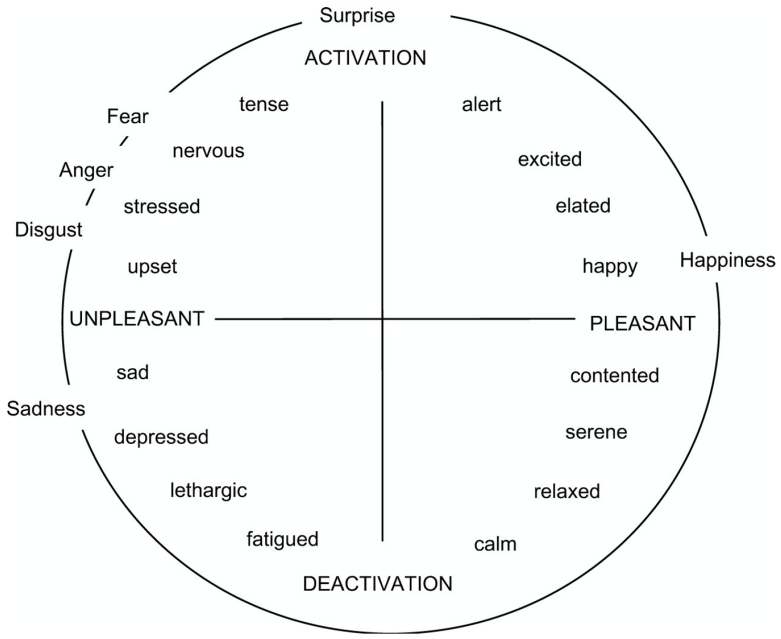


Figure 2.1: Russell's Circumplex Model of Affect.

- **Lövheim's Cube Model** [140]: A three-dimensional model incorporating:
  - **Valence** (pleasure–displeasure)
  - **Arousal** (activation–non-activation)
  - **Dominance** (control–submission)

  This framework is particularly relevant in physiological and neurochemical studies of emotion, as it enables the modelling of emotional states in terms of neurotransmitter influences.

Critical Evaluation of Dimensional Models. Dimensional approaches address some of the limitations of categorical models by representing emotions on a continuous spectrum, which better reflects the gradations and overlaps observed in real emotional experiences [141]. For computational modelling, this flexibility is especially advantageous, as it enables algorithms to capture subtle variations in affect rather than forcing discrete categories. However, dimensional models are not without limitations. Their abstraction into broad axes such as valence and arousal can sometimes oversimplify the richness of emotional phenomena, neglecting discrete emotions that carry unique social

or evolutionary significance (e.g., disgust or pride) [141]. Furthermore, dimensional spaces can vary depending on the methodology used to derive them, raising questions about their universality and comparability across studies [34]. The reliance on self-reported ratings for positioning emotions in these spaces also introduces subjectivity, which may affect the reliability of ground-truth labels.

Rationale for Using Russell's Circumplex Model. Among dimensional models, Russell's Circumplex Model is particularly influential and widely adopted because of its balance between parsimony and explanatory power. The two axes of *valence* and *arousal* capture the majority of variance in affective experience and align closely with underlying neurophysiological processes [32]. Compared to more complex frameworks such as Lövheim's Cube Model, the circumplex offers an intuitive and empirically validated structure that is easier to implement in computational settings and more accessible for participants in self-report surveys. This simplicity makes it especially effective in experimental contexts where participants are repeatedly rating stimuli, as cognitive load is minimised.

Moreover, the circumplex framework facilitates the integration of multimodal data. Physiological signals such as GSR and HRV strongly correlate with arousal, while EEG markers and facial expressions contribute to valence estimation, making the two-dimensional structure well-suited for multimodal fusion in affective computing [142]. While the omission of a dominance/control dimension (as included in models such as Lövheim's) may limit granularity in some applications, the circumplex strikes a practical balance between interpretability, computational efficiency, and empirical robustness. For these reasons, Russell's model was selected as the foundation for emotion labelling in this study. Its broad applicability, empirical support, and efficiency make it a powerful tool in psychological and computational applications. However, when studying emotions in controlled environments, such as laboratories, several challenges arise, particularly in eliciting authentic emotional responses from participants.

In laboratory settings, researchers often face the issue of eliciting and labelling emotions in a way that accurately represents genuine emotional experiences. The challenge lies in triggering genuine and consistent emotions across participants while reliably categorising them. These complexities become especially important when utilising emotion models like Russell's, as accurately labelling emotional states in real-time becomes crucial for understanding the affective responses to various stimuli. This section delves into the techniques and challenges in eliciting and labelling emotions within the lab environment, providing a foundation for understanding the subsequent model development and evaluation.

2.2 Eliciting and Labelling Emotions

Studying emotions in laboratory settings presents several challenges, including how to trigger genuine emotional responses. While emotions are natural reactions to various situations and often occur beyond conscious control, researchers in controlled environments must induce or trigger specific emotional states for study. This is typically achieved through emotional stimuli such as audio [143], video [144], and images [145]. These stimuli are carefully selected to elicit targeted emotional responses, allowing researchers to study how individuals react to specific emotions in a controlled setting. However, the artificial nature of the lab environment, combined with participants' awareness of being observed, can compromise ecological validity and influence emotional authenticity.

Pan et al. found that audio-only stimuli did not necessarily induce stronger emotional responses than visual or audiovisual stimuli [146]. Their research revealed no significant difference between audiovisual and audio-only presentations in terms of emotional response, with visual-only stimuli occasionally being more effective. This suggests that audio stimuli alone may not be as powerful as visual or combined audiovisual stimuli in eliciting emotions. Further supporting this, research on tracking accuracy tasks found visual cues more effective than auditory cues, indicating that visual information can often have a more substantial impact than audio in specific contexts [146].

Similarly, Meike K. Uhrig et al. compared the effectiveness of pictures and film clips in eliciting emotional responses, challenging the belief that films are inherently more effective [147]. They found that pictures, particularly sequences of three congruent images, were more effective in generating stronger emotional responses and arousal [147]. This highlights the importance of carefully selecting emotion-research stimuli for reliable emotion induction.

Other studies have explored the effectiveness of various stimuli for emotion induction. McGinley and Friedman compared imagery, recall, and film clip viewing, finding that film clips were the best-performing technique in predicting emotion-specific patterns of ANS activation [148]. A meta-analysis by Westermann et al. also confirmed that film clips outperformed other emotion induction techniques in eliciting positive and negative emotional states [149].

A study comparing slides and video clips for emotion elicitation revealed that video clips were more effective in inducing arousal, particularly for erotic and fear-inducing content [150]. Participants reported greater self-perceived arousal after viewing fear and erotic video clips compared to slides of the same categories. Additionally, Laroche et al. found that animated images (similar to short video clips) induced higher levels of pleasure than static images in an online context [151].

These findings suggest that dynamic stimuli like videos, due to their multimodal

nature (visual and auditory) and ecological validity, are often more effective than static images in eliciting emotions. However, some studies also indicate that videos may not always be more effective for emotion induction. For example, a survey by Bartolini investigated the impact of film clip duration on emotion elicitation, revealing that longer clips were more effective in evoking positively valenced emotions. In comparison, shorter clips had a more substantial impact on negative emotions [152]. This emphasises the importance of stimulus duration in shaping emotional responses.

Complementing this, research on standardised Chinese emotional short videos showed that clips between 60 to 240 seconds effectively elicit specific emotions, with a mean clip duration of 148.69 seconds [153], [154]. This research also highlighted the importance of considering demographic factors, such as age and gender, when selecting video stimuli.

Critically, while a substantial body of evidence supports the effectiveness of certain stimulus types (e.g., videos, film clips) in eliciting emotions, generalising these findings across populations and contexts remains a challenge. Emotional responses are shaped not only by stimulus features but also by individual differences, cultural norms, and situational contexts. Furthermore, studies relying on self-reports of arousal or valence may not capture more subtle or mixed emotions. Thus, while multimedia stimuli offer strong potential for emotion elicitation, they require careful design, cultural calibration, and complementary validation through physiological or behavioural measures.

As emotion elicitation is already challenging, determining the ground truth emotional state adds a layer of complexity to emotion research. The most common approach to determining ground truth in emotion research is through self-assessment questionnaires. Nevertheless, this approach may not consistently provide accurate results because it depends on people's subjective interpretations of their emotional conditions, which can be affected by factors such as mood, level of self-awareness, and a tendency for social desirability bias and cognitive bias. Cognitive bias is a systematic pattern of deviation from rational judgment, where individuals perceive, interpret, or remember information in a way that is influenced by their beliefs, emotions, or past experiences rather than by objective evidence [155]. While often helpful for quick decision-making, these mental shortcuts can introduce perception, memory, and emotional interpretation errors. Considering cognitive biases during emotional labelling helps account for these distortions, thereby improving the accuracy and robustness of the model by aligning it more closely with how humans process emotional information [156]. Despite its limitations, self-assessment remains a widely used choice due to its simplicity and ease of implementation.

Several types of questionnaires have been used for the emotional detection study. The PANAS (Positive and Negative Affect Schedule), developed by Watson, Clark, and

Tellegen [157], is a psychometric tool designed to measure two independent dimensions of affect: Positive Affect (PA) and Negative Affect (NA). It consists of 20 items, with 10 each for PA (e.g., enthusiasm, alertness) and NA (e.g., distress, fear), rated on a 5-point Likert scale. Scores for PA and NA are calculated separately, with higher scores indicating greater intensity of the respective effect. The PANAS is validated for high internal consistency (Cronbach's alpha: 0.86–0.90 for PA and 0.84–0.87 for NA) and test-retest reliability. Its brevity, reliability, and flexibility across time frames make it widely used in research to assess emotional states.

A study by Bradley and Lang [158] compares two tools for assessing emotional responses: the Self-Assessment Manikin (SAM) and the Semantic Differential (SD) scale. SAM is a non-verbal, pictorial tool that uses simple graphic figures to measure three dimensions of emotion: pleasure, arousal, and dominance. It requires only three judgments, making it efficient and suitable for diverse populations, including non-English speakers and children. In contrast, the SD scale, developed by Mehrabian and Russell [159], uses 18 bipolar adjective pairs (e.g., "happy-sad") to assess similar emotional dimensions. The study found high correlations between SAM and SD ratings for pleasure and arousal. Still, differences in the dominance dimension were noted, where SAM appeared to better capture personal responses to stimuli. SAM was highlighted as a quicker, more accessible method for measuring affective responses across various contexts.

The Cognitive Emotion Regulation Questionnaire (CERQ), developed by Garnefski et al. [160], is a widely used self-report tool designed to assess individuals' cognitive strategies to regulate emotions after experiencing stressful or adverse events. The CERQ consists of 36 items divided into nine subscales, representing both adaptive strategies (e.g., acceptance, positive refocusing, positive reappraisal, putting into perspective, and refocus on planning) and maladaptive strategies (e.g., rumination, catastrophizing, self-blame, and blaming others). Each item is rated on a 5-point Likert scale, with higher scores indicating more significant use of the corresponding strategy.

The CERQ has been extensively validated across various populations and languages, demonstrating strong psychometric properties such as internal consistency, reliability, and construct validity. A shorter 18-item version (CERQ-short) has also been developed for quicker assessments while preserving the original nine-factor structure. The CERQ is particularly useful in clinical and research settings for understanding how cognitive emotion regulation strategies influence mental health outcomes, such as anxiety, depression, and overall well-being. Its ability to distinguish between adaptive and maladaptive strategies makes it an effective tool for identifying emotion regulation patterns and informing interventions.

Izard's Differential Emotions Scale (DES) is a self-report tool designed to measure ten fundamental emotions (e.g., joy, anger, fear, guilt) using 30 items rated on a 5-point

Likert scale [161]. While the DES has been widely used and shows stability in measuring emotional factors, its reliability and validity have mixed findings. Studies demonstrate that the DES generally maintains high intercorrelations and aligns with theoretically defined emotional factors, supporting its construct validity. However, criticisms include low internal consistency for some sub-scales due to the limited number of items per emotion and potential response biases inherent in self-report measures. Factor analyses have supported many of the proposed emotions, but suggest unclear construct validity for certain sub-scales. Additionally, the DES has been criticised for overemphasising negative emotions and excluding low-energy states like serenity, which limits its comprehensiveness. Despite these limitations, it remains helpful in research on emotional states and their relationship to behaviours and psychological conditions.

It is important to critically recognise that while self-report instruments like PANAS, SAM, CERQ, and DES provide accessible means of capturing emotional states, they inherently rely on participants' self-awareness, introspection, and willingness to respond truthfully. As such, they may fail to capture transient, unconscious, or socially masked emotions. Additionally, questionnaire fatigue and interpretation variability can undermine data quality. To overcome these limitations, self-reports should ideally be complemented by physiological, behavioural, or multimodal data sources that can offer more objective or continuous emotion indicators.

These self-assessment tools are essential in emotion research, offering insight into individuals' emotional states. However, their accuracy can vary depending on individual differences and context, making it necessary to scale their ground truth individually [162]. This highlights the importance of multidimensional emotional ratings, which account for the complexity of emotional experiences and provide a more nuanced understanding of emotion detection.

A study by E. David et al. exemplifies this approach with the Multidimensional Emotional Questionnaire (MEQ), a self-report tool designed to comprehensively assess emotional experiences [163]. By evaluating two overarching dimensions of emotional reactivity (positive and negative), three components of reactivity (frequency, intensity, and persistence), and ten discrete emotions, the MEQ captures the richness of individual emotional variation. Additionally, its assessment of emotion regulation abilities ensures a complete analysis of emotional responses. Validated through multiple studies with strong psychometric properties, the MEQ demonstrates how multidimensional emotional ratings enhance the accuracy and reliability of self-assessment tools, ultimately improving individual survey computation in emotion research.

Building on this scaling approach and the reference study by Jongwan Kim et al., we incorporated a multidimensional assessment methodology. In their research, participants watched a series of video stimuli and rated their emotional responses along six

dimensions: excitement, positivity, calmness, anxiety, negativity, and sadness [164].

After each video, participants provided ratings on these six emotional dimensions, creating a comprehensive emotional profile for each stimulus. The researchers constructed a correlation matrix for each participant to analyse this data, encompassing the 32 stimuli across all six emotional ratings.

Based on these insights, we decided to use multidimensional emotional questionnaires for our study, ensuring a more detailed and reliable assessment of emotional responses.

Ultimately, combining rigorous stimulus selection with multidimensional and individualised emotion labelling strategies allows researchers to move beyond oversimplified emotion models. This approach increases ecological validity, captures the depth of emotional experience, and facilitates more accurate emotion detection across diverse individuals and contexts.

2.3 Overview of Emotion Detection Using Physiological Signals

Physiological signals are widely used in emotion detection to enhance accuracy and objectivity, complementing self-reported emotional ratings. These signals include measures like Electroencephalogram (EEG) for brain activity, Electrocardiogram (ECG) and photoplethysmogram (PPG) for heart-related metrics such as HRV, GSR for emotional arousal (see Figure 2.2), Electromyogram (EMG) for muscle activity, respiration amplitude, and Facial Emotion Recognition (FER), all of which provide real-time insights into emotional states and regulation [165], [166].

Emotion detection methods aim to decode human emotions by examining physiological changes in the central nervous system (CNS) and ANS. Various signals reflect underlying emotional states, such as pupil dilation, GSR, HRV, respiration, and brain activity, all driven by the CNS and ANS. These physiological responses offer valuable insights into the connection between emotional experiences and bodily processes.

According to the literature, pupil dilation, GSR, and HRV are strong indicators of emotional arousal [167] (see Figure 2.2), while facial expressions are markers of emotional valence [168]. Brain activity, measured through EEG, provides insights into arousal and valence [169]. Integrating these physiological signals enhances the accuracy of emotion detection.

FER enhances emotion detection by analysing facial expressions to classify emotions such as happiness, sadness, anger, and fear [170]. It improves user experience by enabling adaptive interactions [171], monitors engagement in educational settings [172], enhances surveillance through behavioural analysis [173], [174], and aids businesses

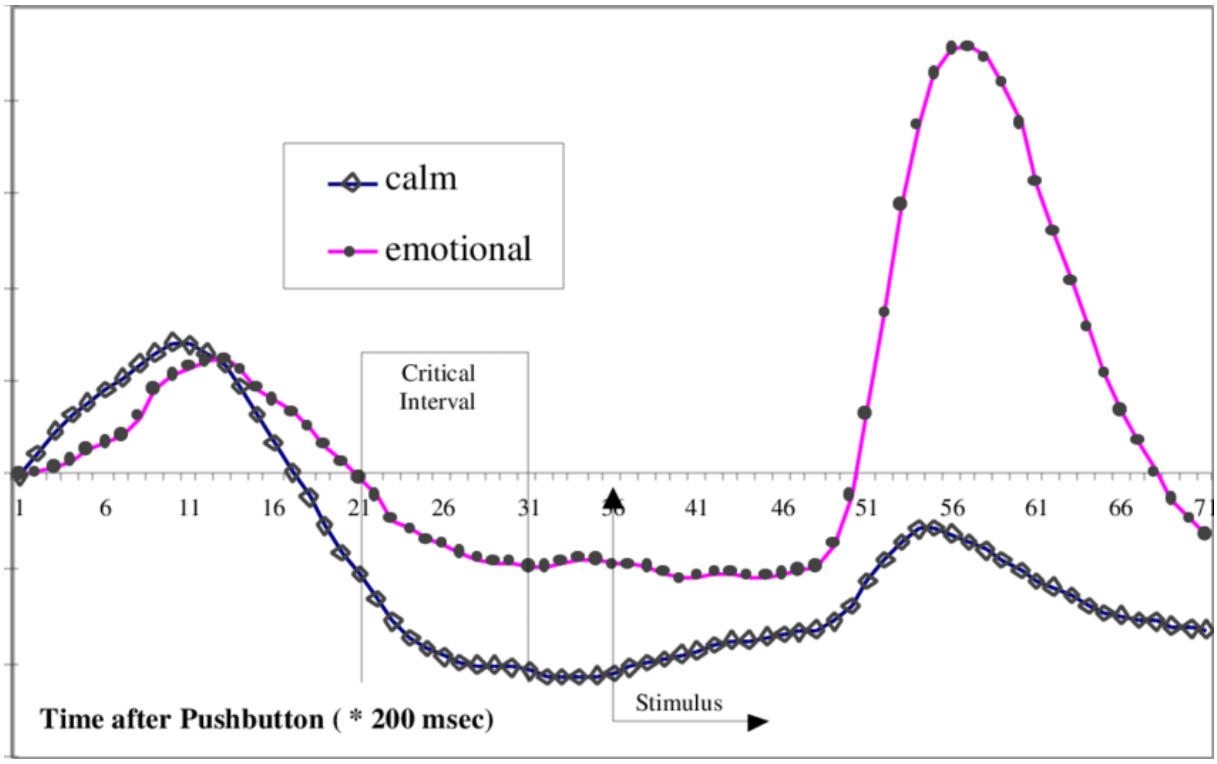


Figure 2.2: GSR response on different emotions [167].

in assessing customer reactions [172], [174].

FER systems can be categorised into static image FER and dynamic sequence FER based on feature representation. Static-based methods focus on extracting spatial features from individual images, encoding only the information in a single frame. These methods are more straightforward and computationally efficient but lack temporal context, making them suitable for tasks where expressions are analysed in isolation [175]. In contrast, dynamic-based methods consider the temporal relationships among contiguous frames in a sequence, capturing the evolution of facial expressions over time. These approaches leverage temporal models, such as Hidden Markov Models (HMMs) or RNNs, to analyse motion and intensity variations in expressions, providing richer emotional context [176], [177]. Dynamic FER is particularly effective in real-world applications where emotions are expressed as continuous sequences, such as video-based interactions or behavioural analysis. Both approaches have their strengths and limitations, with static FER excelling in simplicity and dynamic FER offering greater contextual accuracy.

Traditional methods for FER primarily relied on handcrafted features and shallow learning techniques, such as Local Binary Patterns (LBP) [178], LBP on Three Orthogonal Planes (LBP-TOP) [179], Non-Negative Matrix Factorisation (NMF) [180], and sparse learning [181]. These approaches focused on extracting spatial and texture-based features from controlled environments, often limiting their generalisability to real-world scenarios [171], [182]. However, since 2013, the emergence of large-scale

datasets like FER2013 [183] and Emotion detection in the Wild (EmotiW) [184] has facilitated the transition of FER systems from lab-controlled settings to "in-the-wild" conditions. These datasets provide diverse and challenging real-world facial data, enabling the development of deep learning-based methods that leverage CNNs and other advanced architectures to improve recognition accuracy under varying lighting, occlusions, and head poses [185]. This shift has significantly enhanced FER's applicability in real-world scenarios like HCI and behavioural analysis.

Despite its potential, FER faces challenges such as cultural variability, occlusions (e.g., masks), and privacy concerns. However, integrating physiological signals enhances accuracy and expands applications [186]. Physiological measures like HRV, pupil dilation, and GSR/skin conductance (SC) are involuntary, making them less susceptible to conscious control and providing unbiased emotional assessments [187]. Studies show that physiological responses occur even when emotional stimuli are processed unconsciously, highlighting their reliability in emotion detection [188]. As these signals are less susceptible to conscious control, they offer a reliable means of continuously monitoring emotions, particularly in clinical settings and long-term studies [189], [190]. Additionally, unlike facial expression analysis, physiological measures are not influenced by cultural variability, conscious control, occlusions, and limited emotional granularity, making them a valuable tool for emotion detection [187]. These signals capture distinct emotional dimensions like arousal, valence, and dominance [191]. In a study by S. Jerritta et al., physiological signals are highlighted as essential tools for emotion detection due to their involuntary and reliable nature, reflecting the underlying ANS and CNS responses to emotions [191].

Physiological signals can be detected through various sensors, which are either wearable or mounted on computer screens. Wearable sensors, such as wristbands, chest straps, or adhesive patches, are commonly used for monitoring signals like HR, GSR, aka EDA, aka SC, respiration, and EMG in real time. These devices are compact, non-invasive, and suitable for continuous monitoring in real-world scenarios, enabling applications like stress detection and emotion detection [192]. For example, wearable systems have been developed to monitor HR and SC during activities like driving or exercise, offering robust data collection while allowing free movement [192], [193]. Mounted sensors on computer screens or other fixed devices are less common but can capture physiological signals like gaze patterns or facial expressions for emotion detection in controlled environments. Both wearable and mounted sensors play a vital role in capturing physiological signals for applications in healthcare, sports performance monitoring, and emotion detection [194].

Pupil size, often measured using eye-tracking devices placed on or near the computer screen, plays a crucial role in emotion detection as it reflects both autonomic and cognitive processes associated with emotional states [195], [196]. Eye-trackers can

capture various data, including pupil dilation, gaze direction, blink rates, and fixation points [197]. These provide valuable insights into a person's emotional and cognitive responses during stimulus exposure. Emotionally arousing auditory or visual stimuli consistently result in more significant pupil dilation than neutral stimuli. For example, pupil size begins to dilate within 200 milliseconds following the release of noradrenaline (a neurotransmitter and hormone) in response to emotional arousal [198], [199]. This rapid response makes pupil size an effective physiological marker for assessing emotional reactions. The ANS controls pupil dilation in response to emotional stimuli through the interplay of its two branches [200]. The SNS triggers pupil dilation (mydriasis) via norepinephrine, enhancing vision during arousal or stress [201]. The PNS regulates pupil constriction (miosis) by acting through the oculomotor nerve and the neurotransmitter acetylcholine. This function is most active during calm or restful states, helping to adjust the eyes for near vision and reduce light intake [202]. However, research on the valence-specific effects of emotional stimuli shows mixed results. While some studies suggest that negative stimuli (e.g., crying or fear-inducing images) elicit larger pupil dilations than positive stimuli (e.g., laughter or joyful expressions), possibly due to the survival-related importance of negative emotions [203], others report similar pupil responses to both positive and negative stimuli when arousal levels are comparable [204].

Pupil dilation is generally considered an autonomic response to emotional arousal rather than a direct indicator of cognitive-emotional processing [205]. Its strong correlation with arousal levels makes it a reliable measure for assessing the intensity of emotional experiences, regardless of their valence. This characteristic has made pupil size an invaluable tool in emotion detection research and applications, particularly in contexts requiring non-invasive and real-time measurements. A study by Nicola et al. further explores the significance of pupil size as a metric for understanding brain states, emphasising its utility in assessing arousal, cognition, and neural function [206]. Pupil size, influenced by both light-driven [207] and brain-driven processes [208]–[210], fluctuates in response to ongoing brain activity and serves as a widely recognised measure of arousal and autonomic function. The study highlights the role of pupil-linked arousal in neural and cognitive processes, linking it to neuro-modulators like noradrenaline and orexin [211]. These fluctuations are proposed to reflect interconnected feedback loops within the brain, offering insights into arousal mechanisms. The authors stress the importance of well-defined tasks, neuro-computational models, and physiological probing to refine interpretations of pupil size as an indicator of brain activity. This work underscores pupil size as a low-cost yet powerful tool for basic research, clinical applications, and home monitoring.

GSR and PPG, commonly measured by wearable devices, are two other physiological signals critical for emotion detection.

Galvanic Skin Response (GSR) reflects changes in sweat gland activity regulated exclusively by the sympathetic branch of the ANS [212], [213]. GSR responses can be decomposed into two components: a *tonic* (slow-varying) component that reflects baseline arousal levels and a *phasic* (rapid) component that corresponds to discrete stimulus-evoked responses [214]. Phasic responses typically occur with a latency of 1–5 seconds following an emotional stimulus and last several seconds, providing a temporally precise measure of arousal [215]. While GSR is highly sensitive to emotional intensity and reliably differentiates high- versus low-arousal states (e.g., stress, excitement, or fear) [216], it cannot on its own distinguish between positive and negative valence [217]. Nonetheless, its strong temporal sensitivity makes it one of the most widely used autonomic markers in emotion research.

Photoplethysmography (PPG) measures blood volume changes in peripheral vasculature using light absorption, and it is strongly influenced by both sympathetic and parasympathetic branches of the ANS. From PPG, several cardiovascular indices can be extracted. Heart rate (HR) reflects overall arousal levels, while heart rate variability (HRV) captures the balance between sympathetic activation and parasympathetic regulation, making it informative for both arousal and valence dimensions [218]. For example, reduced HRV has been linked to stress, anxiety, and negative affect, whereas higher HRV is associated with relaxation, positive affect, and greater emotional regulation capacity [219]. Pulse rate variability (PRV), derived from PPG as an alternative to HRV, provides similar insights with high temporal resolution and is particularly suitable for wearable applications [220].

Together, GSR and PPG provide complementary perspectives on emotion: GSR delivers rapid, stimulus-locked markers of arousal, while PPG-derived HRV and PRV indices add information about both arousal and valence through autonomic balance. Their non-invasive, wearable nature enables continuous monitoring in naturalistic environments, making them highly practical for emotion detection in real-world applications such as stress monitoring, adaptive human–computer interaction, and affective healthcare.

EEG measures electrical activity in the brain with high temporal resolution (on the order of milliseconds), making it especially suitable for capturing the fast dynamics of emotional processing [192]. Emotion-related EEG features are typically extracted from both the temporal characteristics of neural oscillations and their spatial distribution across cortical regions. Frequency-domain analysis focuses on the power of different bands: increased *alpha* (8–12 Hz) suppression in frontal regions is associated with higher arousal, while *frontal alpha asymmetry* (greater left vs. right activity) is a well-established marker of positive versus negative valence [221], [222]. Elevated *beta* (13–30 Hz) and *gamma* (>30 Hz) power often correspond to heightened arousal, anxiety, or stress, whereas increased *theta* (4–7 Hz) activity has been linked to emotional memory and regulation [223].

Spatially, EEG allows localisation of emotion-sensitive activity: the **prefrontal cortex** is strongly implicated in valence processing, with left-hemispheric dominance for approach-related positive emotions and right-hemispheric dominance for withdrawal-related negative emotions [224]. The **parietal cortex** contributes to attentional and arousal-related modulation, while **temporal regions** are associated with affective auditory and visual processing [222], [225], [226]. These spatial–temporal markers enable EEG to disentangle dimensions of valence and arousal in real time, complementing slower autonomic measures such as HRV or GSR.

Moreover, event-related potentials (ERPs) derived from EEG provide temporally precise indices of emotion processing. Components such as the **late positive potential (LPP)** are enhanced for emotionally salient stimuli regardless of valence, reflecting sustained attentional engagement [227]. Other components, such as the **N170**, are sensitive to emotional facial expressions, while early components (e.g., **P1, N1**) can reflect rapid automatic differentiation of emotional versus neutral stimuli. These ERP signatures highlight the fine-grained temporal unfolding of emotional responses, from early perception to sustained evaluation.

EEG's millisecond-scale temporal resolution and sensitivity to both frequency-based and spatially distributed markers make it a uniquely powerful tool for studying the neural underpinnings of emotion. When combined with slower but robust autonomic measures such as GSR, HRV, and pupil size, EEG provides a complementary perspective that captures both the rapid and sustained aspects of emotional processing.

Challenges such as intersubject variability and noise in physiological data can affect detection accuracy, which researchers address using advanced techniques like multimodal fusion and domain adaptation algorithms [189]. Combining physiological signals with self-reported ratings or continuous annotations, as seen in datasets like CASE [228], allows for a more comprehensive understanding of emotions by integrating objective measures with subjective experiences.

Researchers increasingly rely on ML approaches to effectively interpret these complex physiological signals in emotion detection [192], [229]. These methods, some of which will be discussed in the next section, enable the extraction of meaningful patterns from high-dimensional physiological data, improving classification accuracy and robustness across individuals [230].

2.4 Machine Learning Approaches in Emotion Detection

Machine learning for emotion detection has become a significant and key area of research, with notable advancements that leverage various data modalities, including

text, physiological signals, and facial expressions, to classify emotions accurately [217]. The flow of emotion detection using ML consists of data collection from physiological signals, data pre-processing to remove the noise or distortions from the signals, feature extraction relevant to emotion detection from physiological signals, and model training and evaluation (see Figure 2.3). The principal methodologies can be broadly categorised into traditional ML and deep learning approaches.

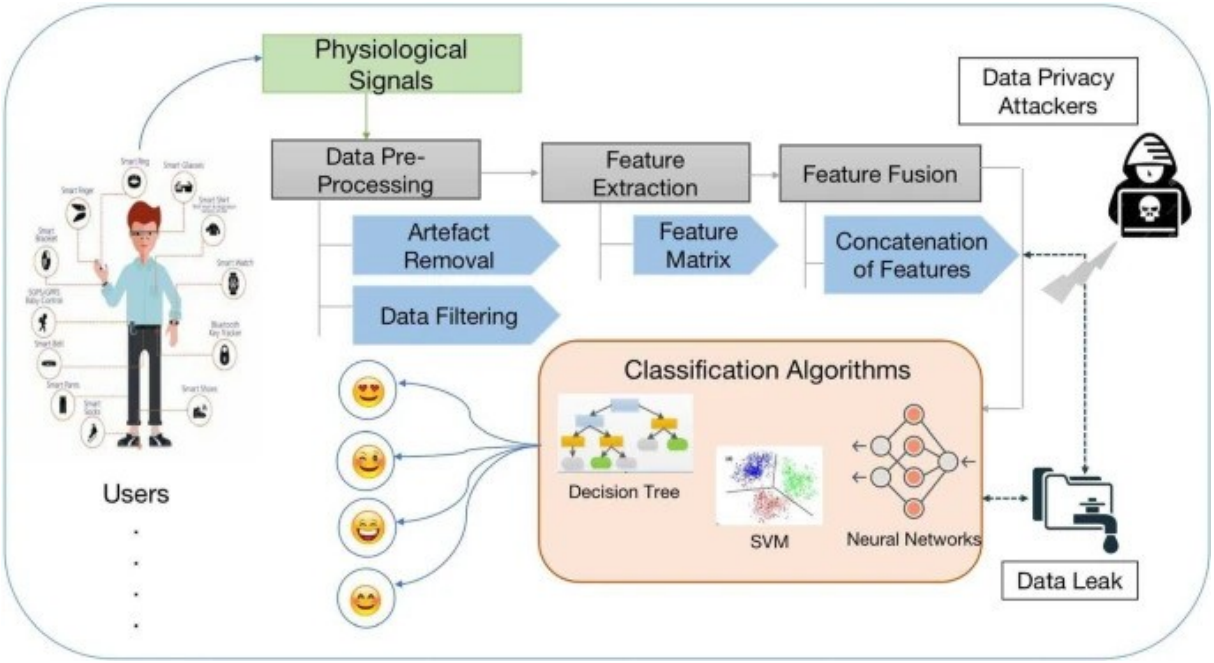


Figure 2.3: Flow of machine learning techniques used in the emotion detection model [231].

Traditional ML Approaches. Traditional methods rely on hand-crafted features extracted from physiological signals, such as time-domain, frequency-domain, and non-linear features. These features are optimised and fed into classifiers like Support Vector Machine (SVM) [232], K-Nearest Neighbors (KNN) [233], Decision Trees (DT) [234], Extreme Gradient Boosting (XGBoost) [235] and Random Forests (RF) [236]. For example, studies using the DEAP dataset [237] have demonstrated the effectiveness of classifiers, such as SVM and boosting algorithms, in recognising emotions from single and multimodal signals, including EEG, GSR, and PPG. Single-modal approaches analyse one signal type at a time (e.g., GSR for arousal), while multimodal approaches combine multiple signals to improve accuracy by capturing complementary emotional information [192], [217].

Deep Learning Approaches. Deep learning eliminates the need for manual feature extraction by automatically learning features from raw data. Techniques such as CNNs [238], Long Short-Term Memory (LSTM) networks [239], and hybrid models [240] are widely

used. Multimodal systems often employ deep learning to fuse feature or decision-level signals. For instance, attention-based LSTM models have been used to integrate EEG, GSR, respiration (RSP), and PPG data for improved classification accuracy [189], [241]. Transfer learning techniques, such as Joint Probability Domain Adaptation (JPDA), have also been proposed to address challenges like intersubject variability and noise in physiological signals [189].

Multimodal Emotion Detection Systems. Combining multiple physiological signals enhances emotion detection by leveraging the strengths of each modality. For example, integrating PPG-based HR with GSR has shown improved accuracy in mental stress prediction compared to single-modal systems [192], [217], [242]. Studies using wearable devices like smartwatches have demonstrated the feasibility of real-time emotion detection in everyday life by combining peripheral signals such as HRV and SC [243]

This research focuses on multimodal emotion detection using physiological signals, offering significant advantages over single-modal approaches. Multimodal systems integrate data from multiple channels, such as EEG, GSR, ECG, and PPG, to provide a more comprehensive and accurate understanding of emotional states [35], [36], [189], [191], [192], [217], [242], [244], [245]. Single-modality methods often suffer from limitations like susceptibility to noise and incomplete emotional information. For instance, EEG alone may capture brain activity patterns but might miss peripheral physiological cues like HRV (ECG) or SC (GSR). By combining modalities, multimodal systems leverage complementary information, improving robustness and accuracy in emotion detection.

2.4.1 Multimodal Approaches to Emotion Detection

Research has consistently shown that multimodal emotion detection outperforms single-modal approaches. Studies demonstrate that multimodal systems achieve higher accuracy by integrating diverse signals to extract informative features while mitigating the weaknesses of individual modalities [246], [247]. Data fusion techniques, such as feature-level and decision-level fusion (see Figure 2.4), further enhance performance by synchronising emotional cues effectively, enabling the detection of subtle or complex emotions [248]–[250].

The multimodal approach is particularly valuable in real-world applications such as healthcare, HCI, and mental health monitoring, as it enables better diagnosis and treatment by providing deeper insights into emotional states through the integration of brain activity with peripheral signals [246]. A key advantage of multimodal systems is their resilience to environmental noise and variability across subjects. Different physiological signals (such as EEG, GSR, and facial expressions) capture distinct aspects of emotional

responses; combining them reduces errors caused by noise in any single modality. For example, EEG may be sensitive to artifacts from eye movements or muscle tension, while GSR can be influenced by environmental factors such as temperature or humidity. By fusing these signals, the system can offset weaknesses in one modality using reliable information from another, leading to more accurate and robust emotion detection [251]. Additionally, multimodal systems account for inter-subject variability, supporting more generalised and adaptable performance across diverse populations and settings [107].

By adopting this multimodal framework in this study, we aim to develop a robust emotion detection system that leverages the strengths of various physiological signals to achieve superior performance and real-world applicability.

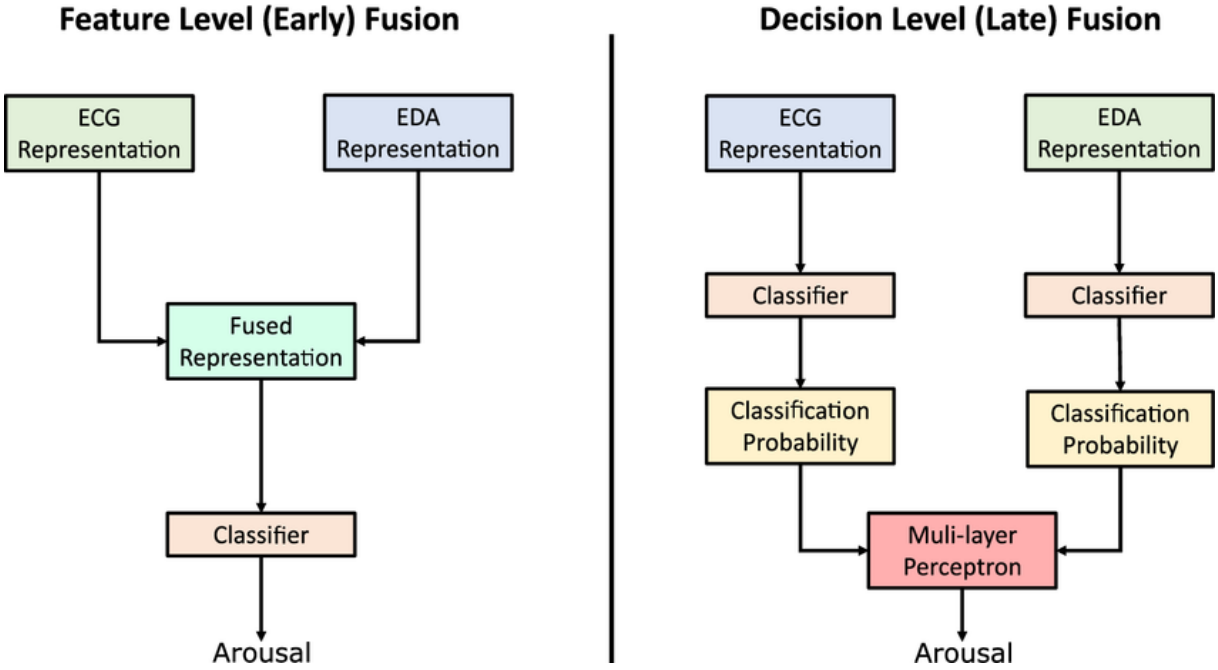


Figure 2.4: Comparison of Multiple Fusion Techniques.

Zhang et al. [252] introduced the CorrFeat algorithm, which extracts correlation-based features between skin conductance (SC) and pupil diameter (PD). Using the MAHNOB-HCI dataset, they performed both binary (high vs. low) and ternary (low, neutral, high) emotion classification for arousal and valence. CorrFeat achieved up to 82.9% (arousal) and 82.1% (valence) in three-class classification, clearly outperforming single-modality baselines. These results demonstrate the complementary nature of SC and PD in capturing autonomic emotional responses, while remaining non-intrusive and wearable-friendly.

Soleymani et al.[250] combined EEG and eye-tracking (including gaze and pupillary responses) for binary arousal-valence classification. While feature-level fusion did not outperform individual modalities, decision-level fusion significantly improved performance, achieving 76.4% for arousal and 68.5% for valence. This highlights that multimodal systems are not only more accurate but also more flexible, as decision-

level fusion provides resilience against missing or noisy data streams. Similarly, Zheng et al.[253] evaluated EEG and eye-tracking fusion. In unimodal settings, EEG achieved 71.77%, while eye-tracking achieved 58.90%. Fusion improved classification to 73.59% (feature-level) and 72.98% (decision-level), confirming that combining neural and ocular signals enhances emotion recognition beyond single-channel performance. More recently, Iacono and Khan [254] further demonstrated this benefit, reporting a mean accuracy of 0.935 ± 0.038 using EEG and eye-tracking fusion for binary arousal–valence detection.

Other two-modal systems have also reported high performance. Bulagang et al. [255] showed that combining ECG and electrodermography (EDG) achieved 95.7% accuracy in binary classification, outperforming unimodal setups and reinforcing the effectiveness of autonomic signal fusion.

The advantages become even more pronounced in three-modal systems. Kumar et al.[256] integrated EEG, ECG, and GSR, comparing multiple machine learning classifiers. Their multimodal approach outperformed unimodal and two-modal baselines, confirming that capturing distinct neural and autonomic features improves classification. Alam et al.[257] achieved near-perfect accuracy with the same three modalities using Random Forest and SVM to predict seven emotional states, showing the strong complementarity between these signals. Ramadan et al. [247] reported one of the highest performances to date by fusing EEG, EMG, and EOG, reaching 95.7% for arousal and 96.41% for valence (two classes). Including muscle activity and eye movements alongside EEG enabled a more holistic representation of emotional states, significantly boosting recognition accuracy, which is clearly higher than the accuracy achieved by Iacono and Khan [254] for two-class prediction using two modalities, which indicates a clear progression emerges across studies: Single-modality systems typically achieve 60–72% accuracy. Two-modal systems improve this to 73–93%, depending on the fusion method and signals used. Three-modal systems consistently reach 95%+, with some reporting near-perfect accuracy.

This trend underscores that multimodal systems provide a more comprehensive view of emotional states by integrating diverse physiological responses. By capturing both neural activity and autonomic dynamics, they mitigate the weaknesses of individual signals, reduce susceptibility to noise, and offer robustness essential for real-world applications.

Despite these advantages, multimodal emotion detection faces several challenges. Inter-subject variability in physiological responses hinders generalisation across users [36], [258]. Real-time processing of multiple signals demands high computational resources, limiting mobile and wearable deployment [252], [259]. Data synchronisation across heterogeneous sensors is another critical challenge, as misalignment can degrade classification accuracy [107], [253]. Furthermore, motion artifacts, lighting variability, and

environmental conditions reduce system reliability outside controlled lab settings [258], [260]. Most studies also focus on basic emotions, neglecting more complex or mixed affective states [254], [261]. Finally, ethical considerations such as privacy, consent, and data security remain a major concern in using physiological signals for emotion recognition [36], [259].

Building on these insights, our study integrates facial expression recognition (FER), galvanic skin response (GSR), and pupil size in a multimodal system. Data were collected from 47 healthy participants using the iMotions platform, which ensured precise synchronisation between stimulus presentation and physiological recording. To address lighting variability, we implemented a pupil-size preprocessing pipeline with cross-participant calibration. We applied feature-level fusion and trained models using XGBoost, enabling robust multimodal emotion regression beyond basic categorical classification. By combining behavioural (FER) and physiological (GSR, pupil) signals, our system provides a more reliable and ecologically valid method for detecting emotional states, addressing long-standing challenges of variability, temporal alignment, and practical feasibility.

2.4.2 Data Pre-processing Techniques

Data pre-processing plays a crucial role in enhancing the quality of physiological signals, ensuring that they can be effectively utilised for emotion detection in affective computing systems. This process involves several key stages: noise removal, distortion correction, normalisation, and segmentation. Different physiological signals, such as pupil size, GSR, and PPG, have distinct preprocessing challenges, and researchers have proposed various methods to address these issues, ensuring more accurate and reliable data for analysis.

FER Data Pre-processing. FER has long been employed in affective computing, traditionally focusing on the six or seven "basic emotions" identified by Ekman and colleagues. Modern FER systems, such as AFFDEX, OpenFace, and Affectiva, typically rely on convolutional neural networks or action unit (AU) activation patterns to detect categorical emotional states on a frame-by-frame basis. These categorical predictions are widely used in human–computer interaction, driver monitoring, and behavioural analytics.

However, there is growing recognition that categorical FER has important limitations. First, facial expressions often encode mixtures of emotions, making discrete labels insufficient for capturing nuanced affective states. Second, categorical outputs do not directly map onto dimensional frameworks of affect, such as Russell's circumplex model, which describe emotions along valence (pleasantness-unpleasantness) and arousal (activation-

deactivation) axes. To address this, several studies have attempted to link facial expressions to dimensional measures. For instance, Cowen and Keltner [262] mapped human-annotated facial expressions into valence-arousal space, showing that expressions cluster in line with circumplex predictions. The AffectNet dataset (Mollahosseini et al., [263]) provides both discrete emotion categories and continuous valence/arousal annotations, enabling models to learn direct regressions from facial images to dimensional affect. Similarly, the Emotional Trace approach (Ayoub et al., [264]) proposed mapping basic expressions into continuous coordinates to visualise trajectories of affective change.

While these studies demonstrate the feasibility of bridging FER with dimensional affect models, most rely either on normative ratings (assigning each discrete emotion a fixed position in the circumplex) or on supervised regression trained on datasets annotated with both labels. In practice, categorical FER outputs are still frequently treated in isolation, limiting their usefulness in multimodal pipelines where dimensional ground truth is required.

Pupil Size Data Pre-processing. Pupil size measurements are influenced by both ambient lighting and emotional arousal [206], [207], [209], [265], introducing an additional layer of complexity when using pupillometry for emotion detection [266]. Pupils typically constrict in bright conditions and dilate in darker environments [267], with size varying by up to 50% [268], complicating the interpretation of emotional responses [207], [269]. This is also the case for our increasingly digital world, where the brightness and contrast of device displays can considerably affect pupil size [270]. Therefore, it is crucial to effectively isolate the changes in pupil size resulting from authentic emotional responses from those influenced by ambient light conditions [271], and literature suggests that subtracting luminosity-induced pupil size changes can effectively isolate the emotional responses [41].

Previous research suggests that accounting for luminosity-induced pupil size changes can help to isolate the emotional responses with varying levels of success [41]. Earlier research used grey screens to measure baseline pupil diameter [205], [272], [273]. Still, this method has proven insufficient for complex stimuli such as videos with dynamic visual elements [207], where changing luminosity and visual elements continually affect pupil responses. In the case of a video, an ideal baseline would be a set of emotionally neutral stimuli of the same luminosity as each video frame, which is challenging to engineer.

In our thesis, we propose a more realistic and scalable approach: we directly predict the component of pupil size influenced by luminosity for each emotional frame, rather than including an emotionally neutral frame with the same luminosity repeated throughout the entire video. This is not the first attempt to use such an approach;

previous studies have attempted to disentangle the effects of emotional arousal and light exposure on pupil size for dynamic stimuli, such as videos. Our research builds upon previous attempts to employ this approach, while addressing some of its shortcomings. For example, Nakayama et al. [274] developed a hyperbolic model to predict pupil size as a function of the luminosity of the computer screen in a dark laboratory. Other researchers have used simpler models. For example, Tarnowski et al. [275] hypothesised a linear relationship between luminosity and pupil size. However, the relationship between pupil size and luminosity is exponentially decreasing [265]. Previous work, including our pupil–luminosity modelling study, demonstrated the potential of exponential models to separate pupil size changes due to luminosity from those due to emotional arousal [109]. The detailed methodology and experimental validation of this approach are described in Chapter 3 in the pupil size analysis section 3.2.6. Nakayama and colleagues [274] also sought to separate pupil dilation related to emotional arousal from luminosity-induced changes in video content. They also applied their method to emotional images and demonstrated a luminosity and arousal effect on pupil size using analysis of variance (ANOVA). We wanted to go one step further: predicting self-reported arousal using pupil size, which was not addressed in Nakayama and colleagues' work.

Raiturkar et al. [276] developed a linear model to predict pupil size based on light intensity, allowing for the isolation of emotional arousal by subtracting the effects of light from actual pupil measurements, and applied it to the analysis of emotional videos. However, they did not record self-reported arousal while watching the movie clips.

Where possible, we have applied the existing models mentioned above to our data, with the advantage that we included a comparatively large sample of 47 participants (compared to a maximum of 10 in the studies above), a relatively rich video collection of 32 emotional video clips, and the fact that we had recorded self-reported arousal for each participant and each video.

In some cases, it was not possible to apply methods from previous studies to our data due to insufficient details regarding the corresponding models. For example, Asano et al. [277], [278] developed models that considered the temporal evolution of pupil size, which we did not do. Their first model consisted of a linear model, and the second of a neural network. We were only able to evaluate the first model, but we were unable to assess the neural network model since we did not have access to its trained neural network. Asano and colleagues [277] wrote that the second method performed better than the first one, and they tested both methods with emotional video clips but did not detect self-reported arousal. A more detailed analysis of the differences between our method and those of the other researchers is reported in the Discussion section 5.

Our research directly addresses these challenges by developing advanced methods for isolating emotional arousal-related pupil dilation from the confounding effects of

ambient light. By implementing more sophisticated baseline correction techniques and accounting for individual variability in pupil reactivity, we aim to improve the reliability of pupil size measurements as a marker of emotional responses, thereby enhancing the accuracy of emotion detection systems.

GSR and PPG Data Pre-Processing. GSR and PPG signals are crucial for emotion detection, reflecting physiological responses to emotional arousal. However, raw signals from both modalities are often contaminated by noise, motion noise, and baseline drift, which can undermine their usefulness in emotion analysis. A significant challenge in preprocessing these signals is removing noise and distortions introduced by body movements, muscle contractions, and environmental factors [279].

For GSR, common approaches include filtering techniques such as low-pass, high-pass, and bandpass filters to remove unwanted frequency components [280], [281]. Wavelet transforms have also been explored to decompose GSR signals into different frequency bands, aiding noise suppression while retaining critical data related to emotional arousal [282]. Similarly, PPG preprocessing employs low-pass, high-pass, and bandpass filters to eliminate noise and baseline drift, with bandpass filters typically used to retain the primary frequency components of the PPG signal (0.5–5 Hz) [72], [283], [284].

Normalisation techniques, such as Z-score normalisation and Min-Max scaling, are commonly applied to standardise both GSR and PPG data, making it easier to compare responses across different individuals [285], [286]. Noise removal methods, such as Independent Component Analysis (ICA) and adaptive filtering, are particularly useful for separating noise from meaningful signals and reducing the impact of motion artifacts [287], [288]. Advanced signal decomposition methods like Wavelet Transform (WT) and Empirical Mode Decomposition (EMD) further enhance signal clarity by breaking down the signals into Intrinsic Mode Functions (IMFs)[289]. More recently, deep learning approaches, such as CNNs and LSTM networks, have shown promise in denoising both GSR and PPG signals. However, they require large datasets for effective training[290].

Furthermore, fixed windowing and event-related segmentation techniques are crucial for organising GSR and PPG data into meaningful intervals, aligning the signals with emotional responses triggered by specific stimuli [213]. These preprocessing techniques collectively enhance the reliability and accuracy of GSR and PPG signals for emotion detection studies.

In conclusion, data preprocessing is critical in enhancing the quality and reliability of physiological signals, including pupil size, GSR, and PPG, for emotion detection. Although various methods, such as filtering, noise removal, and signal normalisation, have been widely explored, challenges remain in dealing with the complex nature of

real-world data, including dynamic visual stimuli, motion noise, and individual differences.

Following preprocessing, the next essential step is feature extraction, where meaningful patterns are derived from the cleaned physiological signals.

2.4.3 Feature Extraction and Selection Techniques

Emotion detection models rely heavily on the quality of the features extracted from physiological signals. No model can perform optimally if the features provided are inaccurate or unreliable; hence, feature extraction and selection become crucial for training any ML model. Feature extraction involves identifying and isolating the most relevant characteristics of physiological signals that correlate with emotional states. In contrast, feature selection focuses on reducing dimensionality by choosing only the most informative features to improve model performance.

Various techniques are employed to extract meaningful features of physiological signals such as EEG, GSR, and PPG-based HR. These include time-domain analysis (e.g., mean, standard deviation, root mean square), frequency-domain analysis (e.g., power spectral density, Fourier transforms), and non-linear dynamics measures (e.g., entropy, fractal dimensions). For example, EEG signals are often analysed for power in specific frequency bands (alpha, beta, theta) to detect emotional arousal or valence. Similarly, GSR and PPG-based HR signals are processed to extract features like SC peaks or HRV, which reflect emotional arousal.

As mentioned in the "Overview of Emotion Detection Using Physiological Signals" section 2.3, pupil size data are typically collected using an eye-tracking device that records ocular metrics, such as gaze position, eye movements, fixation patterns, and pupil dilation. These measurements provide rich information that can be used to extract features relevant to emotion detection, as has been demonstrated in numerous studies. One of the most commonly used features is pupil dilation, which is strongly associated with emotional arousal. Both positive and negative emotions can induce significant changes in pupil size, with negative stimuli often leading to more prolonged dilation [197], [203], [252], [254], [275], [291]–[293]. Another essential feature is fixation duration, which reflects attentional engagement and varies depending on the emotional intensity or valence of the stimulus [197], [275], [291], [293], [294]. In addition, pupil area and position provide spatial and dimensional insights into pupil dynamics during emotional exposure [197], [291]. The researchers have also extracted statistical features from the pupil responses - including minimum, maximum, mean, median, standard deviation, variance and quartile values - to comprehensively describe the distribution and variability of the pupil signal [197], [292], [293]. In addition, saccadic movements and their speed have been used to analyse how the eyes

move across visual scenes, providing cues to cognitive and emotional processing [197], [275], [291], [293]. Finally, dynamic changes in pupil size over time serve as a valuable metric for tracking emotional responses in real time, highlighting the temporal aspect of emotional engagement [295], [296]. These features contribute to a nuanced and multi-dimensional understanding of how pupil behaviour reflects underlying emotional states.

However, our research focuses on pupil size as a key physiological marker of emotional arousal.

Pupil Size as a Stand-alone Feature for Emotion Detection. Several studies have focused on pupil size for emotion detection, confirming its reliability as a physiological marker. In particular, pupil dilation has been shown to correlate strongly with emotional arousal, making it a key feature in emotion recognition systems.

A study by Aracena et al. [199] used the temporal evolution of pupil dilation and gaze coordinates during a fixed time window as input features. The researchers trained machine learning models, including neural networks and decision trees, and achieved a best accuracy of 74.5% for a single subject. However, accuracy dropped to 53.6% when data from multiple subjects were aggregated, highlighting the challenge of individual variability in pupil size responses. This study highlights the importance of gaze position and pupil dilation as they provide complementary information about visual attention during emotional stimulus presentation.

Pupil dilation has also been associated with emotional arousal in various contexts. For example, Oliva et al. [297] found that peak pupil dilation occurred during emotional arousal when participants were exposed to human nonverbal vocalisations, thus correlating pupil size with emotional valence perception.

Another study by Lee et al. [292] extracted 16 statistical features from pupil responses, including minimum, maximum, first quartile (q1), median (q2), third quartile (q3), mean, standard deviation and variance. These features were collected from 30 participants exposed to emotionally evocative content, and the model achieved a classification accuracy of 76% for emotions such as fear, anger and surprise using logistic regression. This demonstrated the value of statistical features in capturing nuanced emotional responses.

Arias et al. [296] further investigated pupil dilation in response to audiovisual emotional speech, showing that pupil responses indicated an emotional mismatch between the visual and auditory stimuli. This pupil dilation occurred as participants searched for emotional cues, particularly in the first fixation on mismatched areas. This suggests that pupil size may reflect attentional shifts in response to emotional stimuli.

In clinical applications, pupil dilation has been used to assess impairments in emotion recognition in stroke patients [298], further supporting its utility in healthy and

impaired populations. In addition, changes in pupil dilation have been shown to occur in response to happy, sad and surprised emotions [299], further emphasising its role in capturing emotional states across a range of affective experiences.

Together, these studies highlight the strong relationship between pupil dilation and emotional arousal, demonstrating its effectiveness as a stand-alone feature for emotion detection. Furthermore, combining pupil size data with other physiological and behavioural signals has improved emotion classification accuracy [291], [292].

Galvanic Skin Response. GSR has been widely studied in the context of emotion detection. GSR reflects changes in the skin's electrical conductance due to variations in sweat gland activity, which are influenced by ANS responses. As a result, it serves as a valuable physiological indicator of emotional arousal and has become a key component in affective computing research. GSR signals consist of two primary components: the tonic component (slow-varying SCL) and the phasic component (rapid fluctuations known as skin conductance responses (SCR) [300]. Properly decomposing these components is essential for effective feature extraction and emotion classification.

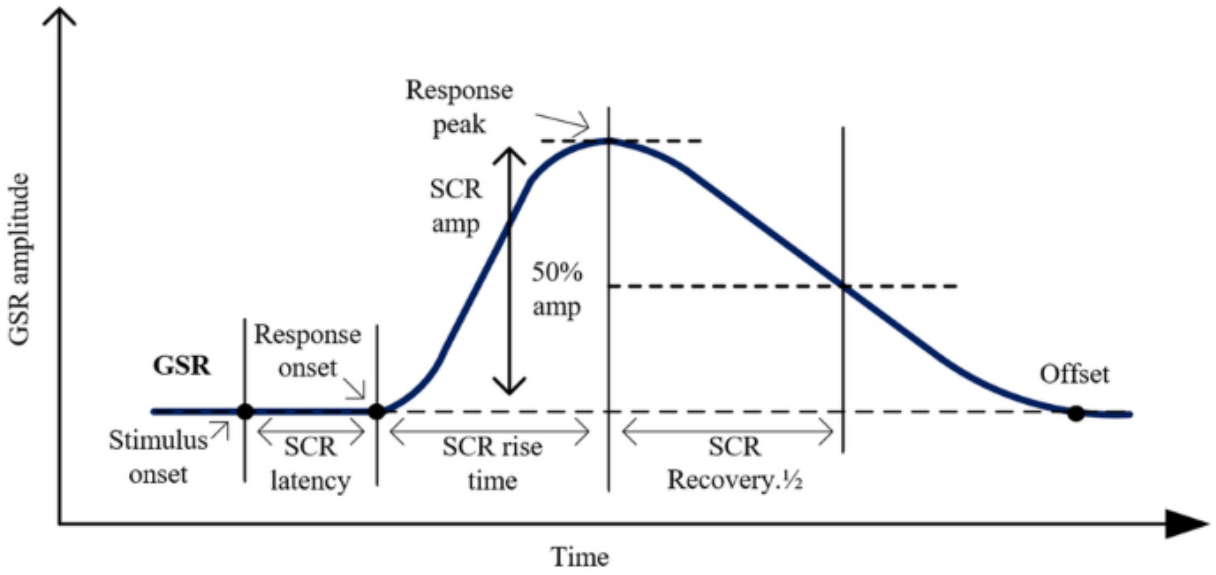


Figure 2.5: GSR Features [167].

Several methods have been employed for feature extraction from GSR signals to support emotion detection. Time-domain features are the most widely used and provide statistical insights into the overall shape and variability of the signal. These include metrics such as the mean, standard deviation, skewness, kurtosis, root mean square (RMS), and higher-order derivatives [301]–[305]. These features help capture fluctuations in skin conductance associated with emotional arousal. The GSR signal is also decomposed into phasic and tonic components to isolate fast, stimulus-driven responses from the underlying baseline activity [301], [305], [306]. Techniques such as high-pass filtering (e.g., with a cutoff frequency of 0.05 Hz)[307] and detrending methods based

on regularised least-squares have been used to achieve this separation[308]. More recently, methods like cvxEDA have gained popularity for decomposing the signal into a smooth tonic component and a sparse phasic component [309], with studies recommending a downsampling rate of 20 Hz for efficient processing [310].

In the frequency domain, features such as Power Spectral Density (PSD)[305] and the Fractional Fourier Transform (FrFT)[311] are used to assess how signal power is distributed across frequencies, capturing oscillatory characteristics linked to emotional states.

Time-frequency methods such as the Short-Time Fourier Transform (STFT) and Wavelet Transform offer multi-resolution analyses that track how frequency content evolves, making them suitable for analysing non-stationary signals like GSR [301], [305]. Decomposition-based approaches, notably EMD, break the GSR signal into IMFs, allowing for adaptive analysis of its complex, non-linear structure [301].

Lastly, event-based features focus on specific responses such as SCR peaks (see Figure 2.5), including their amplitude, latency, rise time, and frequency of occurrence. These measures are crucial for identifying emotionally salient events [192], [252], [305], [306]. Together, these features represent GSR dynamics comprehensively, enabling emotion recognition models to capture spontaneous fluctuations and stimulus-specific reactions across various timescales.

Recent studies have further expanded our understanding of GSR-based emotion detection: A study revealed distinct physiological signatures for different emotional states [312]. Fear responses showed the most extended duration, family bonding emotions displayed slower responses, and humour triggered quick but temporary reactions [312]. Combining SC measurements with other physiological indicators, such as HR, EMG, and brain activity, has enhanced the accuracy of emotion detection systems [312]. As research advances, integrating GSR-based emotion detection with other physiological signals and AI technologies promises to create more sophisticated and accurate emotion-sensitive systems.

Photoplethysmography. PPG signals are widely utilised in emotion detection, with extracted features categorised into time-domain, frequency-domain, geometric indices, and other statistical measures. These features provide insights into HRV, signal morphology, and complex physiological patterns associated with emotional states. However, throughout the research, we focused on PPG-based heart-related features.

Rakshit et al. [72] explored a novel approach to emotion detection using HRV features extracted from PPG signals. Their study demonstrated that PPG-derived features could effectively classify emotional states, often achieving comparable or slightly higher accuracy than ECG-based methods. The research compared PPG with ECG in detecting psychophysical and affective states and found that Pulse Rate Variability (PRV) features

extracted from PPG provided slightly better emotion prediction than traditional HRV features from ECG. The study further highlighted that PPG shape parameters could achieve accuracy levels similar to HRV and PRV features. This suggests that ANS responses to emotional stimuli influence beat duration and PPG waveform shape. ML models such as KNN and SVM were also applied to classify emotional states, demonstrating that PPG technology could be a viable alternative to ECG detection.

Goshvarpour et al. introduced a novel geometric approach by extracting ten features from Poincaré's section analysis of PPG-based interbeat interval (IBI) signals [313]. These features were derived from a 2D phase space reconstruction of PPG-IBI data and included geometric indices, basin geometry characteristics, angular-based features, and quantitative measures from phase-space sections. The geometric indices quantified the spread and distribution of points in the reconstructed phase space. At the same time, the basin geometry characteristics analysed changes in the structure of PPG-IBI phase states under different emotional conditions. Poincaré's sections formed at various angles were also used to capture trajectory variations, and structural properties were extracted to reflect non-linear heart dynamics during emotional responses. Unlike traditional time-domain and frequency-domain features, this approach captured PPG-IBI trajectories' non-linear and dynamic behaviour, making it a promising method for emotion detection. When these features were used in an SVM classifier, the study achieved high classification accuracy, with 96.67% for binary emotion classification and 91.11% for multi-class classification, reinforcing the potential of geometric feature extraction for improving emotion detection.

Additionally, other studies have explored complementary methods, such as entropy-based features [314] and spectral analysis of HRV [260]. However, we focus on integrating the feature extraction approaches established by Rakshit et al. and Goshvarpour et al. to develop a more robust emotion detection model.

Table 2.1 summarises the key PPG-based heart-related features utilised in our study and their corresponding references. By leveraging these well-established methodologies, we aim to refine feature selection and improve the classification accuracy of our emotion detection model.

Feature Selection

Feature selection is an essential step in ML model development, as it helps improve model interpretability, reduce overfitting, and enhance computational efficiency. Selecting the most relevant features can improve the model's predictive performance. Various feature selection techniques exist, ranging from traditional methods to more advanced approaches, each offering distinct advantages depending on the data and the model used.

Table 2.1: PPG-based heart-related Feature Extraction for emotion detection.

Feature Category	**Parameters**	**Description**	**Citations**
Time-Domain Features	Statistical Parameters	Mean, median, standard deviation, maximum, minimum, range	[260]
	HRV	meanRR, medianRR, SDNN, SDANN, pNN50, NN50, RMSSD, SDNNi, meanHR, stdHR	[72], [315], [316]
	Interbeat Interval (IBI)	Mean interbeat interval extracted from PPG	[313]
Frequency-Domain Features	Power Spectral Density (PSD)	Distribution of power across frequency components of PPG signals	[260]
	Spectral Analysis of HRV	VLF, LF, HF, peakVLF, peakLF, peakHF, aVLF, aLF, aHF, aTotal, pVLF, pLF, pHF, nLF, nHF, LFHF ratio	[72]
	HR Estimation	FFT-based frequency analysis to derive HR	[260]
Geometric Indices	Poincaré Plot Measures	Geometric patterns of IBI signal in phase space	[313]
	Basin Geometry	Structural changes in IBI phase states across emotions	[313]
Other Features	Normalised Signal Features	Mean of normalised signals (data_mean), median of 25-mean data (25_mean_median)	[314]
	Entropy Measures	Information entropy (data_entropy) to assess signal complexity	[314]

In general, feature selection methods can be grouped into three broad categories: filter, wrapper, and embedded. Filter methods assess the relevance of features independently of the model and include techniques such as correlation analysis, mutual information, and variance thresholds. These methods are often used for high-dimensional datasets, where feature redundancy or weak relationships can lead to poor model performance. For instance, the Pearson correlation is widely employed to measure the linear relationship between features and the dependent variable, identifying features with strong predictive power and discarding irrelevant ones [317]. This method is particularly useful for reducing dimensionality in regression models, as it helps identify multicollinearity, which can be addressed through techniques like variance inflation factor (VIF) analysis [318].

Wrapper methods, in contrast, evaluate feature subsets based on model performance, using techniques like forward selection, backwards elimination, and recursive feature elimination (RFE) [319]. These computationally intensive methods can yield highly accurate feature subsets for specific models. Embedded methods, such as LASSO (Least Absolute Shrinkage and Selection Operator), Ridge Regression, and Elastic Net, integrate feature selection directly into the model training process, offering both efficiency and accuracy in feature selection [320], [321].

Recent advancements in feature selection, especially for regression models, emphasise hybrid approaches, machine learning-based techniques, and domain-specific optimisations. Hybrid methods combine the strengths of filter, wrapper, and embedded approaches to improve the accuracy and efficiency of feature selection. For example, combining correlation-based filtering with RFE can lead to more refined and practical feature subsets [322]. Machine learning-based techniques, such as Random Forest feature importance [323], SHAP (Shapley Additive Explanations) values [324], and deep learning methods like autoencoders and neural network pruning [325] offer sophisticated ways to identify and prioritise the most significant features. Additionally, domain-specific feature selection methods tailored to genomics, finance, and healthcare fields help enhance model performance and relevance [326].

Several studies have compared the effectiveness of different feature selection techniques. Guyon and Elisseeff demonstrated the efficiency of SVM-based feature selection in regression tasks [327], while Tibshirani highlighted the sparsity-inducing properties of LASSO for high-dimensional data [320]. Kuhn and Johnson examined the trade-offs between RFE and embedded methods, particularly regarding interpretability versus computational cost [328]. Despite its simplicity, the Pearson correlation remains a strong baseline method for feature selection, especially for continuous numerical data [329], [330].

However, feature selection still has challenges, especially when dealing with high-dimensional datasets and ensuring scalability and robustness across different data dis-

tributions. Recent research has focused on automating feature selection processes that can adapt to varying data characteristics. Future work in this area may explore incorporating deep learning and reinforcement learning techniques and developing more interpretable models to further improve the applicability and performance of feature selection in real-world scenarios.

In the context of regression-based models, feature selection plays a critical role in determining the accuracy and efficiency of the model. Whether using traditional methods, advanced approaches or hybrid strategies, the goal remains to identify the most relevant features that improve predictive performance while reducing computational complexity. The choice of method depends on the dataset's characteristics and the specific application, with future developments focusing on automation, scalability and adaptability.

Feature selection becomes even more critical in multimodal emotion recognition systems that combine data from different physiological signals. These systems require careful feature extraction and selection to capture complementary information from different modalities, such as pupil size, GSR and PPG. Poorly selected features can introduce noise and reduce model accuracy, highlighting the importance of effective feature selection in improving model robustness and performance. By refining the features extracted from physiological data, ML models can more accurately detect and interpret emotional responses, improving regression performance and making these models more applicable to real-world emotion recognition systems.

2.4.4 Datasets and Experimental Protocols

Several benchmark datasets are commonly used to train and evaluate models in emotion detection, each with unique characteristics, challenges, and opportunities. For example, the DEAP dataset [237] includes multimodal signals such as EEG, GSR, PPG, and facial expressions, but it is limited by its relatively small sample size and limited emotional diversity as it mainly elicits common emotions within valence–arousal space and lacks coverage of subtle or complex states. Similarly, the DREAMER dataset [331] provides audiovisual stimuli with multimodal recordings (EEG, GSR, PPG, and video), though it also suffers from a small participant pool and limited coverage of emotional states.

The AMIGOS dataset [332] is another multimodal benchmark containing EEG, GSR, PPG, and video recordings. While it encompasses multiple emotional states, such as happy, sad, and neutral, its emotional diversity is limited, and the dataset may not fully capture subtle or mixed emotions. The EmoReact dataset [333] offers multimodal signals, including EEG, GSR, ECG, and video, but is restricted by its small sample size of 32 participants and a narrow range of emotions. In contrast, AffectNet [263] is a large-scale dataset with facial expression labels across millions of images. However, it

lacks physiological signals and is limited to basic emotional categories.

Other datasets also face similar trade-offs. For instance, SAVEE [334] includes audiovisual recordings of emotional speech and facial expressions but only from male speakers, limiting participant diversity. MELD [335] focuses on multimodal emotional dialogues from movies and TV shows, which, while naturalistic, introduce noise from uncontrolled conversational dynamics. Similarly, Affectiva [336] combines facial expression and physiological data but is restricted in terms of the breadth and depth of physiological signals, with a stronger emphasis on facial recognition.

Across these datasets, common limitations include small sample sizes, limited diversity of emotional stimuli, lack of subtle or complex emotional states, restricted participant demographics, and the absence of psychiatric screening to ensure emotionally healthy participants. To overcome these issues, we designed and conducted our own data collection, using a participant pool of 47 individuals. This enabled us to ensure a broader and more representative set of emotional stimuli, as well as have an experimental protocol that directly addressed our initial research questions. However, one limitation of our dataset is that we did not apply pre-screening to exclude participants with potential emotional or psychological conditions, meaning that individual variability in emotional health could influence responses. Despite this, our dataset addresses many of the constraints of existing benchmarks by improving emotional diversity, increasing the participant base, and combining multimodal data sources in a controlled setting, thereby enhancing both quality and relevance for emotion detection research.

2.4.5 Evaluation Metrics and Performance Analysis

Evaluation metrics play a crucial role in emotion detection systems by objectively measuring how well a model can interpret human emotions' complex and context-dependent nature. These metrics help ensure that models are reliable in real-world applications by connecting subjective emotional experiences with measurable AI performance. Since emotion detection is a continuous process, the selected metrics must effectively reflect alignment with the actual valence (positivity) and arousal (intensity) scores. Since our study uses a regression-based approach for emotion detection, we concentrate on evaluation metrics that are appropriate for regression models, such as $R2$, $NRMSE$, Pearson's r, and Concordance correlation coefficient CCC. These are summarised in Table 2.2.

These metrics allow for an accurate quantification of the relationship between predicted and actual emotional states over time.

Key Findings in Performance Analysis. Preprocessing techniques, such as normalisation and filtering, are essential to improve the quality of physiological signals, such as

Metric	Formula	Purpose	Example Performance
$R2$ (R-squared)	$R2 = 1 - \frac{SS_{res}}{SS_{tot}}$	Measures variance explained by the model. Higher values indicate a better fit.	0.892 (arousal), 0.759 (valence) [337]
$NRMSE$	$\frac{RMSE}{y_{max} - y_{min}}$	Normalised RMSE for scale-independent error comparison.	0.035 (arousal), 0.078 (valence) [337]
Pearson's r	$r = \frac{\text{cov}(y,\hat{y})}{\sigma_y \sigma_{\hat{y}}}$	Measures linear correlation between predicted and true scores.	0.88 for valence in EMER (Multi) dataset [338]

Table 2.2: Evaluation metrics, formulas, purpose, and example performance results.

ECG and EDA, increase model robustness, and reduce noise. These preprocessing steps have improved $NRMSE$ values, indicative of model performance [339].

When comparing models, RF significantly outperforms linear models in predicting emotional valence. For example, in valence prediction tasks, RF achieves a higher $R2$ value (0.7) compared to linear models (0.27), highlighting the superior predictive power of RF models [337].

Finally, using $NRMSE$ allows for more reliable model performance comparisons across different datasets. As datasets can have different scales, $NRMSE$ standardises the evaluation, making it easier to assess and compare the effectiveness of different models in different contexts [337].

2.4.6 Impairment of Emotion Recognition in Mental Health

Emotion recognition is essential for effective social interaction, but people with a range of mental disorders face significant challenges in accurately interpreting emotional cues. Emotion recognition impairment (ERI) is seen in a variety of psychiatric and neurological conditions, including depression, schizophrenia, autism spectrum disorder (ASD), acquired alexithymia, traumatic brain injury (TBI) and anxiety disorders.

People with depression tend to misinterpret facial expressions, particularly negative emotions, which can increase feelings of sadness or hopelessness [340], [341]. This impairment has been linked to dysfunction in brain regions such as the amygdala [342]. Emotion recognition deficits are particularly common in schizophrenia, where people struggle to interpret both facial expressions and vocal emotions. These difficulties contribute to social dysfunction and communication problems [18]. People with autism spectrum disorder often find it difficult to understand non-verbal emotional cues, such

as facial expressions, which can hinder their social interactions [343]. People with anxiety disorders may misinterpret neutral or ambiguous facial expressions as threatening, leading to increased stress and avoidance behaviour [344]. Acquired alexithymia, which often results from neurological conditions such as brain injury or trauma, affects the ability to recognise and process emotional experiences. People with acquired alexithymia may have difficulty identifying their emotions and interpreting the emotional expressions of others, which can negatively affect their relationships and psychological well-being [345]. TBI can lead to changes in emotional processing, often affecting the ability to recognise facial expressions. TBI patients may exhibit poor social interaction due to difficulties in emotional recognition, which is particularly important in rehabilitation and social reintegration [346].

Psychological assessment using standardised scales is essential to effectively assess and identify these impairments in clinical settings. For example, scales such as the TAS-20 (Toronto Alexithymia Scale) are used to assess alexithymia [347], while the PHQ-9 (Patient Health Questionnaire) is commonly used to assess the severity of depression [348]. The GAD-7 (Generalised Anxiety Disorder-7) scale helps to measure anxiety levels [349], and the IPDE (International Personality Disorder Examination) is used to diagnose PDs [350]. These scales provide a quantitative measure of psychological conditions, crucial for screening individuals in clinical trials or treatment protocols.

These standardised scales allow clinicians to categorise and assess different mental health conditions, contributing to more accurate diagnosis and treatment. For example, people diagnosed with depression or anxiety disorders often show impairments in emotion recognition that distort their responses to emotional cues [18], [351]. Using these scales with advanced emotion recognition technologies makes it possible to better understand how emotional impairments manifest and tailor interventions accordingly.

2.4.7 The Role of ML in Advancing an Emotion Detection

ML holds great promise for addressing impairments in emotion recognition. ML models can provide accurate, real-time assessments of emotion by analysing multimodal data such as facial expressions, vocal tone and physiological signals (e.g. HR, SC). These systems can significantly benefit clinical settings in the following ways

1. Objective Measurement: ML models provide objective emotion detection, reducing reliance on subjective self-report, which can be inconsistent or biased. This improves the accuracy of emotional assessments, particularly in individuals with mental health disorders [352].
2. Personalised Treatment: ML can help tailor treatment plans based on an individual's unique emotional responses. Emotion recognition systems can monitor responses during therapy or medication trials, allowing clinicians to adjust inter-

ventions as needed, improving treatment outcomes [353].

3. Early Detection and Intervention: ML can detect subtle emotional changes that may signal the worsening of conditions such as depression, schizophrenia or TBI. This early detection allows for timely intervention, potentially preventing further deterioration in emotional health [354].
4. Support for Individuals with ASD: ML-powered emotion recognition systems can help people with autism spectrum disorders recognise and interpret emotions, improving their social interactions and emotional empathy [355].
5. Support for Acquired Alexithymia and TBI: ML-based emotion recognition systems can help identify and interpret emotional cues for individuals with acquired alexithymia and TBI. These tools can help individuals become more aware of their emotional states and improve their empathy, facilitating better social and therapeutic interactions [356].
6. Long-Term Monitoring: ML-based emotion recognition systems can be integrated into remote monitoring tools to provide ongoing emotional support for individuals with limited access to face-to-face care. This particularly benefits patients with chronic conditions or those living in remote areas, providing an accessible and consistent monitoring mechanism for emotional well-being [357].

Impairments in emotion recognition are prevalent in a wide range of mental health and neurological conditions, including depression, schizophrenia, ASD, anxiety disorders, acquired alexithymia and TBI. ML-based emotion recognition systems provide an effective and objective means of assessing and tracking emotional states. These systems enhance clinical diagnosis and treatment and support personalised care, early intervention and long-term monitoring. By integrating these technologies into clinical practice, the accuracy of emotion recognition can be significantly improved, ultimately contributing to better mental health outcomes and improved emotional well-being.

2.5 Systematic Literature Review: Multimodal Continuous Emotion Prediction Using Physiological and Visual Signals

2.5.1 Introduction

Emotion recognition technologies are increasingly critical in affective computing, mental health monitoring, and human-computer interaction. The ability to detect and interpret human emotions has broad applications, from enhancing user experience in human-computer interfaces to supporting clinical diagnosis and mental health interventions [23]–[25]. Traditionally, emotion recognition approaches have focused on

categorical models classifying emotions such as happiness, sadness, or anger. However, these models often fail to capture the continuous and dynamic nature of human emotions. The valence-arousal model, also known as Russell's circumplex model of emotion, provides a dimensional approach that represents emotions along two continuous axes: valence (the degree of pleasantness) and arousal (the intensity of emotion) [358]. Continuous emotion prediction using this model enables a more refined understanding of affective states and their temporal variations. Recent advancements in machine learning, particularly regression-based approaches, have facilitated continuous emotion prediction using diverse input modalities. Among these, FER, GSR, and pupillometry have emerged as prominent physiological and visual signals indicative of affective states [238], [258], [301]. The integration of these modalities into multimodal models holds the potential to improve prediction accuracy by capturing complementary aspects of emotional responses. Despite these advances, effective fusion of physiological and visual signals remains a significant challenge. Various techniques, including feature-level fusion, decision-level fusion, and hybrid approaches, have been proposed to integrate data from different modalities. However, a systematic evaluation of these fusion strategies in the context of continuous emotion prediction is lacking. This systematic literature review aims to identify and assess the most effective techniques for fusing FER, GSR, and pupillometry to enhance continuous emotion recognition accuracy. Particular attention is given to regression-based approaches, since continuous emotion prediction is inherently a regression problem rather than a classification task. Regression models are capable of capturing gradual fluctuations in valence and arousal over time, providing finer temporal resolution and greater ecological validity compared to categorical methods that reduce emotions to discrete labels. By focusing on regression strategies, this review addresses the methodological requirements for modelling emotions as dynamic, continuous processes. Additionally, it explores the existing limitations and challenges in this field, providing directions for future research.

2.5.2 Research Questions

What are the most effective techniques for fusing facial expression data, GSR, and pupillometry to improve emotion recognition accuracy, and what limitations remain?

2.5.3 Methodology

The idea is to do state-of-the-art studies that have built an emotion detection model using audiovisual stimuli using physiological signals, by a fusion model.

Search Strategy. A comprehensive literature search was conducted across four major scientific databases: PubMed, Web of Science, Scopus, and IEEE Xplore. Keywords

included combinations of "continuous emotion prediction," "valence and arousal," "regression," "multimodal," "physiological signals," "facial expression recognition," "GSR," and "pupil size." The search covered studies published from 2009 to 2025.

Google Scholar was not used due to its inclusion of grey literature, limited filtering capabilities, and lack of reproducibility compared to structured academic databases.

Inclusion Criteria.

- Studies published in English between 2009 and 2025.
- Peer-reviewed journal and conference papers.
- Continuous prediction of valence and arousal using regression models.
- Use of at least one of FER, GSR, or pupil size as input modalities.
- Presentation of audiovisual stimuli to participants.
- Reporting of quantitative evaluation metrics (e.g., RMSE, Pearson's r, CCC).

Exclusion Criteria.

- Non-English papers.
- Reviews, theses, books.
- Classification-only models.
- Studies using only EEG, speech, or unrelated modalities.
- Studies using only visual, only audio, or image stimuli.

Rationale for Inclusion and Exclusion Criteria. Inclusion and exclusion criteria were defined to focus the review on studies directly relevant to the research question. Continuous prediction using regression models was prioritised to align with the circumplex model of emotion. Studies using at least one of FER, GSR, or pupil size were selected to explore the most effective physiological and visual modalities for fusion techniques. Audiovisual stimuli were required to ensure multimodal emotional elicitation, comparable to real-world applications. Studies focusing solely on EEG or speech modalities, non-peer-reviewed papers, and those using discrete classification methods were excluded to maintain relevance and methodological consistency.

Study Selection. From 213 records identified, 57 duplicates were removed, leaving 156 unique papers. Titles and abstracts of these papers were screened, resulting in the exclusion of 123 records. Full-text assessment was conducted on 33 papers. Following strict application of inclusion and exclusion criteria, 7 studies were included in the final synthesis.

Data Extraction. Key data were extracted from the included studies, focusing on datasets used, input modalities, machine learning models, fusion techniques, target outputs

(arousal and valence), performance metrics, and primary findings. The study selection process is summarised in the PRISMA flow diagram presented in Figure 2.6.

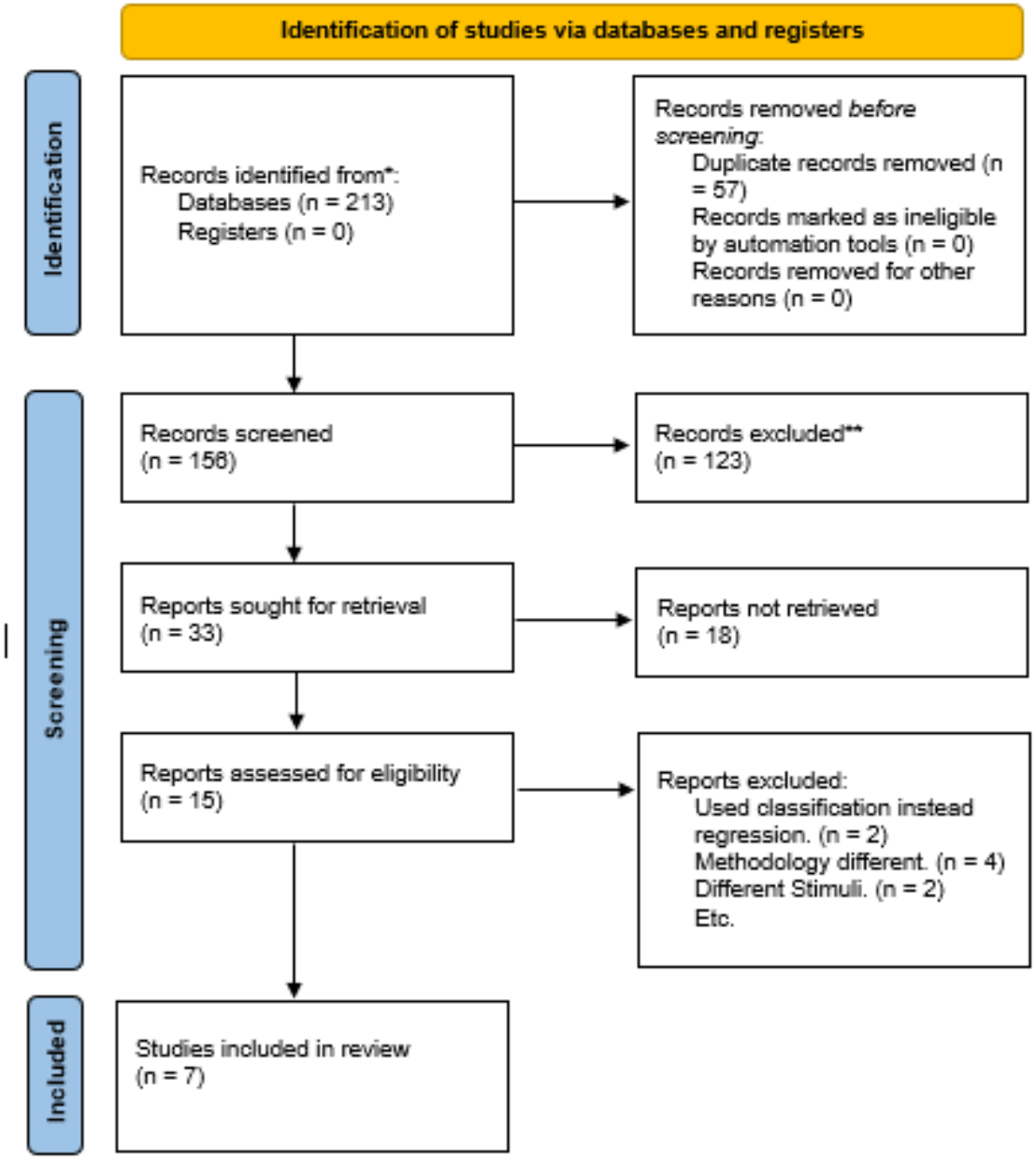


Figure 2.6: PRISMA Flow-diagram.

2.5.4 Results

The details of the included studies, including the data sets, modalities used, fusion techniques, models applied, and performance results, are summarised in Table 2.3.

Table 2.3: Summary of the studies included in the literature review.

Citation	Dataset(s)	Modalities	Method / Model	Metric(s)	Results	Notes
O'Dwyer et al. (2017) [359]	RECOLA	Eye gaze, Pupillometry	LSTM-based regression	CCC	Valence 0.212, Arousal 0.154	Early work on continuous affect prediction from gaze/pupil.
Raju et al. (2021) [360]	RECOLA	FER + Speech	RNN variants with attention	CCC	Valence 0.689, Arousal 0.638	Multimodal fusion, higher CCC than unimodal baselines.
Brady et al. (2016) [361]	AVEC 2016	Video + GSR + Physiological	Feature fusion with regression	CCC	Valence 0.220, Arousal 0.120	Benchmark study; lower CCC due to noisy physiological features.
Patania et al. (2022) [362]	RECOLA	Audio, Video, Physiological	Fusion strategies (deep + shallow)	CCC	Valence 0.424 ± 0.203, Arousal 0.585 ± 0.114	Compared multiple fusion strategies; stronger than unimodal.
Joudeh et al. (2023) [339]	RECOLA	Video + EDA + ECG	Deep learning multimodal fusion	CCC	Valence 0.998, Arousal 0.996	State-of-the-art; very high CCC due to optimised pipeline.
Abhilash et al. (2020) [363]	MOOCs dataset	FER	CNN + XGBoost	Accuracy	NA	Did not perform regression.
Zhang et al. (2024) [364]	Aff-Wild2, Hume-Vidmimic2, C-EXPR-DB	Visual, Audio, Transcripts	MAE features + Transformer fusion + Ensemble	CCC, F1, Acc.	Strong ABAW 2024 results across VA, EXPR, AU	Multimodal, in-the-wild; high compute; not directly comparable to physiology-based methods

Study Characteristics. All 7 included studies used audiovisual stimuli to trigger emotions and aimed to predict continuous valence and arousal. FER was the most common modality (6/7 studies), followed by GSR (3/7) and pupillometry (1/7). Five studies employed multimodal fusion strategies combining visual, audio, and/or physiological signals.

Fusion Techniques and Machine Learning Models. A diverse set of regression-based models and fusion techniques was employed:

- Feature-level fusion: Used in four studies, this technique combines extracted features from different modalities into a single feature vector before model training.
- Decision-level fusion: Two studies adopted this approach, combining predictions from unimodal models.
- Hybrid fusion: One study utilised sequential or hierarchical fusion techniques.

Machine learning models used included: Support Vector Regression (SVR), Extreme Gradient Boosting (XGBoost), Ensemble models, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Multimodal Transformers

Quantitative Performance. Unimodal approaches reported modest performance. In pupillometry, O'Dwyer et al. [359] achieved CCCs of 0.212 for valence and 0.154 for arousal, reflecting the influence of luminance and task context. Brady et al. [361] reported low GSR-only baselines (CCC = 0.220 valence, 0.120 arousal), illustrating the limited predictive power of single biosignals. FER was more effective: Raju et al. [360] obtained CCCs of 0.689 (valence) and 0.638 (arousal) using RNNs with attention.

Multimodal approaches consistently surpassed unimodal baselines, with performance strongly influenced by fusion strategies. Brady et al. [361] demonstrated early fusion of audiovisual and physiological features, but CCCs remained low (0.220 valence, 0.120 arousal). Patania et al. [362] systematically compared late fusion, early fusion, and hybrid fusion schemes, achieving improved CCCs of 0.424 (valence) and 0.585 (arousal), showing that the choice of fusion method materially affects outcomes. Joudeh et al. [339] employed deep fusion of video, EDA, and ECG in an end-to-end learning pipeline, achieving near-perfect CCCs of 0.998 (valence) and 0.996 (arousal) on RECOLA, representing the strongest reported physiological multimodal benchmark. Finally, Zhang et al. [364] advanced ensemble-level fusion by integrating masked autoencoder features with transformer-based fusion across audio, video, and transcripts. Their system achieved state-of-the-art results across valence–arousal, expression recognition, and action unit detection tasks in the ABAW 2024 competition.

Taken together, the literature demonstrates that unimodal biosignals such as pupil size or GSR provide limited predictive power, whereas multimodal fusion approaches yield substantial gains. The performance gap across studies underscores the critical role of fusion strategies: from early and late fusion to deep learning and transformer-based ensembles, progressively more sophisticated fusion frameworks have driven the field toward state-of-the-art affective behaviour analysis.

Datasets and Stimuli. Most studies used the RECOLA dataset, with a minority employing proprietary datasets collected under controlled audiovisual stimulation conditions. All studies were evaluated offline using continuous annotation of valence and arousal.

2.5.5 Discussion

This systematic literature review explored fusion techniques for FER, GSR, and pupillometry data to enhance continuous prediction of emotional states, specifically valence and arousal, using regression models. By focusing on studies employing audiovisual stimuli and reporting quantitative performance metrics, the review highlights trends and limitations relevant to emotion recognition systems designed around continuous, dimensional modelling.

Effectiveness of Fusion Strategies. Across the included studies, multimodal fusion consistently improved continuous emotion prediction compared to unimodal approaches. Feature-level fusion, where preprocessed data streams are combined before input into the regression model, emerged as particularly effective. Studies using this approach (e.g., Brady et al. [361], Zhang et al. [364], Patania et al. [362]) demonstrated

that combining features from FER, GSR, and pupillometry at an early stage allowed machine learning models to capture complementary aspects of affective responses, enhancing the prediction of both valence and arousal. Decision-level fusion, in which predictions from separate unimodal models are combined at the output stage, was also used effectively, particularly in scenarios involving asynchronous data or variable-quality input streams (e.g., O'Dwyer et al. [359], Joudeh et al. [339]). However, decision-level fusion typically limits interaction between modalities during modelling, potentially restricting its ability to fully exploit inter-modal relationships.

More recently, transformer-based architectures and attention mechanisms (as used by Zhang et al. [364] and Raju et al. [360]) showed strong potential in managing complex, time-dependent emotion signals. These models dynamically weighted modality contributions based on temporal relevance, which is particularly important in continuous prediction tasks where the salience of each modality can fluctuate over time. Hybrid fusion strategies, combining feature- and decision-level approaches, remain underexplored but offer potential benefits in balancing model complexity with interpretability.

Role of FER, GSR, and Pupillometry. Facial expression features were consistently found to contribute most strongly to valence prediction, reflecting their clear links to observable emotional expressions. GSR, by contrast, played a more significant role in predicting arousal due to its sensitivity to autonomic nervous system responses during emotional stimulation. However, pupillometry remains significantly underutilised. Despite its potential to capture subtle, dynamic indicators of affective arousal, pupil size was rarely incorporated as a primary modality. Where pupil-related features were used (notably in O'Dwyer et al.), they were often treated statically rather than as a dynamic time-series signal. This represents a notable gap in current fusion strategies. Effective preprocessing, such as z-score normalisation and temporal segmentation, was crucial for integrating modalities with different data structures and sampling rates. Without these steps, modalities like FER, with high-dimensional visual data, risked overwhelming physiological signals like GSR or pupil size during fusion.

Limitations of Current Research. Despite promising advances, several limitations remain across the studies included in this review, such as most studies used controlled datasets, such as RECOLA, recorded under laboratory conditions. This restricts the generalizability of findings to real-world applications, where emotional responses are less constrained and sensor quality may vary. Dynamic pupillometry remains an underused modality. Few studies incorporated pupil size as a continuous, time-resolved signal alongside FER and GSR, limiting understanding of its potential contribution to valence and arousal prediction. All reviewed models were evaluated offline. No study implemented real-time continuous emotion prediction combining FER, GSR, and pupil-

lometry during audiovisual stimulation. Explainability was rarely addressed. While deep learning and transformer-based fusion methods improved prediction accuracy, most models functioned as black boxes, providing limited insight into how each modality contributed over time. Robustness to missing or degraded modality streams was not systematically evaluated, despite the likelihood of such issues in real-world settings.

Implications and Future Directions. Future research should prioritise developing fusion models that integrate FER, GSR, and dynamic pupillometry signals using attention-based or transformer architectures, allowing models to adaptively focus on the most informative modalities over time. Expanding the use of temporal modelling for pupillary signals represents an important research opportunity, given the known affective relevance of pupil dynamics. Additionally, future systems should incorporate explainable AI techniques to improve interpretability. Methodologies should also address modality dropout, ensuring robust emotion recognition even when one or more data streams are unavailable or unreliable. Finally, the development and adoption of standardised, multimodal datasets collected in naturalistic environments with continuous annotation of valence and arousal will be critical for advancing the field and enabling meaningful comparison across models.

2.5.6 Conclusion

This systematic review confirms that multimodal regression-based models using audiovisual stimuli and physiological or visual inputs provide reliable continuous arousal and valence predictions. Feature-level fusion of FER, GSR, and pupillometry features appears most effective for improving emotion recognition accuracy. Further development of hybrid fusion techniques and exploration of underused physiological modalities are recommended.

2.6 Limitations and Future Directions

Emotion detection using physiological signals has significant potential, but several limitations must be addressed for reliable real-world use. Many challenges are common and systemic across modalities, such as pupil dilation, GSR and PPG-based HR, while others are modality-specific. This section first outlines the common limitations across modalities and the specific challenges associated with pupil dilation and GSR. Our analysis focuses on these two modalities due to their increasing prominence in emotion research and inclusion in our empirical investigations.

Common Limitations Across Physiological Modalities

A primary standard limitation is the over-reliance on data collected in tightly controlled laboratory settings. Such settings often fail to reflect the complexity and variability of real-world contexts, reducing ecological validity and limiting the generalizability of findings [192], [217], [292]. In addition, small, homogeneous or biased samples further hinder the development of robust models applicable to diverse populations [36], [258].

Another common problem is the susceptibility of physiological signals to noise and distortions. Wearable sensors such as GSR and PPG are highly sensitive to motion, environmental conditions (e.g., temperature, lighting), and sensor placement [258], [260]. These noises can distort signal quality, making emotion inference difficult in real-world scenarios. Furthermore, most current systems rely on classifying discrete emotional states, even though emotional experiences are inherently continuous and context-dependent. This simplification reduces the ability of systems to capture complex or mixed emotional states [254], [365].

Multimodal ML approaches introduce their own set of complications. Synchronisation and fusion of multiple data streams (e.g. GSR, PPG, EEG) pose technical challenges. Feature-level fusion may introduce redundancy or noise, while deep learning models require large datasets that are often unavailable [217], [252]. Furthermore, the lack of standardised protocols and consistent emotion elicitation methods makes cross-study comparisons difficult [192], [366]. Finally, ethical considerations - user privacy, consent and misuse of emotion data - remain under-addressed despite their growing importance [36], [259].

Modality-Specific Limitations

Pupil Size. Emotion detection by pupil size is particularly sensitive to ambient light conditions, which strongly influence pupil size and can confound emotional signals [274], [297]. In addition, pupil responses vary significantly between individuals due to cognitive load, personality traits, health conditions and genetics [367], [368]. These inter-individual differences challenge the development of generalisable models. Most studies of pupil dilation focus on basic emotional states and are conducted in controlled environments, limiting their applicability to dynamic, real-world settings [203], [292].

Galvanic Skin Response and PPG-Based HR. GSR-based systems often emphasise static features and discrete emotion labels, neglecting the temporal evolution of arousal and the continuous nature of emotions [365], [369]. Furthermore, the GSR is limited in its emotional specificity. While it can reliably indicate arousal, it struggles to dis-

criminate between emotions with similar levels of arousal but different valence, such as fear and excitement [370]. Similar problems apply to PPG-based heart rate signals, which primarily reflect autonomic arousal and are prone to motion noise. External conditions such as temperature, skin colour and physiological delays further reduce the reliability of PPG signals. Inconsistencies in dataset formats and measurement practices compound these challenges, making cross-study validation difficult [302].

Focus and Contributions of This Work

This paper explores the limitations and potential of pupil dilation and GSR-based emotion detection, which are gaining attention for their use in wearable and mobile applications due to their non-intrusive nature and suitability for real-time monitoring. These physiological modalities provide complementary insights into emotional states - while GSR reflects autonomic arousal, pupil dilation is linked to cognitive and affective processing, providing a multidimensional view of emotion.

In our research, we collected data from healthy participants, carefully screening for individuals without mental health conditions such as alexithymia, autism, PDs, anxiety or depression. We used a comprehensive set of sensors, including GSR, PPG, FER, an eye tracker for pupil size, and EEG. The aim was to address several key challenges in emotion recognition. A novel aspect of our study was the development of a method to remove the influence of ambient luminosity on pupil size, allowing us to more accurately isolate the emotional component of pupil dilation.

Following existing methods, we extracted time- and frequency-domain features from the GSR signals and performed min-max normalisation for each stimulus and participant to reduce variability. We converted basic emotions into a continuous arousal-valence representation for facial expression data, allowing for a finer understanding of emotional responses.

Our study also focused on multimodal integration, combining complementary modalities to enhance emotional specificity. By fusing signals such as GSR and pupil dilation with FER data, we aimed to build a more robust emotion detection model. This approach addresses the limitations of using single modality data and improves the accuracy and applicability of emotion detection systems.

This research addresses several critical limitations in emotion detection using physiological signals. By incorporating dynamic modelling, advanced pre-processing techniques and multimodal fusion, we have laid the foundations for future developments in real-time emotion detection. Future work should continue to validate these models in real-world contexts, improve their ecological validity, and explore context-aware systems to enhance the practical use of emotion detection technologies.

2.7 Generalisation and Transfer Learning

The challenge of generalising emotion recognition models across different subjects, environments and scenarios is a key hurdle in emotion recognition research. The effectiveness of emotion detection models relies heavily on their ability to adapt and perform well across different populations, contexts, and changing environmental conditions. However, this generalisation is not always straightforward, especially with multimodal data (e.g. physiological signals, facial expressions, voice), which can vary significantly between individuals or settings [187].

The model can capture the continuous nature of emotions by using regression rather than classification to detect emotional states. Unlike classification, which restricts emotions to a limited set of discrete labels (such as happy, sad, or angry), regression predicts the intensity of emotions continuously, such as arousal and valence levels [371]. This has several advantages:

- Avoid bias towards a few dominant emotions: In a classification model, emotions such as happiness or anger may dominate due to their over-representation in the data sets, leading to biased predictions [372]. Regression, however, can ensure that all emotional variation, including less frequent or subtle emotions, is considered, resulting in a more balanced representation [373].
- Capture subtle emotional differences: Emotions are often complex and nuanced, and small changes in emotional intensity can be significant. Regression models can capture these fine-grained differences, making them more suitable for nuanced applications that require sensitivity to emotional variations, such as mental health assessments or personalised user experiences in HCI [374].
- Better reflect real-world emotional experiences: Unlike classification, which forces emotions into predefined categories, regression models recognise that emotions exist on a spectrum [375]. For example, a person's emotional state may not be strictly 'happy' or 'sad', but may vary between moderate happiness and extreme joy or between mild sadness and deep grief. By modelling emotions as continuous variables, regression more accurately represents these fluid emotional transitions and more closely matches real-world emotional dynamics.

This makes regression particularly useful for applications where gradual emotional changes are critical. Regression provides a more accurate analysis tool in affective computing, where systems need to detect and respond to continuous emotional states in real time, and mental health monitoring, where emotional changes over time can indicate changes in well-being or mental health. It is also essential in HCI, where continuous feedback on emotional states can improve user engagement and system responsiveness.

Furthermore, transfer learning can improve the generalisation of regression models across different subjects or environments. By fine-tuning a model trained on a large,

diverse dataset (e.g., capturing emotional data from a wide range of individuals and settings), it can be adapted to new users or specific contexts with minimal retraining, overcoming the limitations of traditional model generalisation.

Regression approaches offer a more flexible and accurate representation of emotional states, particularly suitable for applications requiring sensitivity to emotional intensity and gradual change. These models can be even more robust and adaptable when combined with transfer learning, ensuring that emotion recognition systems perform well in diverse scenarios and subject populations.

2.7.1 General Strategies to Address Challenges

- **Data enhancement:** In addition to resampling and generating synthetic data, augmenting the data set with different noise patterns and variability can improve generalisation in real-world scenarios.
- **Multimodal Fusion:** Combining data from multiple sources (e.g., physiological signals, facial expressions, and speech data) can help overcome noise-related problems in individual modalities. A robust multimodal fusion approach allows the system to exploit the complementary strengths of each signal type, making the emotion recognition system more accurate and robust.
- **Real-time monitoring and adaptation:** Implementing adaptive learning techniques, where the model continuously improves as more data is collected in real time, can help mitigate the problems of noise and imbalance over time, especially in clinical or dynamic environments.
- **Cross Validation:** Applying cross-validation techniques to different subsets of the data ensures that the model does not over-fit to noisy or imbalanced parts of the dataset, providing more generalisable results.

Addressing these challenges can significantly improve the robustness and accuracy of emotion detection models using physiological signals.

2.8 Conclusion

This literature review highlights the critical role of ML models in advancing emotion detection by leveraging physiological signals and facial expressions. Integrating multimodal data has shown promising improvements in accuracy; however, challenges remain in identifying the most relevant features for robust and real-time emotion detection. A key research direction is the development of optimised feature extraction techniques that enhance regression models for detecting continuous emotional states. Future work should improve feature selection methods, refine fusion strategies for multimodal data, and develop computationally efficient models that maintain high predict-

ive performance. Addressing these aspects will provide more reliable, interpretable, and real-world-applicable emotion detection systems.

Chapter 3

Development of Methodology for Emotion Detection Model

This chapter outlines the methodology used to develop an emotion detection model, structured around two sequential studies and the application of advanced machine learning techniques. It begins with a pilot study that helped us establish effective data collection and analysis protocols using physiological signals. The pilot study's findings informed the design of the main study, in which we implemented refined data collection and processing procedures. Manual pre-processing of physiological signals was performed to improve data quality and ensure compatibility with machine learning algorithms. Finally, we applied advanced machine learning techniques to train and evaluate the emotion detection model.

3.1 Pilot Study

This pilot study examines participants' emotional responses to audiovisual stimuli using facial expression analysis, eye-tracking, GSR, and PPG-based HR. The goal is to assess whether these bio-signals capture emotional arousal and valence variations, aiding the development of a machine-learning emotion detection model.

Unlike traditional self-reported methods, which can be biased, physiological signals offer objective, real-time measures of emotional states. Markers such as Pupil dilation, GSR, and PPG-based HR reflect ANS activity and provide reliable indicators of emotional arousal. Integrating these with facial emotion detection data can create a comprehensive emotion detection system.

3.1.1 Significance and Objectives of the Pilot Study

The pilot study played a crucial role in evaluating the methodology and ensuring the reliability of the data collection process before the main study. Its primary aim was to

assess the feasibility of using multimodal physiological signals, such as pupil size, GSR, PPG, and FER, for emotion detection.

The specific objectives of the pilot study were as follows:

1. To assess the feasibility of using multimodal physiological signals for detecting emotions.
2. To evaluate the consistency of participants' physiological responses to emotional audiovisual stimuli across different modalities.
3. To investigate how stimulus-induced emotion affects each signal type (pupil size, GSR, PPG, FER).
4. To test the effectiveness of different biosensors and data recording techniques.
5. To identify potential distortions or noise in the data (e.g., motion-induced noise in GSR, lighting effects on pupil size).
6. To optimise the timing of stimulus presentation for enhancing emotional engagement.
7. To evaluate participant compliance and refine task instructions.
8. To identify technical challenges and inform refinements in data collection, preprocessing, feature extraction, and model training strategies.

The outcomes of this pilot study laid the groundwork for effectively integrating physiological signal features with machine learning algorithms, ultimately improving the efficiency and reliability of automated emotion detection systems.

3.1.2 Experimental Approach

To examine participants' emotional responses, we designed an experiment exposing them to carefully curated audiovisual stimuli that evoke different emotional states. These audiovisual clips were selected based on validated emotion elicitation databases to ensure diverse emotions [376], [377], including happiness, sadness, fear, and neutrality. During the experiment, simultaneous recording of facial expressions, eye movements, pupil size, GSR, and PPG data was performed to capture real-time physiological changes associated with emotional arousal and valence.

A critical aspect of this study was to analyse the correlations between these physiological signals and self-reported emotional experiences. By comparing participants' subjective ratings with their recorded physiological responses, we aimed to assess each modality's reliability and predictive power in emotion detection.

3.1.3 Participants

A total of 45 participants (20 males, 25 females) were recruited for this study, consisting of students and staff at the University of Essex who voluntarily agreed to particip-

ate. Before the experiment, all participants provided signed informed consent, ensuring their awareness of the study's objectives, procedures, and data usage. Ethical approval for this study was obtained under application number **ETH2223-0088** following the institutional research guidelines of the University of Essex.

There were no strict screening criteria apart from age; participants must be between 18 and 65 to ensure a diverse representation of adult emotional responses. The research complied with ethical guidelines, ensuring voluntary involvement, safeguarding confidentiality, and allowing participants to withdraw without repercussions.

3.1.4 Experimental Setup

The study was conducted in a well-lit laboratory at the University of Essex, School of Computer Science and Electronic Engineering (CSEE). The experimental setup consisted of two screens: an experimenter screen for monitoring the study, an experiment screen for presenting emotional stimuli to the participants, and a mouse for responding to the survey questionnaires.

Figure 3.1: Tobii Pro Nano eye tracker used in the study.

A high-resolution camera (Intel RealSense Module D430 + RGB Camera, 1920 × 1080 resolution, 30 fps) (https://www.intelrealsense.com/depth-camera-d435i/) was mounted on the experiment screen to capture participants' facial expressions throughout the study. Below the screen, a Tobii Pro Nano eye-tracker (https://connect.tobii.com) was installed to record pupil size and the gaze value of eye movements (see Figure 3.1). Participants were required to sit within a 65 cm range from the screen to ensure accurate eye-tracking measurements, as presented in Figure 3.2.

Additionally, GSR and PPG sensors from Shimmer were used to monitor physiological responses. Two GSR electrodes were attached to the index and middle fingers, while a PPG electrode was placed on the ring finger of one hand. These physiological signals and facial emotion detection provided a comprehensive dataset for assessing emotional responses.

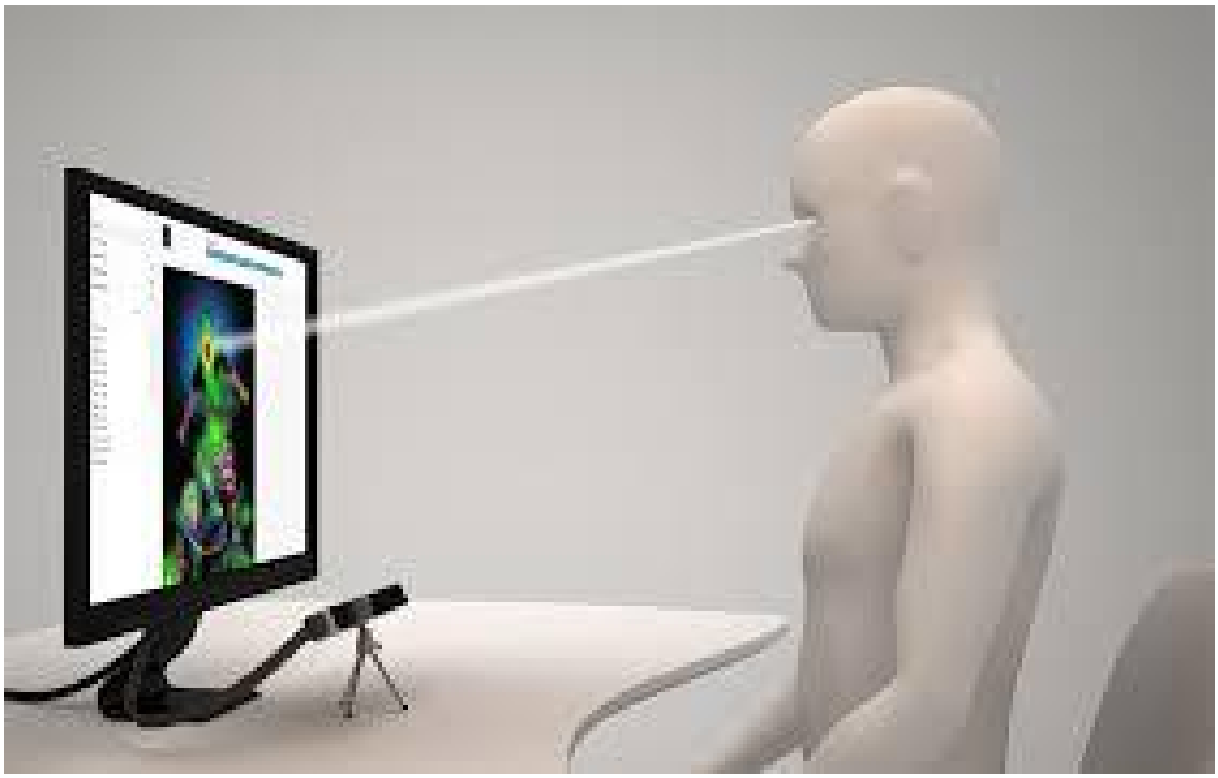


Figure 3.2: Collecting pupil data using an eye tracker [378].

3.1.5 Procedure

The data collection process was conducted using iMotions Software [379], which controlled the presentation of stimuli and synchronised data recording across various sensors. The procedure began with an initial in-built Tobii Pro Nano eye-tracker calibration to ensure accurate tracking, proper alignment, and reliable data capture of the participant's pupil size and gaze positions.

After calibration, participants were shown an instruction screen outlining the experimental procedure. Following the instructions, participants were shown a 60-second grey screen with a black cross at the centre to capture their physiological responses in a neutral emotional state. This phase allowed for measuring their physiological and baseline responses before exposure to emotional stimuli.

The experiment's central part involved presenting 20 audiovisual clips, with an average duration of 35 seconds and a maximum duration of 58 seconds, randomly ordered for each participant (see Figure 3.3). The clips were selected to represent five distinct emotional categories based on Russell's circumplex 2D emotional model [32] (four audiovisual clips per category): high arousal with positive valence, high arousal with negative valence, low arousal with positive valence, low arousal with negative valence, and neutral clips [380]. These categories were designed to capture various emotional states and elicit diverse physiological and subjective responses [1].

[1]Link to audiovisual clips: https://essexuniversity.box.com/s/

It is important to note that iMotions software was employed solely for stimulus presentation and multimodal data acquisition. All subsequent stages of this study, including preprocessing, feature extraction, and the development of machine learning pipelines, were designed and implemented independently by the candidate. This ensured that the original technical contributions of this work lie beyond data collection and focus on the bespoke analytical framework developed.

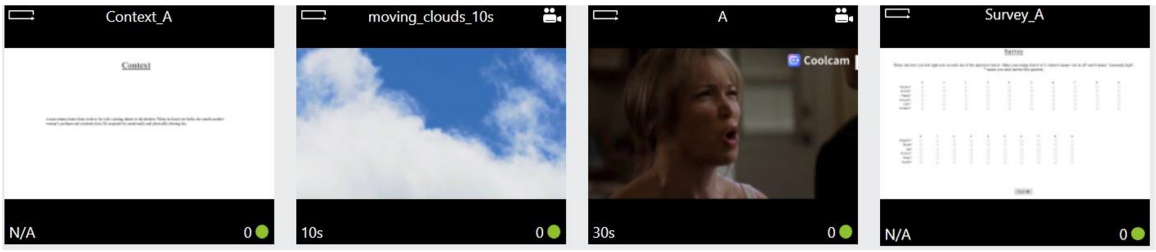


Figure 3.3: The presentation flow of each audiovisual clip.

Before each clip, we provided contextual information to establish a foundation for understanding the scene. Since some audiovisual clips were excerpts from movies, it was necessary to explain the background to elicit the intended emotions effectively. To minimise emotional transfer from contextual information to audiovisual clips, a 5-second neutral video, featuring a cloud-moving animation, was shown between each emotional clip. This neutral stimulus aimed to reset participants' emotional states before presenting the following video clip. After viewing each clip, participants were asked to complete a survey that assessed their emotional responses to the clips. As mentioned in the literature review chapter (2), we used a multidimensional emotional questionnaire to create the survey. The survey consisted of 12 emotion-related questions (12 dimensions), adapted from prior work [35], [164], [381] and extended from [164] to cover a broader range of affective states. Specifically, the questionnaire included the following emotions: **positive, excitement, happiness, amusement, calmness, contentment, negative, anger, fear, anxiety, sadness, and boredom**. Participants rated each of these 12 items on a scale from 0 to 9, where 0 indicated the absence of the emotion and 9 represented the highest intensity. After each audiovisual clip, participants were asked to complete the 12-item survey. The items were presented in a fixed order for consistency, and each rating was recorded through the iMotions interface (see Figure 3.4). This procedure allowed for a comprehensive assessment of the emotional intensity elicited by each stimulus, providing a multi-dimensional self-report complement to the physiological signals.

For transparency and reproducibility, the full text of the survey items and their rating

fovxpkm7k932p85wymsrk84c2c46md8f

Figure 3.4: Survey Questionnaires for Pilot Study.

format are provided in Appendix 6.1.

3.1.6 Emotion Labelling

After completing the experiments, the survey data collected from iMotions were used to compute the emotion labels. Unlike prior studies that directly average or categorise participant ratings, we employed Russell's Circumplex Model of emotion classification, where emotions are organised along two continuous dimensions: valence (pleasantness vs. unpleasantness) and arousal (activation vs. deactivation) [32]. This framework allows emotions to be mapped within a circular space, preserving their theoretical structure.

To derive individualised valence-arousal values from participants' responses, we applied the *Individual Differences Scaling (INDSCAL)* multidimensional scaling technique [164]. This method projected the original 12-dimensional (12D) ratings (12 emotion-related questions per audiovisual clip) into a two-dimensional (2D) circumplex space, while also capturing individual weighting factors. In doing so, INDSCAL provided both a shared group-level perceptual space and participant-specific variations, thereby addressing the limitations of conventional approaches that rely on categorical labels or aggregated averages. This projection of 12D space to 2D space not only ensured consistency with established psychological theory but also offered a mathematically coherent and interpretable framework for emotion labelling, marking an important contribution in the context of multimodal affect recognition.

Importantly, this labelling strategy provided more reliable ground-truth data for model training, ensuring that the subsequent multimodal integration (pupil size, GSR, and FER) was guided by labels that preserved both common emotion structure and individual differences. This strengthened the robustness and ecological validity of our emotion recognition pipeline.

As mentioned above, INDSCAL decomposes the aggregated group similarity matrix into:

- a **common group configuration** $\mathbf{X}$, representing the positions of the stimuli in a shared k-dimensional space, and
- **individual subject weights** $\mathbf{W}_s$, which rescale the dimensions for each subject s.

Formally, given a set of dissimilarity matrices $\{\Delta^{(s)}\}$ for subjects $s = 1, \ldots, S$, INDSCAL solves:

$$\delta_{ij}^{(s)} \approx \left(\sum_{d=1}^{k} w_{sd} (x_{id} - x_{jd})^2 \right)^{1/2}, \tag{3.1}$$

where $\delta_{ij}^{(s)}$ is the perceived dissimilarity between stimuli i and j for subject s, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ik})$ are the coordinates of stimulus i in the group space, and w_{sd} is the weight of subject s on dimension d.

The subject weights $\mathbf{W}_s = (w_{s1}, w_{s2})$ indicate how strongly each participant used the *valence* (dimension 1) and *arousal* (dimension 2) axes. These weights are estimated through least-squares minimisation of the stress function:

$$\text{Stress} = \sum_{s=1}^{S} \sum_{i<j} \left(\delta_{ij}^{(s)} - d_{ij}^{(s)}(\mathbf{X}, \mathbf{W}_s) \right)^2, \tag{3.2}$$

where $d_{ij}^{(s)}$ is the model-predicted distance.

Distance Metrics. To quantify the similarity between 20 audiovisual clip stimuli, we tested multiple distance metrics, including Manhattan and Euclidean distance. Both produced interpretable structures, but the Euclidean distance provided the best fit with the INDSCAL stress function and yielded a more coherent 2D representation consistent with Russell's circumplex model. The results using the Manhattan metric are included in the Appendix for comparison.

Equation for Distance Matrix Calculation. For a set of N stimuli (where $N = 20$) and M scales (where $M = 12$), let:

- $S_i = (s_{i1}, s_{i2}, \ldots, s_{iM})$ be the vector representation of stimulus i across M scales,
- $S_j = (s_{j1}, s_{j2}, \ldots, s_{jM})$ be the vector representation of stimulus j.

The Euclidean distance $D(i, j)$ between two stimuli S_i and S_j is:

$$D(i, j) = \sqrt{\sum_{k=1}^{M} (s_{ik} - s_{jk})^2}. \tag{3.3}$$

This produced a 20×20 symmetric distance matrix for each participant (see Figure 3.5), which was then input into INDSCAL.

Implementation. Since Python does not include a direct implementation of INDSCAL, we used the Statistical Package for the Social Sciences (SPSS) to compute the model. SPSS generated the 2D group configuration and individual subject weights automatically. The output included:

- **2D group space** - the shared valence–arousal coordinates of stimuli,
- **subject weights** - rescaling factors showing how strongly each subject relied on each dimension.

These subject weights enabled personalised interpretations, ensuring each participant's responses were preserved within the common dimensional structure.

A1 LP	A2 LP	A3 LP	A4 LP	A HN	B HN	C LN	F HN	G HP	H HP	J Ne	K Ne	M LN	N LN	O LN	P HP	Q HP	U Ne	V Ne	W HN
0	1.41421	5.83095	4.24264	6.9282	10.4403	5.83095	9.48683	3.87298	6	4.69042	4.24264	7	7.68115	9.16515	7.2111	7.07107	5.56776	6.9282	7.81025
1.41421	**0**	5.47723	3.4641	7.07107	10.5357	6	9.59166	2.64575	5.09902	4.89898	4.47214	7.14143	7.81025	9.27362	6.16441	6	4.79583	7.07107	7.93725
5.83095	5.47723	**0**	5.83095	8	11.1803	7.07107	10.2956	5	5.83095	4.69042	5.83095	8.06226	8.66025	10	6.16441	6.48074	5.91608	6.9282	8.77496
4.24264	3.4641	5.83095	**0**	8.83176	11.7898	8	10.9545	3.60555	6.48074	7.2111	6.9282	8.88819	9.43398	10.6771	6.48074	6.9282	6.7082	8.83176	9.53939
6.9282	7.07107	8	8.83176	**0**	5.91608	5.47723	6.48074	7.14143	6.9282	5.83095	5.47723	3.31662	4.3589	6.32456	8	7.87401	6.55744	3.74166	2.64575
10.4403	10.5357	11.1803	11.7898	5.91608	**0**	8.18535	7.4162	10.583	10.4403	9.74679	9.53939	4	4.69042	5	11.1803	11.0905	10.198	5	5.83095
5.83095	6	7.07107	8	5.47723	8.18535	**0**	9.38083	6.08276	5.83095	4.47214	4	4.79583	5	6.48074	7.07107	6.9282	5.38516	5.47723	5.91608
9.48683	9.59166	10.2956	10.9545	6.48074	7.4162	9.38083	**0**	9.64365	9.48683	8.7178	8.48528	6.55744	8.06226	9.48683	10.2956	10.198	9.21954	6.48074	8.18535
3.87298	2.64575	5	3.60555	7.14143	10.583	6.08276	9.64365	**0**	3	5	4.58258	7.2111	7.87401	9.32738	3.60555	3.60555	3.16228	7.14143	8
6	5.09902	5.83095	6.48074	6.9282	10.4403	5.83095	9.48683	3	**0**	4.69042	4.24264	7	7.68115	9.16515	2	1.41421	1	6.9282	7.81025
4.69042	4.89898	4.69042	7.2111	5.83095	9.74679	4.47214	8.7178	5	4.69042	**0**	2	5.91608	6.7082	8.3666	6.16441	6	4.12311	5.09902	6.85565
4.24264	4.47214	5.83095	6.9282	5.47723	9.53939	4	8.48528	4.58258	4.24264	2	**0**	5.56776	6.40312	8.12404	5.83095	5.65685	3.60555	5.47723	6.55744
7	7.14143	8.06226	8.88819	3.31662	4	4.79583	6.55744	7.2111	7	5.91608	5.56776	**0**	2.44949	4.58258	8.06226	7.93725	6.63325	2.23607	3.74166
7.68115	7.81025	8.66025	9.43398	4.3589	4.69042	5	8.06226	7.87401	7.68115	6.7082	6.40312	2.44949	**0**	6.08276	8.66025	8.544	7.34847	3.60555	3.16228
9.16515	9.27362	10	10.6771	6.32456	5	6.48074	9.48683	9.32738	9.16515	8.3666	8.12404	4.58258	6.08276	**0**	10	9.89949	8.88819	5.47723	7
7.2111	6.16441	6.16441	6.48074	8	11.1803	7.07107	10.2956	3.60555	2	6.16441	5.83095	8.06226	8.66025	10	**0**	1.41421	3	8	8.77496
7.07107	6	6.48074	6.9282	7.87401	11.0905	6.9282	10.198	3.60555	1.41421	6	5.65685	7.93725	8.544	9.89949	1.41421	**0**	2.23607	7.87401	8.66025
5.56776	4.79583	5.91608	6.7082	6.55744	10.198	5.38516	9.21954	3.16228	1	4.12311	3.60555	6.63325	7.34847	8.88819	3	2.23607	**0**	6.55744	7.48331
6.9282	7.07107	6.9282	8.83176	3.74166	5	5.47723	6.48074	7.14143	6.9282	5.09902	5.47723	2.23607	3.60555	5.47723	8	7.87401	6.55744	**0**	4.3589
7.81025	7.93725	8.77496	9.53939	2.64575	5.83095	5.91608	8.18535	8	7.81025	6.85565	6.55744	3.74166	3.16228	7	8.77496	8.66025	7.48331	4.3589	**0**

Figure 3.5: Distance matrix for one participant.

Outcome. Through this process, the original 12D responses were projected into a 2D space, with the x-axis corresponding to **valence** and the y-axis to **arousal**. These derived labels formed the continuous ground truth used for subsequent model training and analysis (see Section 4.2.1).

3.1.7 Data Pre-processing and Analysis of Physiological Signals

Physiological signals were recorded using specialised sensors while participants were exposed to emotional stimuli during the experiment. GSR sensors measured skin conductance changes, a facial camera captured facial expressions, and an eye-tracking system recorded pupil diameter to track visual attention and arousal levels. Each physiological signal requires specific preprocessing steps to ensure data quality and minimise noise. For GSR data, a band-pass Butterworth filter of second order with a cutoff frequency of 5 Hz was applied to remove high-frequency noise and baseline drift. For pupil dilation data, eye blinks, which introduce sudden missing values, were identified using the blink detection algorithm implemented in the iMotions eye-tracking module. In this approach, blinks were flagged when both eyes showed consistent missing data for a short interval, distinguishing them from random tracking noise. These blink events were then replaced with null values to prevent distortions in the signal.

Our study used iMotions software (https://imotions.com/) to collect data for FER and physiological data. iMotions integrates Affectiva's AFFDEX [382] and Realeyes algorithms to detect seven core emotions (joy, anger, fear, surprise, sadness, contempt, and disgust) by analysing facial movements. It captures 20 facial action units (e.g., cheek raiser, lip corner puller), head movements, blinks, and valence, providing a comprehensive emotional profile. Additionally, it synchronises FER with physiological signals like GSR and EEG for multimodal emotion analysis, making it valuable in research, UX testing, and mental health assessment [77].

After initial preprocessing, all physiological data underwent a null imputation process using interpolation to handle missing values. Interpolation is a technique used to estimate values between known data points, ensuring a continuous dataset representation. Specifically, we applied linear interpolation, which assumes a linear relationship between neighbouring data points to estimate missing values. This method computes values at regular intervals between two adjacent known points, preserving the overall trend and structure of the data. For example, if two data points are "1" and "2," and two intermediate values are missing, linear interpolation would estimate them as "1.33" and "1.66" by following a linear trend. This approach was beneficial for handling missing data in our time-series dataset, ensuring that essential patterns and trends remained intact while minimising information loss.

FER Analysis

Facial expressions were analysed using the iMotions software equipped with the AFFDEX 2.0 toolkit. This toolkit applies a deep learning-based facial expression recognition model to each video frame and outputs intensity scores (continuous values ranging from 0 to 100) for seven basic emotions: anger, fear, disgust, sadness, contempt, joy, and surprise. These emotion intensity values were the only outputs taken from the proprietary software, serving as raw input features for our subsequent analysis. Unlike many studies that either use categorical outputs directly (e.g., majority voting across frames) or train black-box regressors from AU activations to valence-arousal values, we, with the help of students, developed a novel pipeline to transform these discrete intensities into a continuous, interpretable representation within a theoretically grounded affective space.

To achieve this, we mapped the seven basic emotions into Russell's Circumplex Model of Affect [32]. Each emotion was assigned an angular position within the circumplex space based on established associations of affective words and prior empirical mappings (see Table 3.1). In this formulation, each emotion contributes both a direction (its circumplex angle) and a magnitude (its intensity value from iMotions). By decomposing the emotion intensities into Cartesian coordinates, summing contributions across emotions, and averaging across time, we obtained a continuous trajectory in valence-arousal space.

This vectorial mapping framework is novel in three respects: (i) it respects the geometry of Russell's circumplex by treating emotions as vectors rather than independent categories, (ii) it allows co-occurring emotions to be naturally combined into a single affective state rather than forcing a dominant label, and (iii) it provides a psychologically interpretable bridge between categorical FER outputs and dimensional affect modeling. To the best of our knowledge, no previous FER-based study has integrated emotion intensities into a circumplex representation in this vectorial form, making our approach both theoretically motivated and practically distinctive within the affective computing literature.

Formally, we represented each emotion as a vector in polar coordinates, where the angle was derived from Russell's model and the magnitude was the iMotions intensity score. These were then converted into Cartesian coordinates as follows:

$$x_i = M_i \cos(\theta_i) \tag{3.4}$$

$$y_i = M_i \sin(\theta_i) \tag{3.5}$$

where:

- M_i = intensity of the i^{th} emotion (from iMotions),
- θ_i = circumplex angle assigned to that emotion (Table 3.1),

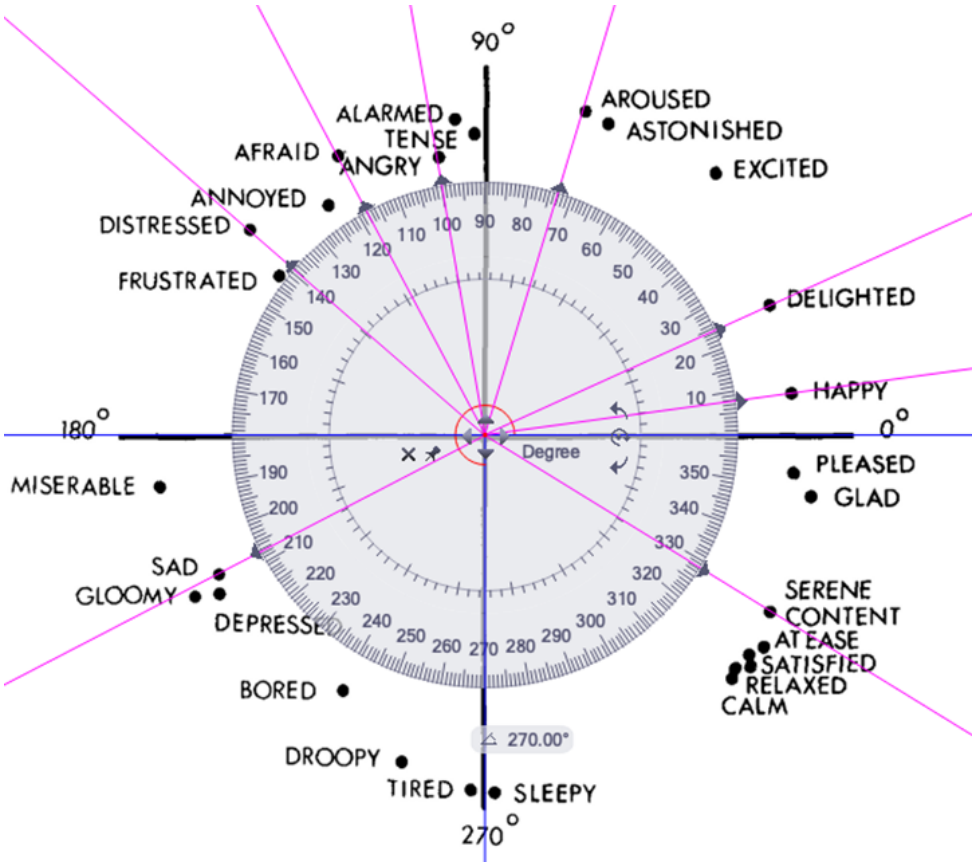


Figure 3.6: Mapping of iMotions' seven basic emotions into Russell's Circumplex Model, with angular positions assigned to each emotion.

- x_i = valence contribution,
- y_i = arousal contribution.

This transformation generated seven emotion vectors per frame of the video recording.

Emotion	Angle (°)
Anger	99.16°
Fear	117.73°
Disgust	138.70°
Sadness	207.70°
Contempt	328.49°
Joy	24.50°
Surprise	73.08°

Table 3.1: Emotion angles on Russell's Circumplex Model

At each timestamp, we then performed a vectorial average of the seven emotion vectors to obtain a single resultant vector. This resultant represented the net valence–arousal position of the participant's face at that time, combining both intensity and direction of multiple emotions into one interpretable affective state.

Because iMotions intensities are unbounded in relation to the circumplex unit circle, some resultant vectors exceeded a magnitude of 1. In these cases, we normalised the

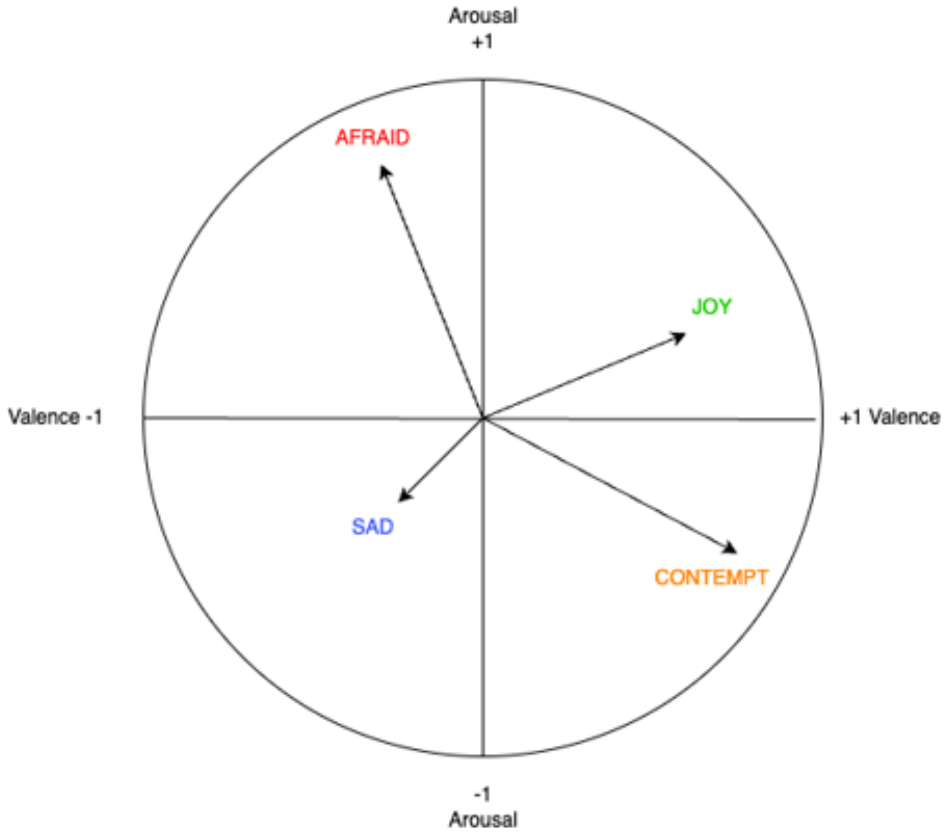


Figure 3.7: Conversion of emotion magnitudes into Cartesian vectors, followed by vectorial averaging across the seven basic emotions at each timestamp.

vector back to unit length, preserving its direction but constraining it within the circumplex. This ensured comparability across participants and stimuli.

After producing time-resolved valence–arousal coordinates for each frame, we computed a second vectorial average across all timestamps within a given audiovisual clip. This yielded one aggregate valence–arousal vector per stimulus per participant, representing the overall facial emotional response to that clip.

From the time series of valence–arousal values, we also extracted descriptive statistical features (mean, minimum, maximum, standard deviation, and kurtosis). These features summarised the dynamics of participants' facial responses and were later used for multimodal analysis.

Finally, to assess the validity of this mapping, we compared FER-derived valence-arousal measures with self-reported ratings obtained through the INDSCAL method (Section 3.1.6). Pearson correlation coefficients quantified the degree of alignment between facially expressed emotions and participants' subjective experiences, thereby linking observable behaviour with internal affective states. Find the clear flowchart of the process in the Figure 3.9.

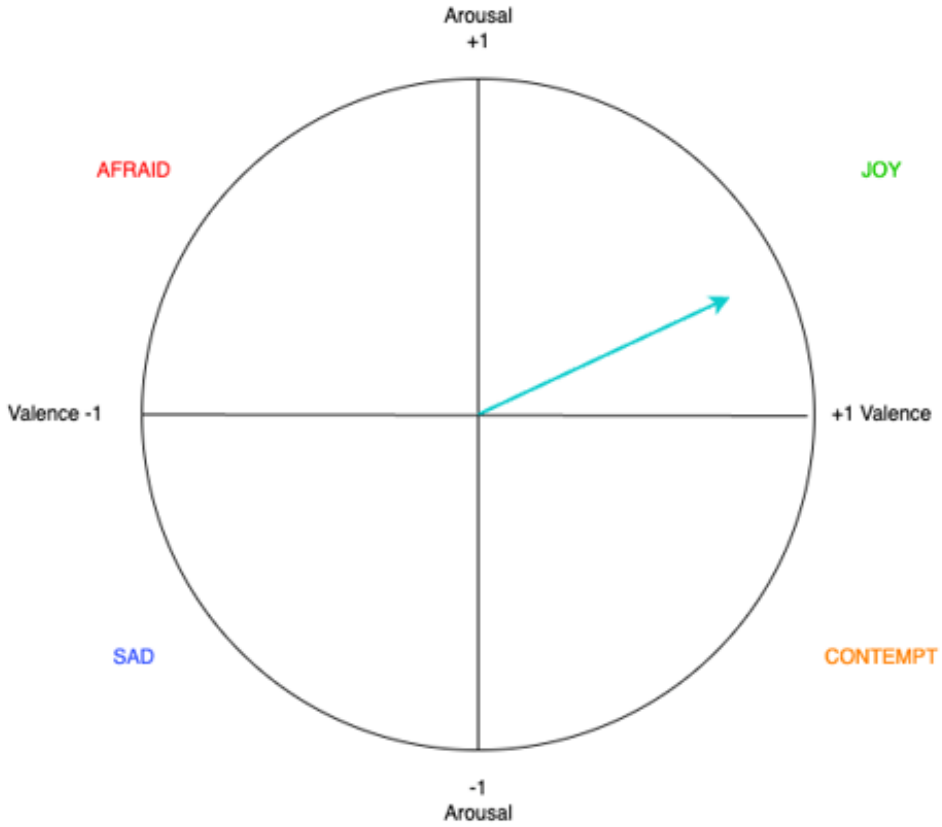


Figure 3.8: Resultant emotional vector after vectorial averaging across timestamps for a given stimulus.

Pupil Size Analysis

We utilised pupil diameter as a key feature for analysis from the eye-tracker data recorded through iMotions software, which provides information on pupil diameter, gaze points, fixation points, and saccades. We mainly used pupil diameter (pupil size) data from this information in the emotion detection model. Specifically, we calculated the average pupil diameter across the right and left eyes to obtain a more stable measurement. Using these pupil diameter values, we extracted various statistical features, including the mean, minimum, maximum, standard deviation, skewness, and kurtosis, to capture fluctuations in pupil size over time.

We normalised these features using a 60-second grey screen baseline pupil diameter measured while participants watched a grey screen to ensure consistency and account for individual differences. This baseline normalisation reduced variability caused by lighting conditions [205], individual physiological differences, and spontaneous pupil fluctuations, allowing for a more reliable comparison across participants and stimuli.

Using these extracted features, we found the Pearson correlation of these features before and after baseline normalisation with self-reported arousal and valence to gain insights into the significance of each feature in predicting emotional states.

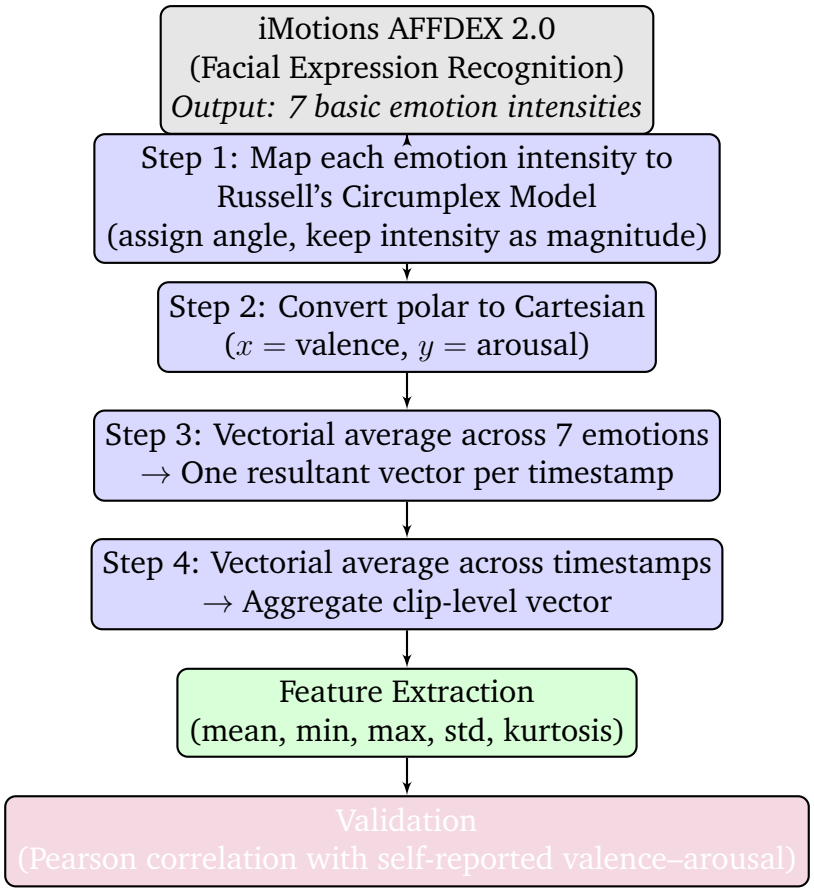


Figure 3.9: FER Processing Pipeline. iMotions AFFDEX 2.0 was only used to output seven basic emotion intensities (grey block). All subsequent stages—including circumplex mapping, vectorial averaging, feature extraction, and validation—were developed in this research (blue, green, and purple blocks).

GSR Analysis

The GSR signal consists of two primary components: the tonic (slowly changing baseline) and phasic (rapid changes in response to stimuli). This document outlines the key feature extraction techniques applied to raw GSR, phasic, and tonic signals using data from iMotions software (https://imotions.com/).

iMotions employs a multi-stage signal processing methodology to decompose raw GSR signals into tonic and phasic components, combining established physiological principles with modern optimisation techniques. The process begins with a median filter (8,000 ms window) applied to the raw signal to isolate the tonic component (SCL), representing slow baseline variations [383], [384]. The phasic component (SCRs) is then derived through convex optimisation, where the raw signal is modelled as the sum of tonic (low-frequency) and phasic (rapid fluctuations) elements, constrained by physiological priors about SCR dynamics [384], [385]. A low-pass Butterworth filter (5Hz cutoff) removes high-frequency noise before peak detection [383]. Event-related SCR identification uses amplitude thresholds (>0.01 μS for onsets, >0.005 μS peak

amplitude) and temporal constraints (minimum 500ms between onset-offset pairs) to distinguish accurate sympathetic responses from distortions [383], [386]. This decomposition aligns with a few studies showing strong phasic-arousal correlations ($p<0.01$) in IAPS emotional stimuli experiments [384]. The pipeline is implemented via R Notebook integrations that automate signal processing while allowing parameter customisation for specific research needs [383].

We first conducted statistical analyses to analyse GSR data to understand its behaviour, particularly the differences between emotionally aroused and non-aroused data. We also examined the significance of variations between high-arousal and low-arousal stimulus responses. These are described in the following section.

Comparing Stimuli and Baseline (60-second Grey Screen). For statistical analysis, we focused on average peak amplitude (APA) as the primary measure of emotional arousal. The average peak amplitude in GSR is calculated by identifying the peaks in the phasic GSR signal, measuring the difference between each peak and its baseline, and then averaging these values over a given period. This measure is essential because it reflects the intensity of emotional arousal, with larger peak amplitudes indicating stronger emotional responses.

GSR signals were continuously recorded throughout the experimental task, and individual peaks in the phasic response were identified for analysis. Peak amplitudes were identified within a defined time window corresponding to stimulus presentation, and the average peak amplitude was computed separately for each participant, which ensured that individual variations in physiological responses were accounted for.

To assess the differences between emotionally aroused and non-aroused data, we first performed Wilcoxon signed-rank tests. In this analysis, we compared the APA aggregated across all participants for each stimulus to the APA of the grey screen stimulus, which we refer to as the baseline stimulus. Based on this analysis, we used the baseline APA to normalise the APA of emotional stimuli for each participant, as follows:

$$\text{Normalised APA} = \frac{\text{APA}_{\text{emotional stimulus}} - \text{APA}_{\text{baseline}}}{\text{APA}_{\text{baseline}}} \tag{3.6}$$

This normalisation allowed a more accurate comparison of participants' emotional responses to stimuli. The normalised APA from different stimuli was then used for further statistical analysis.

Post hoc Friedman Analysis: High vs. Low Arousal Groups. After performing the Wilcoxon signed-rank tests to assess the differences between emotional stimuli and the baseline, we conducted a post hoc Friedman analysis to investigate the differences in GSR responses across two distinct arousal groups: high and low. Grouping was based on

participants' self-reported arousal ratings, where values greater than 0.5 were classified as high arousal (indicating stronger emotional responses), and values less than or equal to 0.5 were classified as low arousal (indicating lower emotional intensity).

For each stimulus, we extracted the average peak amplitude (APA) of the GSR signal across timestamps, since peak amplitude provides a more sensitive indicator of arousal than mean levels, which may smooth out transient but meaningful emotional responses. We then assumed that GSR reactivity would align with participants' self-reported arousal and divided the physiological responses accordingly.

The Friedman test was applied to compare the APA values across different emotional stimuli within each arousal group. This analysis highlighted clear distinctions in APA responses between high and low arousal conditions, providing insights into how subjective arousal levels correspond to physiological activation.

All statistical analyses were conducted in Python, with a significance threshold of $p < 0.05$.

Finally, we proceeded to the feature extraction and analysis phase, where we identified features suitable for training the emotion detection model and examined their reliability by assessing their relationship with self-reported valence and arousal.

GSR Feature Extraction. We used phasic and tonic data provided by iMotions to analyse GSR responses and extracted key statistical features.

We computed both components' mean, minimum, maximum, standard deviation, kurtosis, variance, skewness, mean energy, peak per minute, peak average, and average peak amplitude. To ensure consistency across participants, we applied a 60-second baseline correction, similar to the normalisation process used for APA mentioned in Equation 3.6.

Finally, we integrated all the extracted features with and without normalisation and calculated the Pearson correlation of each feature. We did this to assess the impact of normalisation on feature relationships and ensure that individual differences or variations in scale did not bias the results. First, correlations were computed individually for each participant within their individual space. The results of these analyses are presented in Section 4.

Following the pilot study and other literature review, we recognised the importance of including emotionally healthy participants to establish a reliable baseline of emotional responses that would support a more accurate interpretation of physiological patterns associated with typical emotion regulation. The pilot study also provided valuable insights into the strengths and limitations of our data collection, analysis and feature extraction methods. For example, we found that correcting for pupil size using a grey screen baseline was ineffective. Instead, it is crucial to isolate and remove the effect of luminance to accurately capture the emotional influence on pupil size, and

the development of a method to do that became a significant part of this project (see 3.2.6 and 4.1.3). In addition, we found that relying solely on basic statistical features of the GSR was insufficient, as these features often fail to capture the temporal dynamics and phasic responses that are crucial for detecting short-term emotional arousal. More advanced features, such as peak detection, response latency and signal decomposition, are needed to better represent the nuanced physiological changes associated with emotional states. Further discussion is presented in the results and discussion chapter.

Correlation and Mutual Information Analysis

To quantify the relationship between physiological and facial expression features and participants' self-reported emotional states, we computed three complementary metrics: Pearson correlation, Spearman rank correlation, and mutual information. Pearson correlation was used to assess linear associations, while Spearman correlation captured monotonic relationships independent of linearity. Mutual information was employed to evaluate potential non-linear dependencies between features and emotional targets (Valence and Arousal). For each participant, correlations were computed between individual features and the corresponding emotional ratings, and the results were aggregated across participants to calculate the mean and standard deviation. Heatmaps were generated separately for each metric to provide a visual summary of the relationships across modalities (FER, GSR, Pupil) and emotional targets. These allowed us to refine and enhance our approach, which we leveraged as advantages in the main study. The improvements made are discussed in the next section.

3.2 Main Study

The pilot study provided important insights into the complexities of emotion detection using physiological signals and facial expressions. It highlighted the limitations of facial expressions as stand-alone biomarkers, the significant influence of luminosity on pupil size measurements, and the importance of integrating multiple biomarkers to improve detection accuracy. It also revealed technical and methodological challenges, such as the need for calibration processes to remove the effect of luminosity from pupil size, the necessity of more advanced feature extraction from GSR signals, and the importance of participant screening to ensure consistent physiological baselines. As previous research has shown, underlying mental health conditions such as anxiety, depression, or alexithymia can significantly alter autonomic and emotional responses, introducing variability that can confound interpretation of physiological signals [387], [388].

Building on these findings, the main study incorporates several methodological refinements that constitute novel contributions to multimodal emotion detection. First,

participant screening was systematically implemented to exclude individuals with depression, anxiety, personality disorders, and alexithymia, thereby minimising confounds associated with clinical heterogeneity, an aspect often overlooked in affective computing studies. Second, we developed a new *luminosity isolation model* for pupil size, introducing a calibration procedure based on three primary colours and grayscale references. This is, to our knowledge, the first systematic attempt to correct for lighting effects in pupil-based emotion research, enabling more accurate measurement of pupil-driven emotional responses. Third, our GSR processing pipeline was expanded to include time-, frequency-, and time-frequency-domain features, moving beyond basic statistical descriptors and capturing a richer set of autonomic dynamics. Finally, advanced machine learning models were trained on physiological signals both individually and in integrated multimodal configurations, enabling direct assessment of the benefits of feature-level fusion.

Although EEG data were also collected in the main study, their analysis is beyond the scope of the present work. They will be explored in future research to extend the multimodal framework and further improve the robustness of emotion detection.

Taken together, these methodological advances represent a novel integration of corrected pupil size, enhanced GSR features, and rigorous participant screening into a unified multimodal system, setting a new benchmark for interpretable and clinically informed emotion-aware AI.

3.2.1 Participants

Initially, we screened 103 participants for this study. To recruit participants, we distributed pamphlets, and upon receiving inquiries via email, we provided a screening survey. As part of the screening process, all participants underwent a psychological assessment using four standardised scales to evaluate the potential mental health conditions, such as alexithymia, depression, anxiety, and PDs [387], [388], and the selection criteria of the study.

The purpose of the screening process was to identify participants eligible to take part in the study. The screening survey collected basic demographic information and included psychological scales, applying the following inclusion criteria:

- Age between 18 and 65 years.
- Sufficient proficiency in English to provide informed consent and complete questionnaires.
- No recent neurological disorders (e.g., epilepsy, stroke, traumatic brain injury, or ASD within the past year).

Applying these criteria led to the exclusion of several individuals from the initial pool of 103 screened participants. The remaining participants were then further evaluated

based on their scores on the psychological scale, as detailed in the following section.

Screening Process and Psychological Scales

After applying the inclusion criteria, 55 of the 103 screened participants were retained for the main experiment. These individuals completed all screening procedures and were not at high risk of psychiatric conditions known to affect emotion recognition, such as schizophrenia, anxiety etc.. The selection criteria were explicitly designed to exclude individuals with significant difficulties in emotion recognition and perception, as such impairments could compromise the reliability of physiological and behavioural responses. This resulted in a carefully curated cohort, referred to here as the `Emotionally Healthy` group, representing participants with typical emotional processing abilities and thereby enabling more accurate modelling of normative emotion–physiology associations.

In addition, data were collected from 17 participants identified as being at risk for anxiety, depression, or personality disorders, forming a separate `Clinical` group. Including this group not only increases the ecological validity of our dataset but also provides a unique opportunity to explore how deviations in psychological health may alter emotion recognition and physiological responses.

To the best of our knowledge, very few affective computing studies systematically screen participants for psychiatric risk factors before model training. Our approach is novel in that it explicitly integrates psychological assessment into the data collection pipeline, reducing confounding variance and creating two parallel datasets (`Emotionally Healthy` vs. `Clinical`). This design provides both a cleaner baseline for developing robust multimodal emotion detection models and a pathway toward future investigation of how clinical conditions modulate emotional processing. To assess emotion recognition ability and potential risk factors, the following standardised psychological assessments were administered:

- **Toronto Alexithymia Scale (TAS-20)** – A 20-item self-report questionnaire that measures difficulties in identifying and describing emotions and externally oriented thinking. Scores range from 20 to 100, with higher scores indicating greater alexithymia [347]. The cut-off scores are:
  - $\leq$ **51** – Non-alexithymic
  - **52–60** – Borderline alexithymia
  - $\geq$ **61** – High alexithymia
- **Patient Health Questionnaire (PHQ-9)** – A 9-item scale used to assess depressive symptoms based on DSM-IV criteria [348]. The total score ranges from 0 to 27, with severity levels categorised as:
  - **0–4** – Minimal or no depression

  - **5–9** – Mild depression
  - **10–14** – Moderate depression
  - **15–19** – Moderately severe depression
  - **20–27** – Severe depression
- **Generalised Anxiety Disorder Scale (GAD-7)** – A 7-item questionnaire designed to assess the severity of generalised anxiety symptoms [349]. Scores range from 0 to 21, categorised as follows:
  - **0–4** – Minimal anxiety
  - **5–9** – Mild anxiety
  - **10–14** – Moderate anxiety
  - **15–21** – Severe anxiety
- **International Personality Disorder Examination (IPDE)** – A diagnostic tool for screening PDs based on DSM-IV and ICD-10 criteria [350]. It assesses traits related to various PDs, including:
  - Obsessive-Compulsive Personality Disorder (OCPD)
  - Schizoid Personality Disorder
  - Paranoid Personality Disorder
  - Histrionic Personality Disorder
  - Borderline Personality Disorder (BPD)
  - Dependent Personality Disorder

  Each disorder is assessed through a structured questionnaire, with scoring thresholds indicating the likelihood of a clinical diagnosis.

Participants were classified based on their results:

- **55 participants** (26 males, 29 females) participants were selected as `Emotionally Healthy Group`. They had no severe psychiatric diagnoses affecting emotional recognition (e.g., schizophrenia, bipolar disorder, or autism spectrum disorder), though mild anxiety, depression, or alexithymia was permitted. Specifically, they did not meet the criteria for high alexithymia (TAS-20 $\leq$ 61), moderate to severe anxiety (GAD-7 $\leq$ 5), or moderate to severe depression (PHQ-9 $\leq$ 5). Participants had no diagnosed intellectual disability or significant cognitive impairment, no recent neurological disorders (e.g., epilepsy, stroke, ASD, or TBI in the past year), and were proficient in the language of the questionnaires. They provided informed consent and were able to complete all required questionnaires and bio-signal recording tasks.
- **17 participants** were recorded under the `Clinical Group`, as they met the criteria for the risk of having psychiatric conditions such as OCPD, schizoid personality disorder, generalised anxiety disorder, depression, and alexithymia.

Among the Clinical Group, several participants exhibited the risk of comorbid conditions, with the most frequent being alexithymia, obsessive-compulsive PD, schizoid PD,

anxiety, and depression. A small subset also had the risk of PDs such as histrionic, paranoid, and borderline PD. This classification allowed us to conduct preliminary analyses on how these conditions influenced emotional recognition, perception, and physiological responses to emotions.

After selecting participants from each group, we invited them to the laboratory to perform the experiment with their consent and ethical approval.

Ethical Considerations

All participants provided `informed consent`. Under application numbers **ETH2324-0962**, **ETH2324-1491**, **ETH2425-0176**, where the ethical approval from the pilot study were updated by modifying the changes, granted in compliance with institutional research guidelines.

To ensure diverse representation, participants aged 18 to 65 were recruited. The study adhered to the ethical principles of voluntary participation, confidentiality, and the right to withdraw at any stage without consequences. Following ethical considerations, we performed the final experiment described in the following sections.

3.2.2 Experimental Setup

We used the same experiment setup as in the pilot study section 3.1.4. We conducted all the experiments in the well-lit laboratory at the University of Essex, using the same high-resolution camera, Tobii Pro Nano eye-tracker, GSR, and PPG sensors from Shimmer 3.10.

Additionally, EEG data were recorded using the g. Nautilus system with a 32-channel cap, providing standard scalp coverage and high-density monitoring of brain activity. Electrode impedances were maintained within acceptable ranges to ensure reliable signal quality. Although EEG offers rich information about neural responses to emotional stimuli, it was not included in the subsequent analysis. The primary aim of this study was to develop a practical and easily deployable emotion detection model using signals that are simple to collect, such as facial recordings, eye-tracking, and GSR. This design prioritises ease of use for clinical or therapeutic applications, enabling a less complex yet accurate system suitable for real-world deployment.

This comprehensive approach combined physiological signals, FER, and EEG data to provide a rich dataset for understanding participants' emotional reactions.

3.2.3 Procedure

Data was collected using iMotions Software [379], which enabled the presentation of stimuli and synchronised data recording across various sensors. The process began

Figure 3.10: Experiment set-up.

with calibrating the Tobii Pro Nano eye-tracker to accurately track participants' pupil size and gaze position. This calibration was crucial for obtaining reliable and precise data throughout the experiment, as it ensured the eye-tracker was aligned with the participants' eyes.

Once the calibration was complete, participants saw an instruction screen that provided an overview of the experimental procedure and what to expect. They then listened to a meditation audio track while viewing a black screen, designed to help them relax and establish a baseline for low-arousal data. To verify their relaxation, a questionnaire slide was presented, asking them to rate their level of relaxation after the meditation.

Following this, participants viewed a 27-point calibration video designed explicitly

for pupil size analysis (more details on pupil size analysis can be found in the relevant section 3.2.6). This was succeeded by a series of monochrome and multicoloured, emotionally neutral test images to further evaluate pupil size responses. After the calibration phase, the main experiment began (see Figure 3.11). Given the experiment's length, a short break was provided midway, allowing participants to rest before continuing. After the break, participants watched another 27-point calibration video, identical to the one shown at the start, before resuming the remaining audiovisual clips from the main experiment.

The experiment's first phase randomly presented 32 audiovisual clips to each participant [2]. As highlighted in the literature, self-assessment is influenced by various factors, including cognitive bias. Our approach aims to develop a multimodal model that takes into account self-assessment and cognitive bias. Instead of standardising the clips, we selected them based on insights from previous studies and a preliminary survey on well-known clips capable of eliciting the desired emotions. Cognitive biases were considered during the emotional labelling to enhance the model's robustness. Compared to the pilot study, this study included additional clips and replaced a few that were assessed by participants in contradictory ways.

The selected clips were carefully chosen to represent four distinct emotional categories: high arousal with positive valence, high arousal with negative valence, low arousal with positive valence, and low arousal with negative valence. Each category contained eight audiovisual clips designed to evoke diverse emotional states and elicit physiological and subjective responses [32]. A 5-second grey screen featuring a black cross in the centre was displayed between audiovisual clips to reduce emotional carry over between audiovisual clips. This neutral stimulus aimed to reset the participants' emotional state before introducing the next emotional stimulus. In a pilot study, we used a cloud video that appeared to elicit low arousal and mildly positive valence, consistent with other findings in the literature. We initially selected such stimuli based on prior studies that had categorised them as neutral. However, based on our pilot observations and participants' responses, we recognised this potential bias. Therefore, in the main study, we removed the cloud videos and replaced them with a uniformly grey screen to minimise unintended emotional influence. After watching each audiovisual clip, participants were asked to complete a survey evaluating their emotional reactions and familiarity with the clips. As outlined in the literature review chapter (2), we employed the multidimensional emotional questionnaire, which consisted of 12 questions related to emotions extended from [164] (see Figure 3.12). These questions evaluated emotions such as **positivity, excitement, happiness, amusement, calmness, contentment, negativity, anger, fear, disgust, sadness, and boredom** (more inform-

[2]Link to audiovisual clips: `https://essexuniversity.box.com/s/7bquteyzf5qeagud8uzat2fglntmgu3f`

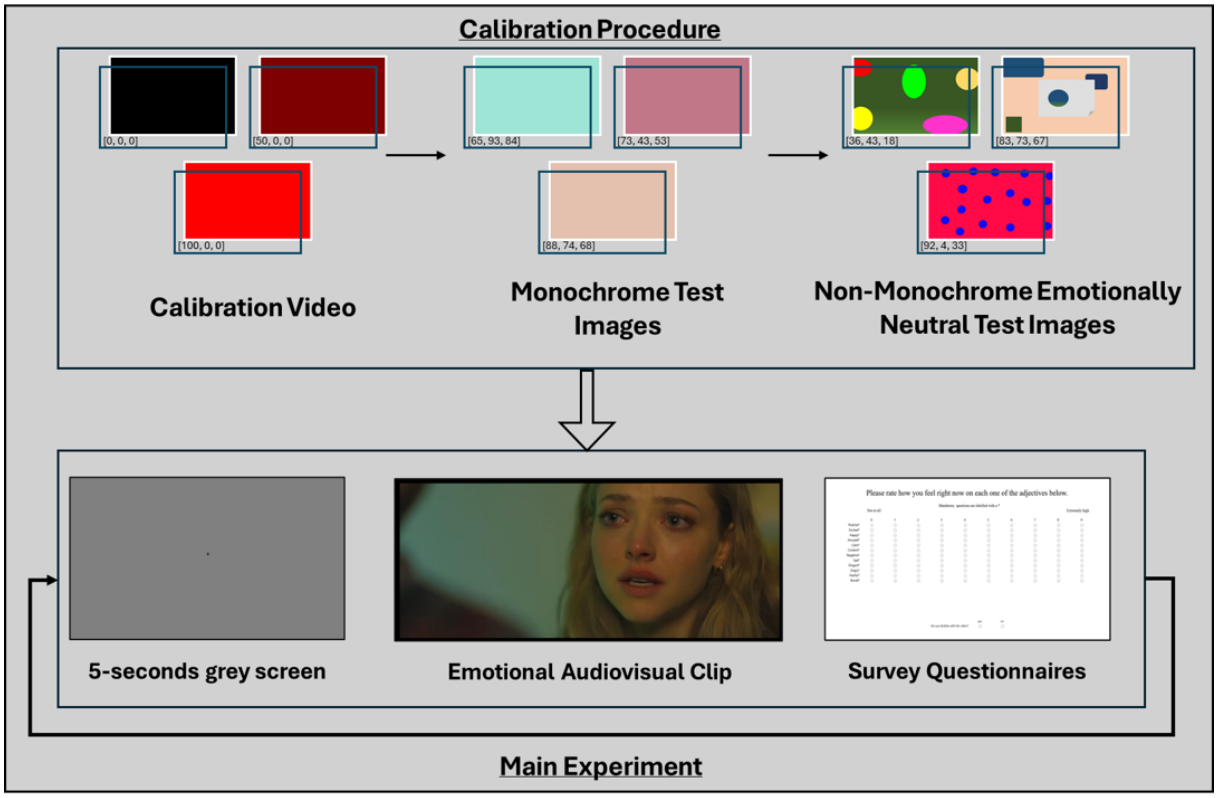


Figure 3.11: Experiment Presentation flow chart.

ation about survey questionnaire selection is in the pilot study experiment procedure section 3.1.5) Each of the 12 emotional questions was rated on a scale from 0 to 9, with 0 signifying no emotion and 9 indicating the highest level of emotion (see Figure 3.12). This scale allowed for a comprehensive assessment of the emotional intensity felt by participants while watching each audiovisual clip. Integrating physiological measurements and self-reported emotional responses created a substantial dataset for examining emotional reactions among participants.

Criteria to discard participants after the experiments. In addition to the initial screening, participants were excluded for several reasons following the experiment. Some exhibited poor Tobii calibration, which compromised the accuracy of gaze and pupil size measurements. Others demonstrated insufficient engagement, such as consistently failing to look at the screen or falling asleep during the task. Additionally, some participants provided unreliable responses, such as giving identical ratings to all questions or providing random responses. Additionally, some individuals appeared disengaged during the experiment, such as not looking at the screen or being distracted by their surroundings. We also observed low inter-rater reliability, with Spearman's correlation coefficients below 0.4, indicating significant inconsistency in participants' responses. This correlation, calculated as described in the following section, helped us to identify and address these issues. As discussed in the literature review, survey ques-

Figure 3.12: Survey Questionnaires.

tionnaires can lack reliability, complicating the assessment of participant attentiveness. It was essential to remove those participants with stated issues.

While it was relatively easy to identify participants who provided identical responses throughout or were visibly inattentive, detecting those who answered insincerely or selected responses randomly was more challenging. To address this, we applied a few inter-rater reliability assessment techniques, including Krippendorff's alpha, Cronbach's alpha, and Spearman's correlation, with the help of a collaborating student. Among these methods, Krippendorff's alpha and Spearman's correlation gave similar results, but Spearman's correlation required significantly less computation. We therefore used Spearman's correlation to measure the consistency of one participant's responses with those of others.

To compute **Spearman's correlation**, we used the survey questionnaire responses from each participant for each audiovisual clip after discarding the noticeably unreliable participants, as described earlier. Each questionnaire consisted of **12 emotion-related questions**, and we calculated the correlation between each participant's responses across all audiovisual clips for each emotion separately.

Let:

- M be the total number of participants
- K be the total number of clips
- $E = 12$ be the number of emotions

For each emotion, we computed an $M \times M$ **correlation matrix** representing the Spearman correlation coefficients between all participants for all clips. This resulted in a total of **12 correlation matrices** (one for each emotion), denoted as:

$$C_e \in \mathbb{R}^{M \times M}, \quad e = 1, 2, \ldots, 12$$

Next, we stacked these 12 matrices and computed the average across all emotions to obtain a single overall correlation matrix:

$$C = \frac{1}{12} \sum_{e=1}^{12} C_e, \quad C \in \mathbb{R}^{M \times M}$$

To determine each participant's reliability, we computed the average correlation of each participant with all other participants, resulting in a final $M \times 1$ correlation vector:

$$C_{\text{final}}(i) = \frac{1}{M-1} \sum_{j=1, j \neq i}^{M} C(i, j), \quad \forall i \in \{1, \ldots, M\}$$

If any participant's average correlation value was **below 40%** (i.e., $C_{\text{final}}(i) < 0.40$), we removed that participant and recomputed the correlations iteratively until all remaining participants had correlation values above the threshold.

This iterative process ensured that only participants with consistent and meaningful responses were retained for further analysis.

After identifying participants with low correlation scores, we reviewed their experimental recordings for signs of drowsiness or inattention, including prolonged eye closures, frequent blinking, and consistently low pupil activity. Participants exhibiting both low correlation and sustained indicators of drowsiness or inattention were removed from the dataset.

Ultimately, we retained 47 participants (26 males, 21 females) (Average age: 27.79 ± 10.36), data in the `Emotionally Healthy Group`. In the `Clinical Group`, we retained 14 participants (9 males, 5 females) (Average age: 26 ± 4.47), data.

All participants were asked to complete survey questionnaires after each video clip to assess their emotional and arousal responses, without being prompted to identify specific moments within the video. We conducted an additional brief survey to better identify emotionally salient intervals that could inform our analysis.

3.2.4 Identifying Emotionally Salient Intervals

Based on initial survey responses, it became evident that different audiovisual clips elicited varying levels of emotional arousal across participants. A major challenge in analysing these responses was the variability in the proportion of emotional content across clips: some maintained a uniform emotional tone, while others featured emotional peaks only in specific segments. For example, a scary clip might begin with a calm or neutral tone and build up to a sudden, intense climax. In such cases, one would reasonably expect a corresponding physiological reaction, such as a noticeable pupil dilation, only at the climax rather than throughout the entire clip.

Given these natural variations in emotional pacing and arousal levels, we deliberately chose not to strictly control all variables. The goal was to develop an emotion detection system that could be robust enough to be used in real-world settings, where emotional stimuli are rarely uniform and fluctuate dynamically. Additionally, the clips varied in length (ranging from 10 to 100 seconds) and emotional pacing. Some were monotonously boring, others consistently engaging, and some presented emotional highs at particular moments. This variability made it difficult to associate a single arousal label with the entire clip, prompting the need for a more granular analysis.

To better understand these dynamic emotional transitions and to capture changes in arousal over time, we conducted a targeted follow-up study called the Emotion-Induced Interval Study (EIIS). Ten participants from the original group of 55 Emotionally Healthy participants were selected for consistency. In this study, participants were asked to rewatch the complete set of audiovisual clips and mark specific time intervals during which they experienced noticeable changes in arousal, whether increases (high

arousal) or decreases (low arousal). These user-defined periods were termed Emotion-Induced Intervals (EII).

The goal was to capture the temporal dynamics of emotional arousal and allow participants to break down their emotional experiences into finer segments within each clip. This helped reveal fluctuations in arousal that might otherwise be lost in a single post-clip rating. Analysis of the EIIS data confirmed that many audiovisual clips exhibited clear internal emotional transitions.

These observed transitions allowed us to divide certain audiovisual clips into two distinct segments, low-arousal and high-arousal portions, based on the reported EIIs. As a result, six clips were split, increasing the total number of analysable segments from 32 to 38. Each segment could be treated as a separate entity, better aligned with its evoked emotional response. This segmentation was particularly beneficial for analysing physiological signals that respond rapidly and spontaneously to stimuli, such as pupil size.

We identified salient intervals for each clip to determine which time intervals most represented emotionally salient moments. A salient interval was defined as the period with the highest level of agreement among the ten participants. On average, the overlap between the participant-indicated intervals and the final selected interval was 74% ± 15%, indicating a strong level of consensus. For emotionally flat clips (e.g., boring), the entire duration was often chosen as the salient interval. In contrast, emotionally intense clips (e.g., scary or suspenseful) typically had shorter, well-defined salient segments.

This interval-based approach also helped address a limitation in the main study. As noted in the Procedure 3.2.3, participants were asked to report their emotional state at the end of each clip but were not required to specify when they felt those emotions during the clip. Therefore, using the EIIS framework allowed us to retrospectively annotate and segment clips based on emotionally significant periods, improving the precision of arousal labelling and subsequent model training.

The outcome of this segmentation process forms the basis for more accurate emotion labelling, which is discussed in the following section. By associating physiological signals with specific, participant-validated emotional intervals, we enhanced the reliability of emotion detection and built models that better reflected the temporal nature of emotional experience.

3.2.5 Emotion Labelling

The data labelling process followed the same approach described in the pilot study's Data Labelling section (3.1.6). Audiovisual clips were categorised into emotional quadrants according to arousal (high/low) and valence (positive/negative). For instance, a high-arousal positive clip was labelled as HPclip number, where "HP" indicates the

category and the number identifies the specific clip within that category.

We used INDSCAL [164], a multidimensional scaling method that accounts for individual differences, to project the 12-dimensional (12D) self-report ratings (12 emotional questions per clip) into a two-dimensional (2D) valence–arousal space. In addition to INDSCAL, we also evaluated Factor Analysis (FA) as an alternative dimensionality reduction technique for mapping the 12-dimensional emotional questionnaire responses into a two-dimensional valence-arousal space. For both methods, we first computed pairwise dissimilarity matrices using Euclidean distance between stimuli. **INDSCAL was implemented in SPSS to generate a group configuration and individual subject weights, while FA was applied using standard factor extraction methods. The resulting two-dimensional representations from both methods were subsequently compared using clustering quality metrics (silhouette score, Davies-Bouldin index) and within-stimulus variance, as detailed in Section 4.2.1.**

The INDSCAL approach, aligned with Russell's Circumplex Model of Affect, produced a balanced mapping of the clips across the four quadrants of the affective space. In constructing the similarity matrices required for INDSCAL, we explored multiple distance metrics to measure dissimilarities between stimuli. Both Euclidean and Manhattan distances were tested, with Euclidean ultimately selected as it produced the most consistent and interpretable results in the 2D valence–arousal space. For completeness, a comparison of Euclidean and Manhattan outcomes is presented in the Appendix 6.1.1.

Initially, 32 audiovisual clips were included in the study. However, after applying the EIIS method (3.2.4), several clips were split into two segments when they exhibited distinct arousal shifts (e.g., transitioning from low to high arousal within the same clip). This segmentation increased the total number of labelled stimuli to 38. While the second segment of these split clips did not always have a neutral baseline, we retained them because the goal was to capture naturalistic emotional transitions. These transitions are important for analysing how physiological signals respond dynamically to shifts in arousal, which better reflects real-world emotional experiences.

As a result, the final dataset contained $(N = 38)$ stimuli, each represented in a (38×38) distance matrix constructed from the 12D self-reports ($(M = 12)$ scales). The INDSCAL projection then yielded individualised and group-level coordinates in the 2D space, with the x-axis interpreted as valence and the y-axis as arousal. These derived labels formed the ground truth for subsequent model training and analysis.

3.2.6 Data Pre-processing and Analysis

Before analysing the data, we took several steps to ensure its quality and reliability. This involved cleaning and normalising the data to remove any noise and outliers, and converting it into a format suitable for analysis. Once the pre-processing was complete,

we used various analytical techniques to explore the relationships between physiological signals and emotional responses, measured in valence and arousal for each participant and across the group. This process included extracting key features from the data, applying mathematical models, and conducting correlations to reveal patterns and insights essential for emotion detection.

Pupil Size Analysis

We developed a non-linear exponential model to predict pupil size as a function of luminosity, aiming to separate the confounding effects of ambient light from pupil responses to emotional arousal. The study progressed through the following steps:

1. We first validated the exponential model introduced in [109] under controlled dark-laboratory conditions with 10 participants.
2. Next, we evaluated the model's robustness in a well-lit laboratory, testing its ability to generalise across varying luminosity levels.
3. We then quantified the pupil size attributable to emotional arousal by subtracting the predicted luminosity-driven component from the total pupil size. For this, participants viewed 32 video clips with varying emotional content and reported their arousal level for each.
4. Finally, we compared the exponential model against hyperbolic and linear alternatives, demonstrating that the exponential model provided a superior fit for describing the relationship between pupil size and luminosity.

This approach effectively removes luminosity-induced co-founders in pupil-based emotion research, particularly in low-light conditions, filling a significant gap in existing methodologies.

As discussed in the literature, pupil size reflects emotional arousal but is strongly influenced by ambient luminosity [268]. Because this relationship is inherently non-linear [265], simple corrections are insufficient. Our exponential modelling framework enables accurate separation of emotional and luminosity effects [109].

In practice, we first constructed a predictive model of pupil size as a function of luminosity. We then applied this model to remove luminosity effects from the raw pupil data, which contained both emotional and light-driven components. The resulting arousal-isolated pupil size signals were subsequently used to train a machine learning model for emotion detection, initially relying exclusively on pupil-based features.

Here, we divided this approach into two study phases:

1. Development of the Luminosity Effect Prediction Model (LEPM): In this study phase, we developed a model to predict pupil size changes purely caused by luminosity. By subtracting the predicted pupil size from the observed one, we isolated the pupillary response attributable to emotional stimuli.

2. Development of the Arousal Detection Model (ADM): To prove that the LEPM developed in the previous study phase was accurate and had ecological validity, we trained a simple linear model on participants' pupil size data while watching emotional audiovisual clips and used it to predict arousal with and without the LEPM algorithm.

Development of the Luminosity Effect Prediction Model. We developed a luminosity-based pupil size prediction model to isolate emotional pupil responses and subtracted its predictions from the observed pupil sizes.

In this study, the model predicts pupil size based on the stimulus's luminosity and subtracts this expected value from the actual pupil measurements taken during emotional stimuli. This method improves the accuracy of pupil-based emotion assessments and enhances the reliability of emotion detection systems by providing a clearer picture of emotional arousal.

The development process involved several phases:

1. The first phase is to predict the intensity of the luminosity stimulus reaching the eye as a function of screen luminosity, such as when someone is watching a video.
2. The second phase consists of determining the size of the pupil as a function of the intensity of the light stimulus.
3. The third phase consists of combining the predictive models developed in the first two steps to obtain a single model that can predict pupil size as a function of screen luminosity in an unlit laboratory using monochrome images or video frames (red, green, blue, and grey) with varying luminosities. The model was tested on monochrome and non-monochrome emotionally neutral images or video frames. Third, we tested the model in a well-lit laboratory environment.

Finally, in the second part of our study, we developed the emotion detection model and verified that it improved emotion detection. Below, we describe the three phases in more detail.

Prediction of light intensity. Each pixel of an image consists of three pixels, one red, one green, and one blue (RGB system), and is characterised by three RGB intensity values, one per colour, expressed as a percentage of the maximum possible value. Here, we will refer to a monochrome image as an image in which all pixels have only one colour. In contrast, a non-monochrome image may contain pixels of different colours. Additionally, here, by primary colours, we refer to red, green, and blue (RGB) colours. We will speak, for example, of a primary-colour monochrome image to indicate an image with all pixels only red for a red monochrome image, only green for a green monochrome image, or only blue for a blue monochrome image. Images are commonly non-monochrome and non-primary because the pixels can take on different colours and

contain colours that are not necessarily the primary RGB colours. In this study, by grey, we mean the colour resulting from combining red, green, and blue equally, e.g., 65% red, 65% green, and 65% blue.

The brightness of a monitor is a function of the RGB values of all pixels. In the case of a monochrome image, we can write:

$$L(r, g, b) = k \cdot f(r, g, b) \tag{3.7}$$

Where L is the luminosity value, k is a constant (scaling factor) depending on the brightness/contrast settings of the monitor, and $(r, g, b) \in [0\%, 100\%]$ are the RGB intensity values of a monochrome image displayed on the screen[109]. In a commercial monitor, the relationship between RGB intensity values and the luminosity values of the screen is non-linear. For example, considering a single colour, e.g., blue, the luminosity corresponding to an RGB intensity value of 60% is not twice as high as that corresponding to a value of 30%. Instead, the increase in luminosity as a function of RGB intensity is slower at lower intensity values and dramatically faster as the intensity approaches maximum [109]. Furthermore, as more colours are used, this non-linear relationship changes. Therefore, the function f cannot be easily described analytically, and we tabulated its values in a look-up table to calculate the brightness of images or frames displayed on a computer screen. The input of our look-up table was an RGB value corresponding to three integers ranging from 0 to 100. The various brightness values were measured empirically. We created a set of images with different RGB values. The luminosity of each image was measured in an unlit lab using an LX1010BS digital lux meter at a 65 cm distance, following standard eye-tracking protocols [109]. If we wanted to consider all values, we would have 100^3 entries and would have to generate 100^3 images. To make the search computationally less expensive, we sub-sampled this space, producing 1330 images of different colours and luminosities and measured with a professional lux meter. The set of images was composed as follows: the first image of the set was utterly black and each pixel had an RGB value of 0%, 0%, 0% (all colours off); the last image was white with maximum luminosity and each pixel had an RGB value of 100%, 100%, 100% (all colours at maximum intensity); all the other images had intermediate values uniformly distributed in the range [0%, 100%].

When querying the table, if the input matched a point, the corresponding brightness value was returned; otherwise, a weighted average of the eight nearest points was calculated, with weights inversely proportional to their distances. In the case of images composed of only one of the fundamental colours, instead of eight nearest neighbours, we used only two [109].

We calculated the values in the lookup table using a Dell Precision M6500 (1920x1080) monitor set to 100% brightness and 100% contrast, which we will refer to

as the "reference monitor" [109]. When the contrast and brightness values are changed, the equation 3.7 remains valid, apart from the scaling factor k, which must be recalculated each time (see calibration procedure below).

Different monitor models have different brightness capabilities. However, when testing five monitors of various sizes and brands, we consistently observed the same non-linear relationship in equation 3.7), which remained valid. In contrast, we had to recalculate the scaling factor k only for each monitor, with a maximum error of 3 lux [109]. This suggests that regardless of monitor manufacturer and model, the non-linearity in the relationship between luminosity and RGB intensity values is a common characteristic of different monitors. Therefore, the calibration procedure described later allowed us to compensate for differences between possible monitor settings and between different monitors, provided the monitor settings did not change during the experiment. For the reference monitor, the value of k was 1 [109].

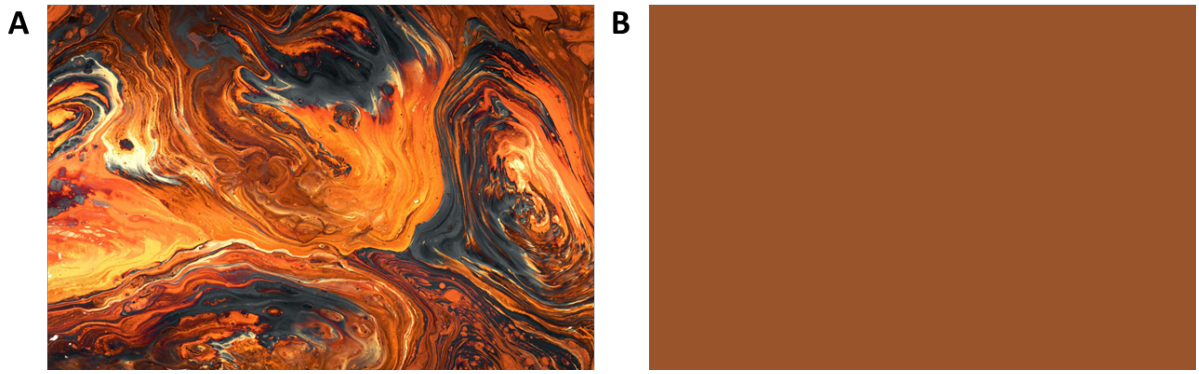


Figure 3.13: Example of original multi-colour image (A) and its corresponding monochrome image (B).

When an image or video frame is not monochrome, one can calculate the brightness of each pixel and average the values obtained. However, we observed that the luminosity of an image is similar to that of a monochrome image with the average RGB value of the original image, calculated by averaging the RGB values of all its pixels. To verify this, we selected 100 images: 50 taken from the internet and 50 generated by assigning a random RGB value to each pixel. We then generated the corresponding monochrome images using the average RGB value of each original image, as shown in Figure 3.13. The luminosity of the original images differed from the luminosity of the corresponding monochrome images by 1.5 lux on average, with a maximum difference of 3 lux, which is tolerable for our purpose. For computational reasons, using monochrome images is preferred as it is much faster than calculating the luminosity for each pixel and then averaging.

The predicted luminosity value was then used to predict pupil size, as described in the next phase below.

Determining the pupil size. The second step consisted of developing a model to predict the pupil size of a subject looking at a monochrome image on the screen. This model incorporated the function f, defined in the previous step, which transforms RGB values into luminosity. This luminosity function then served as an input to another function responsible for mapping luminosity to pupil size. Notably, this second function is highly non-linear and exhibits a decreasing trend [265].

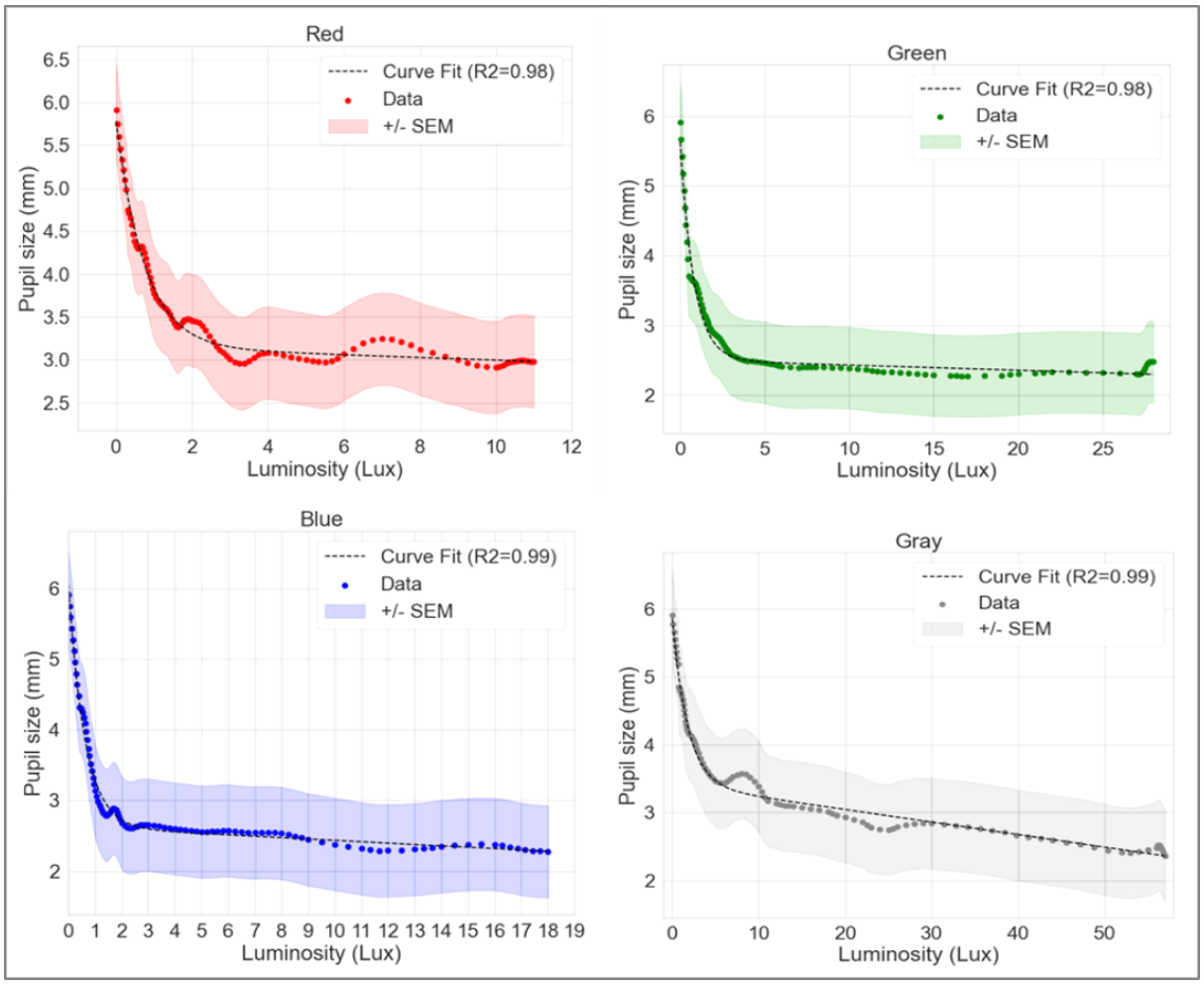


Figure 3.14: Pupil size as a function of luminosity for red, green, blue, and grey (dotted line for experimental data and continuous line for the fitted curve).

To determine the pupil size for different luminosities, we recorded the pupil size of ten participants while watching four monochrome videos, with each video only showing one fundamental colour (red, green, blue) and grey. Keeping the colour constant in each video, only luminosity could change from frame to frame. Our goal was to assess the pupil response either to one of the fundamental wavelengths (red, green, blue) or to their uniform combination (grey). We started from the lowest intensity (0%, 0%, 0%) to the highest intensity (e.g., for red 100%, 0%, 0%, green 0%, 100%, 0%, blue 0%, 0%, 100%, and for grey 100%, 100%, 100%), increasing the brightness by 1% every second. Each of the four videos lasted 101 seconds and was presented on a full screen on the

reference monitor (1920x1080). We measured pupil size for each frame at a distance of 65 cm between the screen and the participant using a Tobii Pro Nano eye tracker in an unlit laboratory. At the same time, we measured luminosity during each video using a lux meter [109]. For frames displaying an eye blink (indicated as -1 in the dataset), the pupil size was substituted with the median pupil size value for that particular frame. Subsequently, we calculated the average pupil size data across all participants for each level of luminosity. Figure 3.14 displays the average pupil size (aggregated across participants) plotted against luminosity for each colour (scattered points). The dotted lines represent the fitted models. As is evident, pupil size diminished exponentially with rising luminosity [109], [265]. It does not take 101 measurements to fit a model with four coefficients, but using 101 points allowed us to evaluate whether the model of the equation 3.8 was the most appropriate. Figure 3.14 shows that the chosen model appropriately describes the data (R2 $\geq$ 0.98).

To model the relationship between pupil size (PS) and luminosity (L) for each colour (red, green, blue, and grey), we use different equations based on the relationship between them and apply Bayesian Information Criteria (BIC) to obtain the best model with the lowest BIC. The result shows (see in Figure 3.2) that, across all four colours, the Mixed model (Exponential + Linear) gives the lowest BIC, meaning it best balances fit quality and complexity. The Exponential-only model is consistently the second-best, performing reasonably well but not as good as Mixed. The Linear and Gaussian models perform poorly in all cases, with much higher BIC values. For Grey, the Gaussian model failed to converge, confirming it's not an appropriate choice for monotonic pupil responses. Therefore, we finalised the following equation:

$$PS = a_i \cdot e^{-b_i \cdot L} + c_i \cdot L + d_i, \quad i \in \text{red}, \text{green}, \text{blue}, \text{grey} \tag{3.8}$$

where PS represents the predicted pupil size, L is the luminosity value, and a_i, b_i, c_i, d_i are the coefficients defining the curve for each colour. The term $e^{-b_i \cdot L}$ models the exponential decay of pupil size in response to increasing luminosity, $c_i \cdot L$ captures the linear response of pupil size, and d_i represents the baseline pupil size when $L = 0$.

The calculation process begins with the initialisation of coefficient values (P_0) set to $[1, 0, 0, 1]$, where $a_0 = 1$, $b_0 = 0$, $c_0 = 0$, and $d_0 = 1$. This serves as a starting point for optimisation. These initial values were determined based on the overall shape of the pupil size-luminosity curve across all colours. Pupil size measurements are recorded for different luminosity levels across the four colours, and the data is preprocessed to remove noise and inconsistencies.

Next, non-linear least squares regression is used to fit Equation (3.8) to the recorded pupil size data, minimising the difference between observed and predicted pupil sizes. The fitted coefficients (a_i, b_i, c_i, d_i) for each colour are obtained and stored in Table 3.3.

Table 3.2: Model fitting results across colours. RSS = residual sum of squares, BIC = Bayesian Information Criterion.

Colour	Model	Parameters	RSS	BIC
Blue	Exponential	$a = 3.54,\ b = 1.44,$ $d = 2.46$	1.324	-434.6
	Linear	$c = -0.101,\ d = 3.54$	50.515	-64.1
	Gaussian	$a \approx -2.0 \times 10^4,\ \mu = 11.0,$ $\sigma = 776,\ d \approx 2.0 \times 10^4$	33.664	-96.6
	Mixed (Exp+Lin)	$a = 3.44,\ b = 1.62,\ c = -0.019,$ $d = 2.63$	**0.787**	**-483.5**
Red	Exponential	$a = 2.70,\ b = 1.22,$ $d = 3.04$	0.970	-466.6
	Linear	$c = -0.142,\ d = 4.11$	29.221	-120.5
	Gaussian	$a \approx -2.25 \times 10^4,\ \mu = 6.97,$ $\sigma = 524,\ d \approx 2.25 \times 10^4$	13.692	-189.3
	Mixed (Exp+Lin)	$a = 2.63,\ b = 1.34,\ c = -0.015,$ $d = 3.15$	**0.887**	**-471.1**
Green	Exponential	$a = 3.16,\ b = 1.10,$ $d = 2.39$	1.816	-402.0
	Linear	$c = -0.049,\ d = 3.30$	43.561	-79.4
	Gaussian	$a \approx -1.67 \times 10^4,\ \mu = 17.15,$ $\sigma = 1159,\ d \approx 1.67 \times 10^4$	26.557	-121.1
	Mixed (Exp+Lin)	$a = 3.13,\ b = 1.23,\ c = -0.007,$ $d = 2.50$	**1.592**	**-410.9**
Gray	Exponential	$a = 2.65,\ b = 0.24,$ $d = 2.75$	5.568	5.68
	Linear	$c = -0.0378,\ d = 4.18$	29.977	174.4
	Gaussian	— (did not converge)	—	—
	Mixed (Exp+Lin)	$a = 2.44,\ b = 0.56,\ c = -0.018,$ $d = 3.41$	**1.035**	**-163.0**

Once the optimal coefficients are determined, Equation (3.8) is used to predict pupil size at any given luminosity L. The calculation follows these steps: computing the exponential term $e^{-b_i \cdot L}$, multiplying by a_i to obtain the non-linear decay component, computing the linear response $c_i \cdot L$, adding the baseline pupil size d_i, and summing all terms to obtain the predicted pupil size.

This mathematical model is applied in further analyses, including emotion detection and individual calibration procedures. It ensures accurate pupil size estimation by combining exponential and linear components while adapting to colour-specific variations.

Table 3.3 shows the coefficient values of the model presented in Figure 3.14 [109]. It was necessary to recalibrate this first estimate of the coefficients for each new participant. In a real-world scenario, it is essential to redetermine the coefficients before

Table 3.3: Values of the four coefficients in Equation 3.8, for each colour.

colour	a_i	b_i	c_i	d_i
Red	2.631718881	1.337185719	-0.015263019	3.150067507
Green	3.125971983	1.232499771	-0.007369488	2.503629301
Blue	3.443025449	1.616759396	-0.019305937	2.62718504
Grey	2.446582212	0.563893338	-0.018479723	3.414006057

starting an experiment and tailoring them to the individual experimental participant. To achieve this, each participant must complete a calibration procedure before any experiment by watching a video of 27 monochrome frames [109]. These frames consisted of red, blue, green, and grey colours, each displayed for 4 seconds, resulting in a total duration of 108 seconds. The images were created using all possible combinations of three RGB intensity levels (0%, 50%, and 100%), ranging from the darkest (0%, 0%, 0%) to the brightest (100%, 100%, 100%). Out of 27 monochrome frames, nine frames with primary colours—red, blue, green, and grey (three per colour but considering that only one image is needed 0%, 0%, 0%, being common to all colours, for red two more frames were - 50%, 0%, 0% and 100%, 0%, 0%, for green two more frames were - 0%, 50%, 0% and 0%, 100%, 0%, for blue two more frames were - 0%, 0%, 50% and 0%, 0%, 100%, for grey two more frames were - 50%, 50%, 50% and 100%, 100%, 100%) were used to calibrate the pupil size prediction model for each participant. In contrast, the remaining frames were utilised as test frames. During calibration, the coefficients presented in table 3.3 serve as initial values and were then varied until, for each colour, the curve described by equation 3.8 passes as close as possible to the three detected points. We tried other methods, but this is the one that gave the best results during the test. This calibration process accounts for individual variations in pupillary responses to colour and luminosity and the difference between different monitors simultaneously (described in the following paragraph), establishing a reliable baseline for accurate subsequent measurements.

Recalculation of the Fitted Model using Calibration data. To improve our model for predicting individual pupil sizes, we recalibrate a general model—defined by Equation (3.8) with the coefficients in Table 3.3—using participant-specific calibration measurements. Although the general model serves as a baseline, it does not capture individual variations in pupil response. Therefore, we use calibration data collected at three standard intensity levels (0%, 50%, and 100%) for each colour (red, green, blue, and grey) and interpolate additional points to refine the model.

Step 1: Generating Additional Calibration Points. Since our calibration data are originally available at only three intensity levels, we divide the calibration range into

two intervals:

- **Interval 1:** 0% (minimum) to 50% (maximum)
- **Interval 2:** 50% (minimum) to 100% (maximum)

Within each interval, we select nine additional intensity levels equidistant in terms of luminosity between the two endpoints. This results in 21 intensity levels per colour, including the original 0%, 50%, and 100% intensities. These additional points ensure a more structured and precise interpolation, enhancing curve-fitting accuracy.

Step 2: Predicting Pupil Size Using the General Model. Using Equation (3.8) (hereafter referred to as the **general model**), we compute the predicted pupil size at all 21 intensity levels.

For the first interval (0% to 50%), the overall predicted difference is:

$$\Delta_{\text{general}} = \text{Pupil Size}_{0\%} - \text{Pupil Size}_{50\%} \tag{3.9}$$

For any intermediate intensity $x\%$ within this interval:

$$\Delta_{\text{general}}(x) = \text{Pupil Size}_{0\%} - \text{Pupil Size}_{x\%}, \quad \forall x \in (0\%, 50\%) \tag{3.10}$$

This step is then repeated for the second interval (50% to 100%).

Step 3: Rescaling Using Participant-Specific Data. For each participant, measured pupil sizes are available at 0%, 50%, and 100%. In the first interval, the measured difference is:

$$\Delta_{\text{participant}} = \text{Measured Pupil Size}_{0\%} - \text{Measured Pupil Size}_{50\%} \tag{3.11}$$

To adapt the predictions to the subject, we rescale the intermediate differences. For any intensity $x\%$:

$$\Delta_{\text{rescaled}}(x) = \Delta_{\text{participant}} \times \left(\frac{\Delta_{\text{general}}(x)}{\Delta_{\text{general}}} \right), \quad \forall x \in (0\%, 50\%) \tag{3.12}$$

The rescaled pupil size at $x\%$ is:

$$\text{Rescaled Pupil Size}_{x\%} = \text{Measured Pupil Size}_{50\%} - \Delta_{\text{rescaled}}(x) \tag{3.13}$$

The same procedure is applied to the second interval (50% to 100%), using the measured pupil size at 50% as the new reference point.

Step 4: Recalibrating the Model. Once rescaled pupil sizes are computed for all 21 intensity levels, these new data points are used to refit Equation (3.8) for each

participant. This recalibrated model now incorporates both the structure of the general model and individual variations in pupil response.

By applying this recalibration process separately for each participant and each colour, the final emotion detection model is refined to achieve more accurate and customised pupil size estimations.

Determining pupil size from RGB values. Finally, to compute pupil size from RGB values, the two models, luminosity prediction and pupil size prediction, were combined according to the following Equation:

$$\begin{aligned} PS &= a \cdot e^{-b \cdot k \cdot f(r,g,b)} + c \cdot k \cdot f(r,g,b) + d \\ &= a \cdot e^{-g \cdot f(r,g,b)} + h \cdot f(r,g,b) + d \end{aligned} \tag{3.14}$$

where the pupil size is expressed as a function of the RGB (see equation 3.7) intensity value k, $g = b \cdot k$ and $h = c \cdot k$.

To fit the model, we need to compute four coefficients: a, d, g, and h. These coefficients need to be computed for each participant and each experiment using the calibration procedure mentioned in 3.2.6, given that each participant has a different response to light, each monitor is different, and each experimental setting is different (e.g. distance of the participant from the screen, monitor's settings, etc.).

The model described by equation 3.8 enabled us to predict the pupil size of an individual observing a primary colour or a grey image with brightness levels ranging from 0 to 100. Our subsequent objective was to expand this capability to predict pupil size in response to monochrome non-primary colour images of any arbitrary colour and luminosity. For example, an image with RGB values (100, 75, 80) represents a dark pink hue, with red being the dominant component. We pursued this goal with two approaches, which later turned out to be partially complementary. The first approach, which we called "grey-based," consisted of considering the pupil variation dependent only on the total brightness of the image and not on the particular colour. The second approach, called "colour-based," on the other hand, consisted of considering the brightness of each primary colour in the image differently. We expected that the first approach would work best for figures where the three primary colours were relatively balanced and that the second approach would work best for images in which one of the three primary colours was prevalent.

In our first approach, which we called the "grey-based" approach, we considered the pupil size variation independent of the specific colour, and we used Equation 3.14 only for the grey colour:

$$PS_{\text{grey}} = a_{\text{grey}} \cdot e^{-g_{\text{grey}} \cdot f(r,g,b)} + h_{\text{grey}} \cdot f(r,g,b) + d_{\text{grey}} \tag{3.15}$$

where PS is the pupil size, a_{grey}, g_{grey}, h_{grey}, d_{grey} are the coefficients for grey colour, and $f(r, g, b)$ is the function of luminosity for RGB value. The “grey-based” approach calculated the image’s luminosity on the screen as a weighted average of the luminosity of the eight nearest images in our look-up table, as explained above. We then calculated the relative pupil size as if the image were composed equally of the three primary colours (grey image, according to the definition of grey used in this paper). In practice, we used equation 3.8 with `i= grey` (see equation 3.15). To evaluate the "grey-based" approach and the calibration procedure, we considered only the three grey-scale images of the calibration video: (0, 0, 0), (50, 50, 50), and (100, 100, 100). The other images were used to try alternative methods. We tested 20 subjects across five different monitor screens with varying resolutions, ranging from a minimum of 1920×1080 to a maximum of 3840×2160. The test measurements included 18 ($27 - 9 = 18$) images from the calibration video and nine additional monochrome test images.

We performed all experiments in the unlit laboratory, and the maximum luminosity reaching the eyes of our participants was 57 lux, which is the maximum luminosity of our reference screen (see also Figure 3.14). Since we did not have measurements taken at the highest luminosity level, we assumed that the pupil size at 100 lux, which corresponds to typical daylight conditions, was 80% of the pupil size recorded at the maximum screen luminosity in the calibration video [265].

In the "colour-based" approach, we fitted the model represented by the Equation 3.14 independently for each colour to obtain the contribution to the pupil size given by the luminosity at each colour:

$$\begin{aligned} PS_{\text{red}} &= a_{\text{red}} \cdot e^{-g_{\text{red}} \cdot f(r,0,0)} + h_{\text{red}} \cdot f(r,0,0) + d_{\text{red}} \\ PS_{\text{green}} &= a_{\text{green}} \cdot e^{-g_{\text{green}} \cdot f(0,g,0)} + h_{\text{green}} \cdot f(0,g,0) + d_{\text{green}} \\ PS_{\text{blue}} &= a_{\text{blue}} \cdot e^{-g_{\text{blue}} \cdot f(0,0,b)} + h_{\text{blue}} \cdot f(0,0,b) + d_{\text{blue}} \end{aligned} \tag{3.16}$$

where PS is the pupil size, a_{red}, g_{red}, h_{red}, d_{red} are the coefficients for the red colour, a_{green}, g_{green}, h_{green}, d_{green} are the coefficients for the green colour, a_{blue}, g_{blue}, h_{blue}, d_{blue} are the coefficients for the blue colour.

To find all the coefficients we fitted the three independent models described by Equations 3.16 using the pupil sizes recorded during calibration, and corresponding RGB intensity values of the following images: one black image (0, 0, 0), two red images (50, 0, 0), (100, 0, 0), two green images (0, 50, 0), (0, 100, 0), two blue images (0, 0, 50), (0, 0, 100). Then, we used the same procedure described in the 3.2.6.

We fitted the three models for each participant and related monitor. Then, we predicted the pupil size for each test image as if presenting one colour per time: given a monochrome test image with RGB intensity values (r, g, b), we considered three different monochrome images, with RGB intensities (r, 0, 0), (0, g, 0), (0, 0, b), in three

different moments. The idea was to disentangle the effect of each wavelength (colour) on pupil size. For example, for the test picture with RGB intensity (64, 86, 45), we pretended to have only the red component (64, 0, 0), then the green one (0, 86, 0), and then the blue one (0, 0, 45). Using the Equations 3.16 we obtained three different predicted pupil sizes PS_{red}, PS_{green}, and PS_{blue}. We computed the final pupil size as a weighted average of these three contributions:

$$\begin{aligned} PS = &(\frac{\mathrm{r}}{\mathrm{r}+\mathrm{g}+\mathrm{b}}) \cdot PS_{\mathrm{red}} \\ &+ (\frac{\mathrm{g}}{\mathrm{r}+\mathrm{g}+\mathrm{b}}) \cdot PS_{\mathrm{green}} \\ &+ (\frac{\mathrm{b}}{\mathrm{r}+\mathrm{g}+\mathrm{b}}) \cdot PS_{\mathrm{blue}} \end{aligned} \tag{3.17}$$

where PS is predicted pupil size. For example, for a picture with RGB values (64, 86, 45), we have:

$$PS = (\frac{64}{195}) \cdot PS_{red} + (\frac{86}{195}) \cdot PS_{green} + (\frac{45}{195}) \cdot PS_{blue}$$

The “colour-based” approach was tested simultaneously with the “grey-based” approach (on the same 20 subjects, on the same five different screens with different resolutions, etc.).

Since the "grey-based" method performed more accurately on images where the three primary colours had similar intensities, while the "colour-based" method excelled when one colour was dominant, we combined both approaches to leverage their respective strengths. Consequently, the pupil size was calculated as a linear combination of the values obtained with the “grey-based” approach and those obtained with the “colour-based” approach:

$$\begin{aligned} PS = K \cdot (&a_{\mathrm{grey}} \cdot PS_{\mathrm{grey}} + a_{\mathrm{red}} \cdot PS_{\mathrm{red}} \\ &+ a_{\mathrm{green}} \cdot PS_{\mathrm{green}} + a_{\mathrm{blue}} \cdot PS_{\mathrm{blue}}) + C \end{aligned} \tag{3.18}$$

with the constraint

$$a_{\mathrm{grey}} + a_{\mathrm{red}} + a_{\mathrm{green}} + a_{\mathrm{blue}} = 1 \tag{3.19}$$

We named this approach the "combined" approach (see Figure 3.15). Given the constraint in Equation 3.19, it is expected that, after fitting, the value of the multiplicative coefficient K is close to 1 and that of the intercept C is close to 0. This would mean that the value of the pupil size PS is given by the values PS_{grey}, PS_{red}, PS_{green}, and PS_{blue} added together in different percentages.

The “combined” approach was tested simultaneously with the “grey-based” and the "colour-based" approaches (with the same 18 participants, the same five different

screens with different resolutions, the same images, etc.).

For each participant, first, we calculated the values PS_{grey}, PS_{red}, PS_{green}, and PS_{blue} using the nine images of the calibration procedure, as described above. Then, we trained and tested the model in Equation 3.18 (see Figure 3.15).

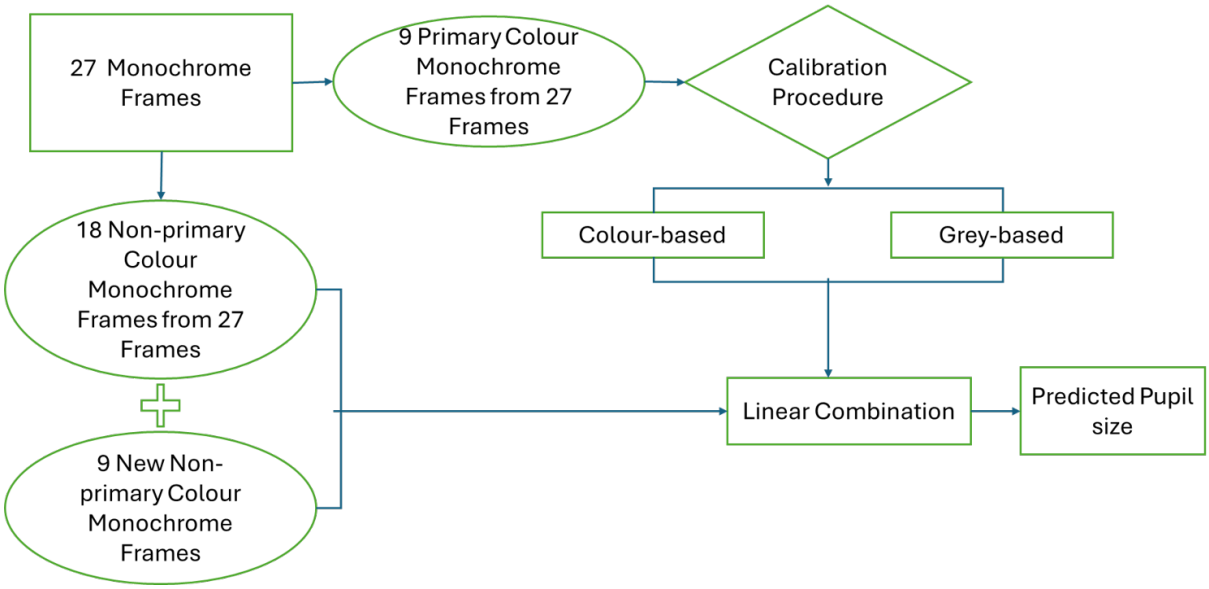


Figure 3.15: Testing the combined approach.

Since the best results were obtained with the "combined" method, we used only this approach from then on and throughout this paper.

Testing the Luminosity Effect Prediction Model in a well-lit laboratory and with non-monochrome images. Finally, we tested our LEPM model on non-monochrome images in both dark and light environments. We recruited an additional 10 participants and used the 27 monochrome test images mentioned above, plus 46 non-monochrome emotionally neutral images (see Figure 3.16). Each image was displayed for 4 seconds, resulting in a video of 4 minutes and 52 seconds plus 36 seconds for the calibration procedure (9 images). All images were displayed for 4 seconds, and to process them, we calculated the mean pupil size over these 4 seconds. In the case of test images, we trimmed the first half-second to eliminate any potential influence from the preceding image. However, we did not apply this trimming to calibration frames because they were created with a gradual increase in luminosity. Conversely, test images are presented randomly, allowing for sequences where a high-luminosity image may follow a low-luminosity image or vice versa. We used five different Dell monitors and Dell laptops with varying brightness levels and a resolution of 1920×1080. Pupil size was measured using a Tobii Pro Nano eye-tracker. The results obtained with the non-monochrome images were not as promising as those obtained with monochrome images. In particular, images that contained a very bright region within a dark figure yielded much worse

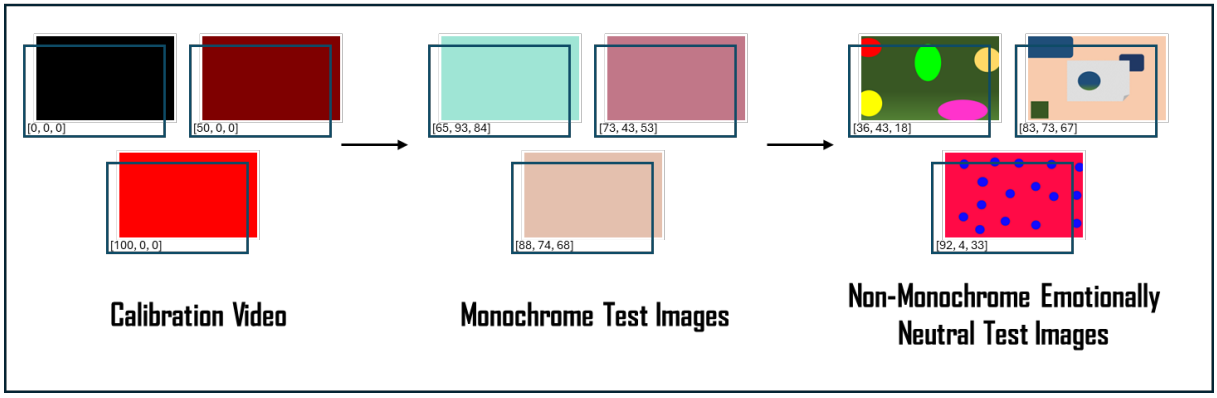


Figure 3.16: Experiment flow of LPEM validation including Pupil Size Calibration Procedure.

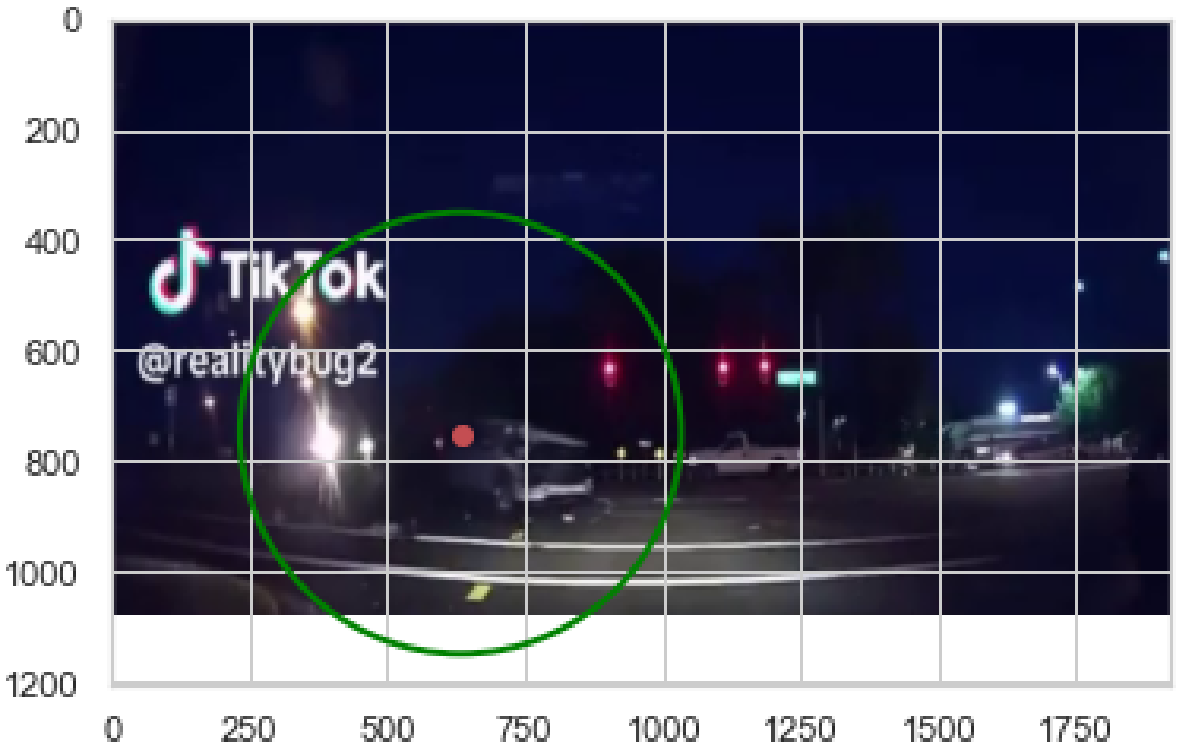


Figure 3.17: Visualisation of a video frame from an emotional audiovisual clip with a 300-pixel radius circle (green circle) indicating the participant's gaze location (red dot).

results than monochrome images. Considering that the average colour and brightness of the image were not effective strategies in these cases, an image could be dark on average while the participant was looking at a bright region, potentially inducing a more pronounced pupillary constriction than that elicited by an image with an average brightness level. Therefore, it was necessary to determine where participants were looking, as indicated by the recorded eye-tracking data. Specifically, we analysed a region with a 300-pixel radius centred on the point where the gaze was directed at each instant. If the average brightness of the specific region exceeded that of the overall

image, we used the region's average instead. For example, in the Figure shown in 3.17, the participant is not looking far from the bright light source (the red dot represents the eye-gaze location). As a result, the brightness affects pupil size. Therefore, calculating the average luminosity of that specific region is more reliable than averaging over the entire frame, as the overall image is relatively dark and would result in a lower average luminosity. This method provided a more accurate representation of the subjective luminance and led to improved results, which we present in the Results Section 4.2.2. All the code related to the models presented in this section was written in Python.

Development of Arousal Detection Model. The second part of the study aimed to demonstrate that the pupil size value discounted from the luminosity part was immediately usable for detecting the level of emotional arousal without the need to use complex pre-processing, advanced ML or even deep learning models. We used the pupil size data measured during the data collection process for 32 audiovisual clips, split the data for those clips due to EIIS, and did data pre-processing, as explained in the following.

Pupil Data Pre-process. To preprocess the recorded pupil size data, we implemented a robust data imputation process that ensures the dataset's reliability and accuracy. The first step involved identifying and marking blink-related distortions. Data points identified as blinks by the eye tracker were substituted with null values. To account for the potential pre- and post-blink effects on pupil size measurements, an additional window of two milliseconds before and after each blink was also marked as null. This precaution minimises the influence of abrupt changes associated with blinking on the subsequent analysis.

Following this, we performed data imputation to fill the null values. The missing data points were replaced using a gradual interpolation method. Specifically, we employed a linear interpolation approach that considers the trends in the data immediately before and after the null segments. By interpolating in this way, we maintained the natural trajectory of pupil size changes over time, preserving the physiological relevance of the data while mitigating noise introduced by blinking.

This pre-processing step is critical as blinks and their associated distortions can significantly distort pupil size measurements, potentially leading to inaccurate conclusions about emotional arousal. The careful handling of missing data ensures the integrity of the dataset, enabling the extraction of meaningful emotional responses from the processed pupil size data.

Arousal Effect Prediction. After pre-processing the pupil size data, we applied the LEPM to predict the pupil size for each video frame across all video clips without influencing ambient and screen luminosity.

We trained the model described by the Equations 3.18 and 3.19 by calculating PS_{grey}, PS_{red}, PS_{green}, and PS_{blue} frame by frame and the coefficients a_{grey}, a_{red}, a_{green}, a_{blue}, K, and C for each video clip. We hypothesised that since there was also an arousal component (in addition to the light component), the fitting would be worse and that the error in the fitting would be due specifically to the effect of arousal, given that the model was designed to capture only the component due to luminosity. In other words, we hypothesised that the measured pupil size would be equal to the pupil size predicted by the LEPM model as an effect of luminosity plus a residual:

$$\begin{aligned} PS_{\text{measured}} = [K \cdot (a_{\text{grey}} \cdot PS_{\text{grey}} + a_{\text{red}} \cdot PS_{\text{red}} \\ + a_{\text{green}} \cdot PS_{\text{green}} + a_{\text{blue}} \cdot PS_{\text{blue}}) + C] + Residual \end{aligned} \tag{3.20}$$

where

$$\begin{aligned} PS_{\text{luminosity}} = [K \cdot (a_{\text{grey}} \cdot PS_{\text{grey}} + a_{\text{red}} \cdot PS_{\text{red}} \\ + a_{\text{green}} \cdot PS_{\text{green}} + a_{\text{blue}} \cdot PS_{\text{blue}}) + C] \end{aligned} \tag{3.21}$$

so that

$$PS_{\text{measured}} = P_{\text{luminosity}} + Residual. \tag{3.22}$$

The Residual was the portion of pupil size that cannot be explained by luminosity, i.e., the portion due to arousal:

$$Residual = PS_{\text{arousal}}. \tag{3.23}$$

Hence, to extract arousal-related information for each video, we subtracted the predicted pupil size due to luminosity from the pre-processed measured pupil size:

$$PS_{\text{arousal}} = PS_{\text{measured}} - PS_{\text{luminosity}}. \tag{3.24}$$

The resulting difference represents the emotional arousal level of participants watching the video clips, as measured by pupil size.

ADM Testing. To test the model, we compared its predictions of a participant's arousal while watching audiovisual clips, based on the measured pupil size, against the arousal ground truth. For each clip and each participant, we had the self-reported arousal value (ground truth) $\text{Arousal}_{\text{self-reported}}$, the recorded pupil size value PS_{measured}, and the pupil size value corrected for luminosity and due only to arousal PS_{arousal}. We then calculated the Pearson correlation between the pupil size corrected and not corrected for luminosity, and the self-reported arousal value (see results section). Then, we computed the average pupil size for each clip and participant across the salient intervals (EII) obtained during EIIS (see section 3.2.4, with and without our luminosity correction.

While our results demonstrated the potential of pupil size as a reliable indicator of emotional arousal, we wanted to prove that PS_{arousal}, derived from our model, can be used without further processing or the use of complex machine learning techniques, becoming accessible to the entire scientific community, while still obtaining robust results. For this purpose, we used the following procedure. We used a leave-one-participant-out (LOPO) cross-validation approach, temporarily eliminating one participant from our dataset as if they were a new participant.

We then calculated the pupil size corrected for luminosity PS_{arousal} and self-reported arousal $\text{Arousal}_{\text{self-reported}}$ for each video and each participant remaining in our dataset. We then assumed a simple linear relationship between the size of the pupil corrected for luminosity (component due to arousal) and self-reported arousal:

$$PS_{\text{arousal}} = a \cdot \text{Arousal}_{\text{self-reported}} + b \tag{3.25}$$

We fitted the model 3.25 to all the videos and all the participants (except one). We obtained a good fit (see Results).

By inverting Equation 3.25, we obtained an estimate of self-reported arousal $\widehat{\text{Arousal}}_{\text{self-reported}}$ as a function of pupil size corrected for luminosity:

$$\widehat{\text{Arousal}}_{\text{self-reported}} = \frac{PS_{\text{arousal}} - b}{a}. \tag{3.26}$$

We then repeated the same procedure for pupil size, which was not corrected for luminosity. In this case, in Equations 3.25 and 3.26, PS_{arousal} must be replaced with PS_{measured}. We predicted the self-reported arousal value for the eliminated participant starting from the pupil size, using Equation 3.26, and for all the videos. We compared the obtained values with the ground truth for the luminosity-corrected and uncorrected pupil size. We repeated this procedure, eliminating one participant at a time, and finally calculated the average results for our sample.

This method shows that pupil size, isolated from luminosity, directly plays an integral part in the emotion detection model without any complex or advanced machine learning model. Instead, we couldn't use this kind of simple feature directly to train a simple machine learning model for FER and GSR, so we used advanced machine learning techniques described in the following sections.

3.3 Advanced Machine Learning Techniques for FER, Pupil Size, and GSR

In this section, we present the training and evaluation of our multimodal emotion detection models using advanced machine learning techniques. Unlike many existing

studies that either focus on a single modality or rely on basic statistical descriptors, our framework integrates FER, luminosity-corrected pupil size, and GSR within a unified pipeline. This multimodal configuration is itself novel, as it combines complementary information channels: FER captures expressive behaviour, pupil size reflects cognitive and affective arousal once corrected for lighting effects, and GSR indexes autonomic activation. To the best of our knowledge, this is the first systematic study to jointly model these three signals for continuous valence–arousal prediction.

The pipeline begins with extensive feature engineering tailored to each modality. For FER, we introduced a vectorial mapping of raw emotion intensity outputs into valence–arousal coordinates, grounded in Russell's circumplex model (see Section 3.1.7). This approach preserves psychological interpretability while producing continuous features that can accommodate multi-emotion mixtures. For pupil size, we developed a novel luminosity-isolation model based on calibration with primary and grayscale colours, effectively disentangling ambient light effects from genuine affective changes. For GSR, we extracted a comprehensive set of features spanning time, frequency, and time–frequency domains, going beyond the simple statistical measures typically reported in the literature. Together, these diverse features enabled the model to capture both fast-changing and longer-term dynamics of autonomic activity, providing a richer representation of emotional processes.

For this analysis, feature-level fusion was employed to integrate multiple data modalities, followed by supervised learning using robust algorithms like gradient boosting regressors. While tree-based models like XGBoost are generally more resilient to heteroscedasticity [389], the non-constant variance in error residuals, than traditional linear models, significant variations in physiological responses between participants could still affect model stability and potentially lead to overfitting. To mitigate this, a leave-one-participant-out (LOPO) cross-validation regime was used. This robust evaluation strategy ensures subject-independent generalisation by training the model on data from all participants except one, and then validating on the held-out participant's data [390]. This approach accounts for inter-participant variability and provides a more reliable assessment of performance across diverse populations, which is particularly crucial for real-world clinical and applied contexts where consistency is essential. The synergistic combination of gradient boosting's inherent strength in handling varying error distributions and the LOPO cross-validation's ability to provide a less-biased generalisation estimate strengthens the reliability of the reported outcomes.

Our approach makes three key novel contributions: (i) it introduces a previously unexplored multimodal signal configuration that integrates FER, luminosity-corrected pupil size, and GSR; (ii) it advances preprocessing with innovations such as a vectorial FER-to-circumplex mapping and a calibration-based luminosity correction model for pupil data; and (iii) it employs rigorous subject-independent evaluation to establish

reliable and generalisable benchmarks. Collectively, these contributions yield a framework that is not only more accurate but also psychologically interpretable and broadly applicable, thereby laying the groundwork for future multimodal affect-aware systems.

3.3.1 FER, Pupil Size and GSR Feature Extraction

First, we extracted a comprehensive feature from FER, pupil size and GSR. For FER, we used the basic emotion detected by the iMotions software, and for pupil size, we used corrected and non-corrected luminance. For GSR, we used the pre-processed GSR from iMotions. Once the features were extracted and cleaned, we trained machine learning models to predict emotional dimensions, specifically, arousal and valence. This pipeline allowed us to assess the predictive power of each modality individually and in combination.

FER Features Extraction

To extract features using FER features for predicting arousal and valence, we followed the approach outlined in the pilot study (see Section 3.1.7). We first generated a vector representation of the seven basic emotions detected by iMotions software on the valence-arousal plane for each video segment, where (x, y) corresponds to (valence, arousal). We extracted statistical features from these vectors, such as mean, maximum, minimum, and standard deviation, for the full video and emotionally salient intervals (EII), as defined in EIIS 3.2.4.

Analysing features from the entire video alongside those from emotionally significant intervals provides a deeper understanding of emotional expression. While the whole video captures broader emotional trends and baseline states, salient intervals focus on moments of heightened emotion. This comparison enables us to evaluate which temporal scope offers more precise emotional predictions. Later, we trained a machine learning model using features from the entire video and salient interval, which is described in the model training section 3.3.2

Pupil Size Feature Extraction

Pupil size signals were corrected for luminosity variations to isolate emotion-related changes. Building on findings by Celniak et al. [293], highlighting pupil dynamics' emotional significance, we extracted higher-order statistical features from both corrected and uncorrected signals for the emotionally salient intervals (EII) mentioned in the EIIS 3.2.4. Since pupil responses are spontaneous and involuntary, focusing on segments where emotional content is present rather than analysing the entire video is more meaningful. This targeted approach enables a more accurate and relevant representa-

tion of features, thereby enhancing the effectiveness of subsequent emotion prediction models.

Similar to a pilot study, Such characteristics are crucial for reflecting the subtle pupil size changes associated with different emotional states. Using these features, we trained an emotion detection model in the model training section 3.3.2.

GSR Features Extraction

As detailed in the pilot study on GSR analysis section 3.1.7, we utilised GSR data recorded and pre-processed using iMotions software. While iMotions provides various features, including phasic, tonic components and peak detection, we focused specifically on the phasic, tonic, and calibrated (noise-free) GSR signals for feature extraction [383], [391]. This selection ensured the reliability and relevance of the extracted features in our analysis. Using the GSR data, we extracted emotion-related features for each participant across all audiovisual stimuli. Unlike pupil size, GSR responses are slower due to the nature of sweat gland activity [212], [258], [392], making it unsuitable to analyse only specific intervals corresponding to elicited emotions in the audiovisual clip. Instead, we utilised GSR data from the entire duration of each audiovisual clip to extract relevant features.

To enhance the accuracy of emotion detection, various physiological signal features were extracted across multiple domains. Time-domain features capture baseline and momentary fluctuations in skin conductance, providing insights into general arousal and variability. Peak analysis features help quantify response intensity and speed by examining specific SCR characteristics. Signal dynamics track changes in response patterns over time, while temporal decomposition focuses on variations in the latter phase of physiological signals. Spectral analysis examines the complexity and frequency power of signals, particularly in low- and very-low-frequency bands associated with emotional and cognitive processing. Additionally, advanced statistical features offer non-linear variability and stability measures, further refining the emotional characterisation. Finally, Mel-Frequency Cepstral Coefficients (MFCC) features capture the signal's textural properties, contributing to a more robust emotion detection framework.

Relying solely on statistical features for GSR emotion detection can be limiting, as it may overlook the complex, non-linear patterns inherent in emotional responses. By incorporating a broader range of features, including spectral, temporal, and dynamic properties, we capture the full complexity of physiological reactions, allowing for a more accurate and holistic model for emotion detection. These diverse features comprehensively represent emotional responses, which are crucial for training robust and effective ML models.

Time-Domain Features.

- Tonic Mean: Average baseline SCL, reflecting general arousal [305], [391].
- Tonic Standard Deviation: Variability in the baseline skin conductance [305], [391].
- Phasic Mean: Average magnitude of momentary skin conductance responses (SCRs) [305], [391].
- Phasic Standard Deviation: Variation in the magnitude of SCRs [305], [391].
- Phasic Skewness: Asymmetry of the SCR distribution [305], [391].
- Phasic Kurtosis: "Peakedness" of the SCR distribution [305], [391].

Peak Analysis Features.

- SCR Rise Time: Average time between SCR onset and peak, indicating response speed [305], [391].
- SCR Recovery Time: Time for 63% return to baseline (skin conductance half-life) [305], [391].
- SCR Area Under Curve (AUC): Total phasic activity, representing response intensity [305], [391].
- Non-Specific SCR Rate: Number of non-specific SCRs per minute, indicating arousal frequency [305], [391].

Signal Dynamics

- Phasic Gradient Mean: Average rate of change in SCRs [305], [386].
- Tonic Gradient Standard Deviation: Variability in baseline conductance changes [305], [386].

Temporal Decomposition.

- Second Half Mean: Average of later-phase responses [305], [386].
- Second Half Standard Deviation: Variability in later-phase responses [305], [386].
- Second Half Peak Rate: Frequency of peaks in the latter half of the signal [305], [386].

Spectral Analysis.

- Spectral Entropy: Measure of signal complexity [305], [393].
- Power VLF (0.04-0.15Hz): Power in a very low-frequency band, related to thermoregulation [305], [393].
- Power LF (0.15-0.4Hz): Power in the low-frequency band, associated with cognitive/emotional processing [305], [393].

Advanced Statistical Features.

- Phasic Entropy: non-linear complexity measure of SCRs [305], [386].
- Phasic 90th Percentile: Threshold for extreme responses [305], [386].
- Tonic Quantile Difference: Measure of baseline stability [305], [386].

MFCC Features.

- MFCC 0-12: Mel-frequency cepstral coefficients, capturing signal "texture" [386].

These features comprehensively analyse GSR signals, encompassing various skin conductance changes related to emotional and cognitive processes [305], [386], [391], [393].

Before feeding the features into the model, we apply Min-Max scaling to normalise them. This step ensures that all features are on the same scale (0 to 1), preventing models from being biased toward features with larger numerical ranges. The formula used for min-max normalisation is:

$$X' = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}} \tag{3.27}$$

where, X is the original feature value, X_{min} is the minimum value of the feature, X_{max} is the maximum value of the feature, X' is the normalised value.

This normalisation reduces the influence of different feature scales, making the data more suitable for ML models and statistical analysis. Before integrating GSR with other physiological signals, we wanted to assess its effectiveness in detecting emotions. To do this, we trained a regression model using only GSR features described in the model training section 3.3.2.

After feature extractions, we moved forward to do machine learning model training, where we experimented with several machine learning models, including SVR and XGBoost. Based on comparative performance and feature analysis, we adopted XGBoost as the core regression model for its ability to model complex, non-linear feature interactions and robustness against multicollinearity, manage feature interactions, and avoid overfitting. The objective was to predict continuous emotional states- arousal and valence- using extracted physiological features.

3.3.2 Model Training Methodology

We developed and evaluated ML models using both unimodal (individual signals) and multimodal (combined signals) input. The training process was divided into three main stages. First, we trained separate models using features from FER, pupil size, and GSR to assess signal effectiveness. Then, we combined all three signals into a unified feature set to train a multimodal model. Finally, we assessed model performance using standard metrics across LOPO cross-validation and nested K-Fold validation for hyperparameter tuning.

Model Initialisation

We initialised the XGBoost Regressor with the following hyperparameters:

- **Objective**: $'reg : squarederror'$ – This objective function is used for regression tasks, minimising the squared error between the predicted and actual values.
- **colsample_bytree**: 0.3 – This parameter controls the fraction of features to sample when building each tree, helping prevent over-fitting and improving model generalisation.
- **learning_rate**: 0.1 – The learning rate shrinks the weight of each tree. A smaller value makes the model more robust but requires more converging trees.
- **max_depth**: 5 – The maximum depth of the trees. A value of 5 allows for sufficient complexity while avoiding over-fitting.
- **alpha**: 10 – L2 regularisation term on weights helps reduce over-fitting by penalising significant coefficients.
- **n_estimators**: 100 – The number of trees to train in the model, balancing between model complexity and computational efficiency.

This configuration is commonly used to control overfitting while maintaining a powerful, flexible model capable of capturing non-linear relationships in the data. We implemented a two-level cross-validation strategy, LOPO and Nested 5-fold Cross-validation, to ensure generalisability across participants and avoid overfitting.

LOPO Cross-Validation. We adopted an LOPO approach to account for participant variability. In this approach, for each iteration, data from all participants except one are used for training, and the data from the left-out participant is used for validation. The model is trained on the training set (comprising data from all participants except the one left out), allowing it to learn the relationships between the features and the target variable (Arousal or Valence). After training, the model predicts the Arousal values for the left-out participant, and performance metrics are evaluated using the actual values of that participant. This procedure is repeated for each participant, ensuring that every individual is used as a training and validation data point. This provides a comprehensive evaluation of the model's generalizability.

Cross-Validation within Each Fold. For each iteration of the LOPO method, we further employed K-Fold Cross-Validation with 5 folds to optimise the model's hyperparameters and reduce over-fitting. The K-Fold Cross-Validation was performed using the following configuration:

- **n_splits**: 5 – The training data is divided into 5 subsets (or folds). The model is trained on 4 out of the 5 folds and validated on the remaining fold, ensuring robust performance evaluation.

- **shuffle**: $True$ – This ensures that the data is randomly shuffled before splitting into folds, which helps mitigate any potential bias in the fold divisions.
- **random_state**: 42 – This random seed ensures the reproducibility of the results by maintaining the same random splits across multiple runs.

During the cross-validation process, the XGBoost model is trained on 4 of the 5 folds; the remaining fold is used for validation. Cross-validation helps select the model's optimal hyperparameters. Performing multiple iterations on different subsets of the data ensures that the model does not overfit any particular fold or set of features.

Performance Evaluation. After each iteration of the LOPO method, the following performance metrics were computed to assess the model's predictive accuracy for participants who were left out:

- **Coefficient of Determination ($R2$)**: Measures how well the model explains the variance in the target variable.
- **normalised RMSE ($NRMSE$)**: Provides an error measure normalised by the range of the target values, making it easier to compare results across different datasets or variables. If NRMSE is low, that means there is a tiny difference between the predicted and actual value.
- **Pearson Correlation (r)**: Measures the linear relationship between predicted and actual values. A higher value indicates stronger predictive performance.

These metrics were computed for each LOPO cross-validation and averaged across all iterations (across all participants) to obtain the overall model performance.

Model Performance and Hyper-parameter Selection. We employed a nested validation strategy combining LOPO cross-validation with internal 5-fold cross-validation to ensure reliable and universal predictions. Hyperparameters were optimised based on performance across all folds, and the best configuration was used to train the XGBoost model on all participants except the one left out in each LOPO iteration. This approach comprehensively evaluated model performance and supported robust generalisation, as discussed in the literature review under generalisation and transfer learning 2.7.1.

3.3.3 Unimodal Training

We applied a consistent modelling pipeline across different feature sets, using the specified model training parameters to predict arousal and valence. Specifically, we trained separate models on FER features extracted from the full video and salient intervals, pupil size features corrected and uncorrected for luminance, and GSR features extracted from the full video. For each emotional dimension (arousal and valence), an XGBoost regression model was trained using LOPO cross-validation across parti-

cipants and internal fivefold cross-validation to ensure training stability and generalisability. Model performance was assessed using $R2$, Pearson correlation coefficient r and $NRMSE$. The results, presented in section 4, highlight the contributions of FER, pupil size, and GSR features to emotion recognition.

Having established the individual predictive power of FER, pupil size, and GSR through unimodal models, we explored whether combining these signals could enhance emotion recognition performance.

3.3.4 Multimodal Feature Fusion: Integrating FER, Pupil Size, and GSR for Machine Learning Models

To investigate the synergistic effect of multimodal physiological signals on emotion prediction, we adopted a feature fusion approach, as outlined in the literature (see Chapter 2). This technique integrated all extracted features—specifically from full-video FER analysis, pupil size data corrected for luminosity, and GSR—into a single unified feature set, based on their performance in unimodal evaluations. This integration captures a richer representation of physiological responses, allowing the model to learn complementary patterns across modalities that may not be evident when each is analysed independently.

The same XGBoost regression model architecture used for unimodal training (FER, pupil size, and GSR separately) was applied to this fused dataset (see Figure 3.18. This consistency allowed for direct performance comparison between unimodal and multimodal models.

Correlation Analysis

Before model training, we computed correlation matrices using Pearson and Spearman coefficients to examine linear and non-linear relationships among the integrated features and their correlation with the target variables (Valence and Arousal). This analysis provided valuable insights into inter-feature dependencies and highlighted which features exhibited the strongest associations with emotional dimensions. While traditional feature selection methods often exclude highly correlated features, we retained all features to preserve the multimodal dataset's full expressive power. Correlation analysis was used purely for interpretability rather than dimensionality reduction.

Model Training and Over-fitting Prevention

We used XGBoost, a gradient-boosted decision tree algorithm, due to its ability to model complex, non-linear interactions and its robustness to feature multicollinearity. To ensure the model generalised well and did not over-fit the training data, several over-

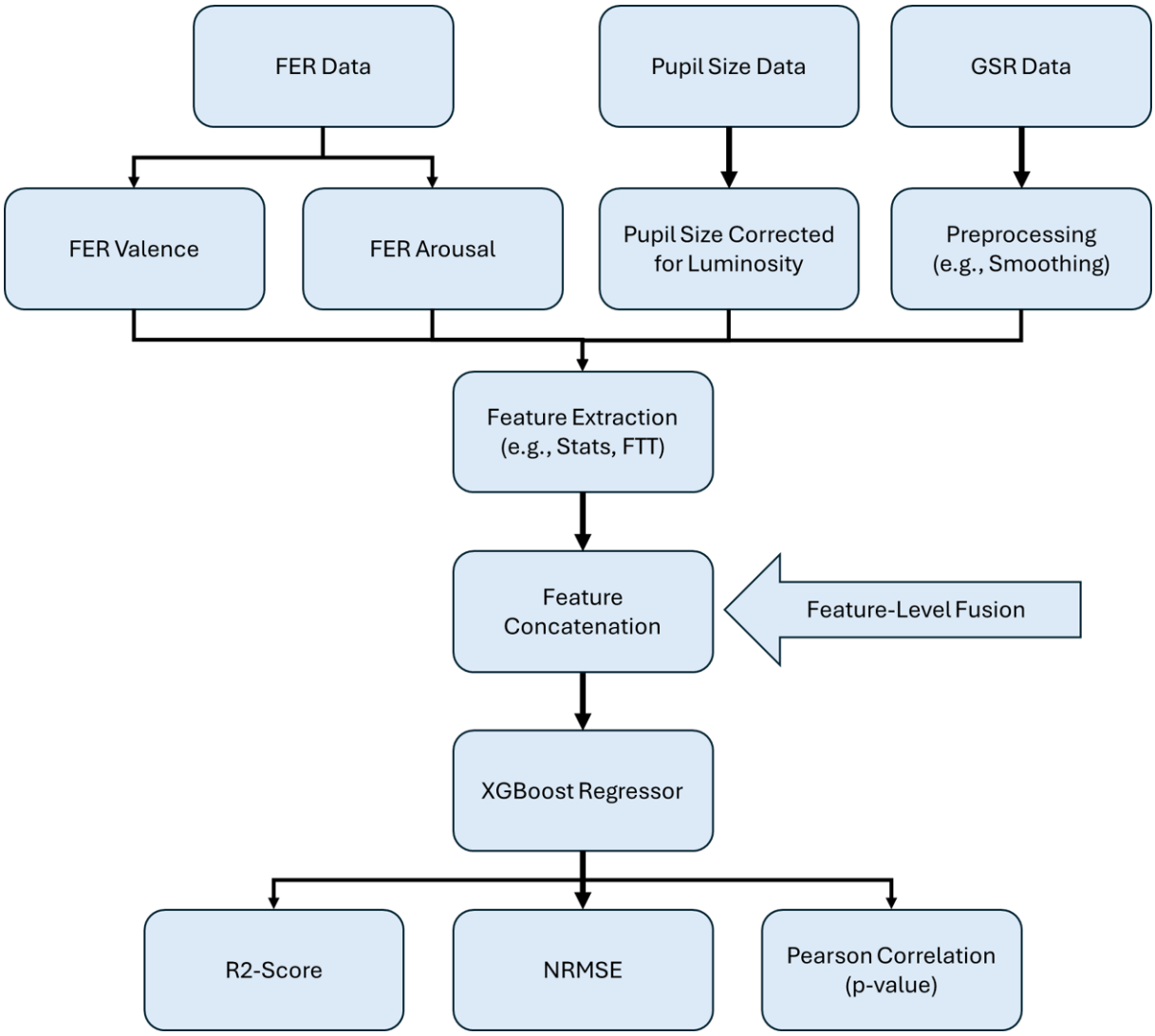


Figure 3.18: Flow Chart of Training the ML Model using Feature-Level Fusion Technique.

fitting prevention strategies were employed, like early stopping, where during training, we implemented early stopping based on validation loss, halting the process when the performance did not improve for 10 consecutive rounds. This helped prevent unnecessary boosting rounds that could lead to over-fitting. Regularisation, where we applied L1 (alpha) and L2 (lambda) regularisation penalties to control model complexity and reduce the risk of over-fitting due to high-dimensional feature fusion. We employed LOPO cross-validation at the participant level, where each participant's data was held out once as a test set while the model was trained on the remainder. A five-fold cross-validation was conducted within each training set to tune the model and validate its internal consistency.

Prediction and Evaluation

We trained the model separately using the fused feature set to predict self-reported arousal and valence. For each LOPO iteration, predictions were generated for the excluded participant using the trained model. Evaluation metrics included R2-score, Pearson correlation, and $NRMSE$. The final performance results and comparative analysis with unimodal models and other existing literature are detailed in Section 4. This multimodal approach provides insight into how integrating various physiological signals enhances emotion detection and supports the hypothesis that feature fusion leads to more robust and generalisable models.

Chapter 4

Results

This section clearly and comprehensively lays out all the analysis results from the pilot and main studies. It provides a detailed comparison and interpretation of the findings, delving into the data collected, the statistical outcomes, the trends observed, and any significant patterns that emerged during the research. We discuss how the results from the pilot study influenced the design and execution of the main study, highlighting the key adjustments and methodological tweaks we made to enhance the reliability and validity of our findings. Additionally, this section provides a closer examination of the similarities and differences between the two studies, addressing any discrepancies that may have arisen. We critically analyse the insights gained to provide a well-rounded understanding of their implications, limitations, and how they contribute to the broader research goals.

4.1 Results of the Pilot study

This section presents a comprehensive analysis of the results obtained from various study angles, utilising data from 45 participants. It covers the outcomes of emotion labelling for the INDSCAL group space and examines selected individual spaces, shedding light on how emotions are mapped across participants. We also detail the findings of FER feature extraction and analysis, highlighting significant patterns in facial emotions. Additionally, we examine the results of pupil size feature extraction, which helps us understand the physiological responses associated with cognitive and emotional states. The section continues with the extraction and analysis of GSR features, which play a crucial role in assessing the activity of the autonomic nervous system. To further enrich our understanding, a correlation matrix is presented to examine the relationships among all extracted features for a subset of participants in the pilot study. This enables a more nuanced understanding of how multimodal data interact. Finally, we conclude this section with reflections on the lessons learned from the pilot study, highlighting

key takeaways and methodological improvements that informed the design of the main study.

4.1.1 Emotion Labelling

This section presents the results of our emotion labelling analysis for 20 audiovisual clips (stimuli), evaluated using both the INDSCAL group space and selected individual spaces. The analysis provides insight into how emotions are organised within a shared perceptual space and how consistently participants' subjective experiences align with this representation.

The INDSCAL group space revealed distinct clusters corresponding to the expected quadrants of Russell's circumplex, with stimuli clearly separating along valence and arousal axes. This validates the suitability of our labelling framework, as the emergent spatial structure reflected theoretically grounded emotion dimensions rather than arbitrary statistical groupings. Beyond the group-level findings, individual INDSCAL spaces uncovered meaningful variability in how participants positioned the same stimuli, demonstrating that while a common perceptual structure exists, individual differences remain important in shaping emotional experience.

To assess the reliability of these mappings, we also analysed the variance of ratings across participants for each stimulus. Stimuli with low variance indicated strong consensus in emotional interpretation, whereas higher variance highlighted clips that elicited divergent reactions, often reflecting more ambiguous or mixed affective content. This dual analysis, linking group-level consensus with individual-level variation, provides a richer picture of emotion labelling than traditional averaging approaches. By combining dimensional mapping with individual weighting, our framework offers a unique and interpretable ground truth for training and evaluating emotion recognition models.

The image in 4.1 represents the INDSCAL group space for emotion labelling across all participants, where each coloured dot corresponds to a specific labelled emotion. The x-axis (valence) measures the positivity or negativity of an emotion, with negative values indicating unpleasant emotions and positive values representing pleasant emotions. The y-axis (arousal) measures emotional intensity, where higher values indicate highly aroused emotions (e.g., excitement or fear), while lower values correspond to calmer emotional states (e.g., sadness or relaxation). The annotations next to each point indicate stimulus names.

As illustrated in Figure 4.1, the stimulus names include category representations following an underscore for their belonging categories. As mentioned in the methodology of the pilot study, we selected audiovisual clips based on five categories using Russell's circumplex model. For instance, "HP" denotes high arousal with positive valence (rep-

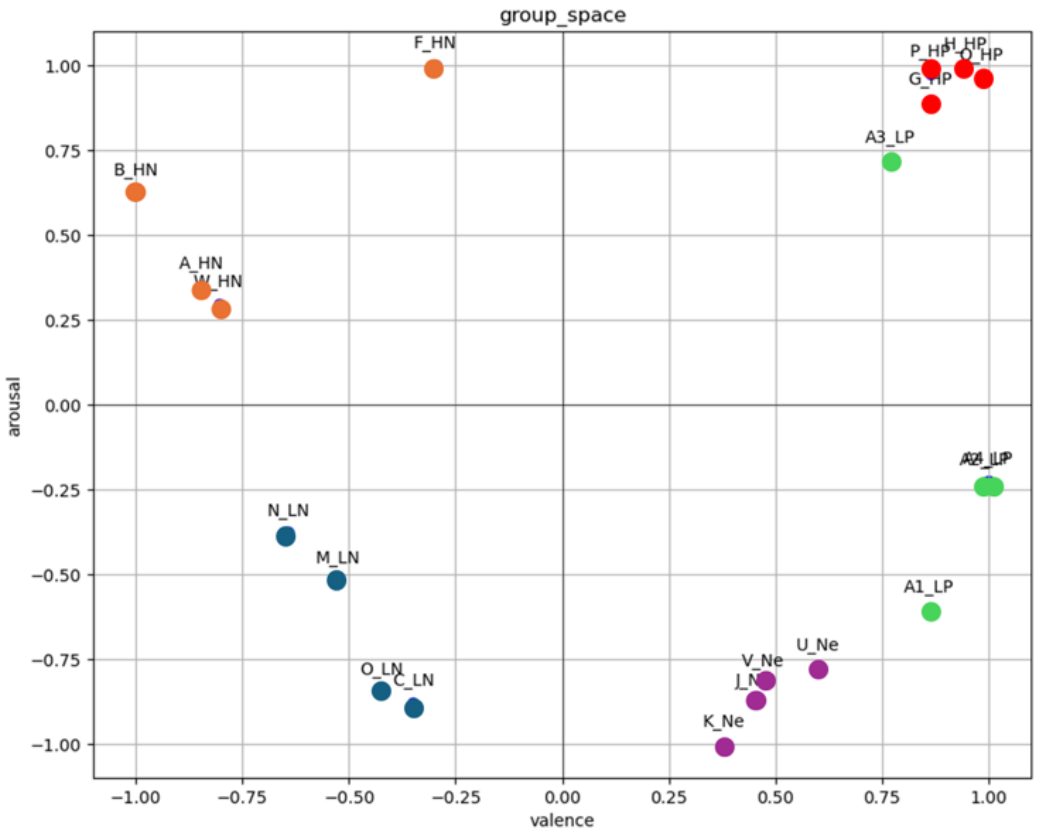


Figure 4.1: Plot of group space across all the participants for labelled emotion for HP (red), HN (orange), LP (green), LN (blue), and neutral (violet) stimuli. The plotted positions represent the mean ratings across participants for each stimulus.

resented by red-coloured circles), “HN” represents high arousal with negative valence (orange-coloured circles), “LP” stands for low arousal with positive valence (green-coloured circles), “LN” corresponds to low arousal with negative valence (blue-coloured circles), and “Ne” signifies neutral (violet-coloured circles). This naming convention was essential for accurately categorising and labelling the stimuli during the analysis process.

Examining the distribution of points, distinct clusters emerge based on emotional characteristics. In the top-right quadrant (positive valence, high arousal), labels such as `P_HP`, `O_HP`, and `G_HP` (red) appear, suggesting emotions that are both positive and highly energetic, such as excitement or joy. In contrast, the top-left quadrant (negative valence, high arousal) includes labels like `B_HN`, `A_HN` and `W_HN` (orange), which likely represent high-arousal but unpleasant emotions, such as anger or fear. Moving to the bottom-left quadrant (negative valence, low arousal), we observe labels such as `N_LN`, `M_LN`, `O_LN`, and `C_LN` (blue), which likely correspond to emotions like sad-

ness or fatigue. Meanwhile, the bottom-right quadrant (positive valence, low arousal) includes points like A1_LP, A2_LP and A4_LP (green), which represent emotions that are positive but calming, such as contentment or relaxation.

A key observation from this visualisation is that emotions appear well-structured within the valence-arousal space, with clear separations between high-energy and low-energy emotions, as well as between positive and negative affective states. Some emotions, such as F_HN (orange) and A3_LP (green), are positioned at extreme values, indicating vigorous emotional intensity. Additionally, neutral emotions U_Ne, J_Ne, V_Ne and K_Ne (violet) are clustered near the centre of the graph, confirming their balanced nature in both valence and arousal dimensions.

Table 4.1: Variance of Valence and Arousal Ratings across Participants for Each Stimulus

Stimulus	**Valence Variance**	**Arousal Variance**
A1_LP	0.017224	0.004890
A2_LP	0.024564	0.000485
A3_LP	0.012465	0.007959
A4_LP	0.025127	0.000589
A_HN	0.046391	0.001572
B_HN	0.060410	0.005354
C_LN	0.012164	0.011501
F_HN	0.011371	0.014568
G_HP	0.016224	0.013243
H_HP	0.020777	0.016688
J_Ne	0.002492	0.011042
K_Ne	0.001276	0.014479
M_LN	0.021637	0.004444
N_LN	0.029047	0.002109
O_LN	0.015633	0.010818
P_HP	0.016028	0.016136
Q_HP	0.022601	0.015807
U_Ne	0.006399	0.008328
V_Ne	0.002988	0.009572
W_HN	0.042534	0.001212

The INDSCAL group space organises emotions into distinct regions, aligning with well-established emotional models. The clustering of emotions suggests that the extracted labels successfully capture variations in emotional perception across participants. Importantly, by incorporating both mean positions and variance measures, we can assess not only where stimuli are placed in the emotion space, but also the level of agreement or disagreement among participants (see Table 4.1). The analysis revealed that valence ratings showed greater variability than arousal ratings, indicating less consensus in how pleasant or unpleasant stimuli were perceived. For example, B_HN exhibited the highest variance in valence, suggesting mixed perceptions among participants, while neutral

clips such as K_Ne and J_Ne showed the lowest variance, reflecting stronger agreement. In contrast, arousal ratings remained relatively stable across all stimuli, with only minor differences in variance. This highlights that while participants generally agreed on the arousal level of the clips, their interpretations of valence were more subjective and varied across stimuli.

These findings underscore the importance of considering both group-level patterns and individual variations when studying emotion processing, especially for training emotion detection models that account for such diversity, where we referred to this as self-reported arousal and self-reported valence. This variability can be partly attributed to differences in participants' emotional health, as noted in the literature review (Chapter 2), and could potentially be reduced by restricting the analysis to emotionally healthy participants.

To train an emotion detection model, it is essential to consider labelled emotions and physiological responses. As mentioned in previous chapters, physiological signals indicate emotional reactions. Incorporating these signals alongside labelled emotions would enhance the reliability of the emotion detection model.

To use physiological signals in an emotion detection model, we need to clean the data, extract relevant features, and analyse them to obtain reliable features. The following section presents the processing, feature extraction, and analysis results for physiological signals, including FER, pupil size, and GSR.

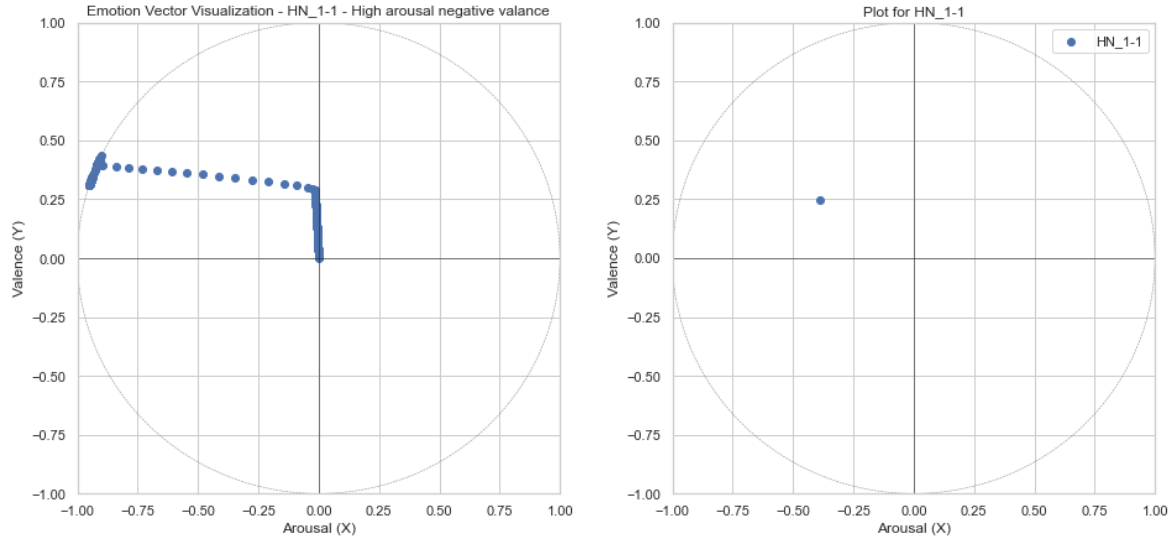
4.1.2 Results of FER Analysis

FER proved to be a valuable modality for characterising emotional responses in our study. While the raw outputs from iMotions (AFFDEX 2.0) provided reliable detection of the seven basic emotions, the strength of our approach lies in how these signals were subsequently utilised. Rather than treating the basic categories as isolated outcomes, we transformed them into continuous valence-arousal trajectories, enabling a more fine-grained representation of emotional dynamics across time.

The results demonstrate that this transformation produces interpretable emotional trajectories that capture subtle shifts in affective state, which would be obscured by categorical labels alone. In particular, the FER-derived valence and arousal coordinates aligned with expected quadrant distributions from Russell's circumplex model, supporting both the validity of the mapping and its capacity to represent co-occurring or mixed emotional states. This outcome highlights the advantage of our approach: by embedding FER outputs into a dimensional space, we bridge categorical recognition with continuous affect modelling, offering richer and more flexible input for multimodal fusion.

The results presented here show the vectorial representations of the basic emotions identified by FER. We analysed 20 audiovisual clips, each lasting between 30 and 58

seconds, during which FER data for basic emotions were recorded at each timestamp. Figures 4.2 and 4.3 illustrate the valence-arousal space representations of these emotions for participant ITA04, first at each timestamp and then as the final average vectorial result, which reflects the levels of arousal and valence for the respective stimuli. Figure 4.2 illustrates stimuli that belong to the high arousal, negative valence category

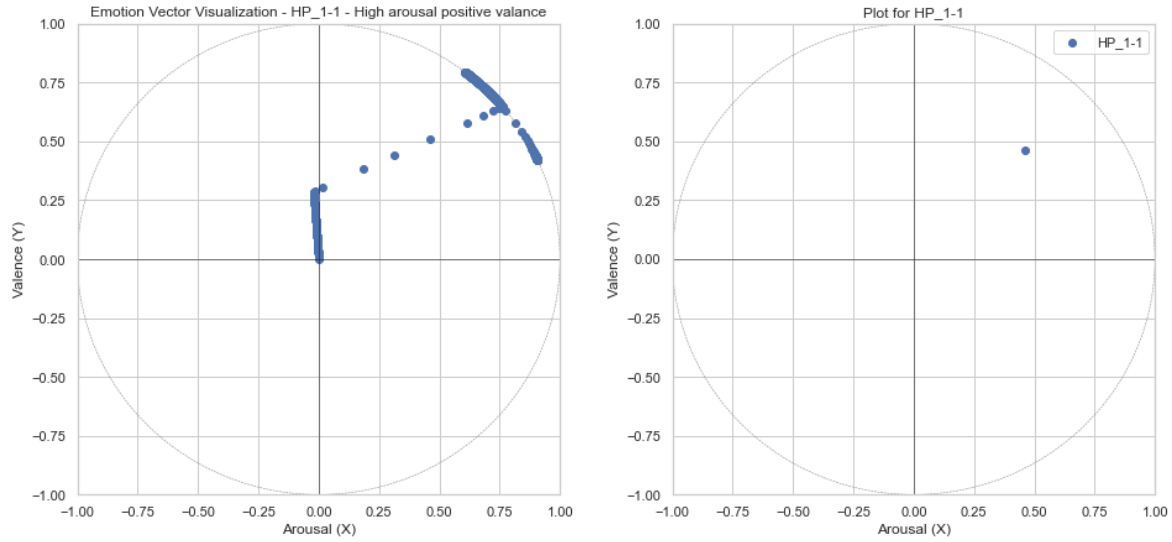


(a) (b)

Figure 4.2: FER vectorial representations (VR) for participant ITA04 and one clip for high arousal negative valence (HN). (a) VR at each timestamps of the clip, (b) average VR across timestamps.

for participant ITA04. As shown in the first part of the figure, the participant's facial expressions predominantly remained within the HN quadrant for most of the clip timestamps. Additionally, the average value at the end of the clip also falls within the HN quadrant, further validating the reliability of the vectorial representations produced by our method. Similarly, Figure 4.3 illustrates high-arousal, positive-valence stimuli for participant ITA04. As shown in the first part of the figure, the participant's facial expressions predominantly remain within the HP quadrant for most of the clip timestamps. Additionally, the average value at the end of the clip also falls within the HP quadrant, further validating the reliability of the vectorial representations produced by our method.

Next, we extracted statistical features of FER arousal and valence, including the mean, maximum, minimum, and standard deviation. We then analysed the Pearson correlation between these extracted features and the self-reported arousal and valence to examine the linear relationship between the features and the self-reported values. The correlation analysis between FER-derived features and self-reported emotional states (arousal and valence) was conducted for all 45 participants.



(a) (b)

Figure 4.3: FER vectorial representations (VR) for participant `ITA04` and one clip for high arousal positive valence (HP). (a) VR at each timestamps of the clip, (b) average VR across timestamps.

FER Features Correlation with Valence and Arousal

The top 10 FER features, ranked by mean Pearson correlation across participants, are shown in Figure 4.4. For **Valence**, the strongest positive correlations were observed for `FER_Mean_Arousal` ($r = 0.253 \pm 0.231$), `FER_Min_Arousal` ($r = 0.188 \pm 0.236$), and `FER_Max_Arousal` ($r = 0.179 \pm 0.251$), indicating that features associated with arousal-related expressions contributed more to perceived valence than the valence-specific features themselves. Conversely, some valence-related features, such as `FER_Kurtosis_Valence` ($r = -0.139 \pm 0.236$), showed small negative correlations, suggesting limited predictive value for valence in isolation.

For **Arousal**, the highest correlations were found in `FER_Std_Arousal` ($r = 0.191 \pm 0.245$) and `FER_Max_Arousal` ($r = 0.156 \pm 0.265$), highlighting that variability and peak intensity in arousal-related facial expressions are moderately associated with participants' reported arousal. Interestingly, certain valence-related features, such as `FER_Kurtosis_Valence` ($r = -0.128 \pm 0.247$), also contributed weakly, suggesting some cross-over influence between valence and arousal features.

The correlations are moderate at best, with many features showing low or near-zero mean correlations across participants. The error bars in Figure 4.4 indicate substantial inter-participant variability, highlighting that the relationship between FER features and subjective emotional ratings is not consistent across all participants. These findings suggest that while FER features provide some information about emotional state, single features alone are insufficient for accurate prediction of valence or arousal, and

a combination of multiple features or modalities may be necessary for robust emotion estimation.

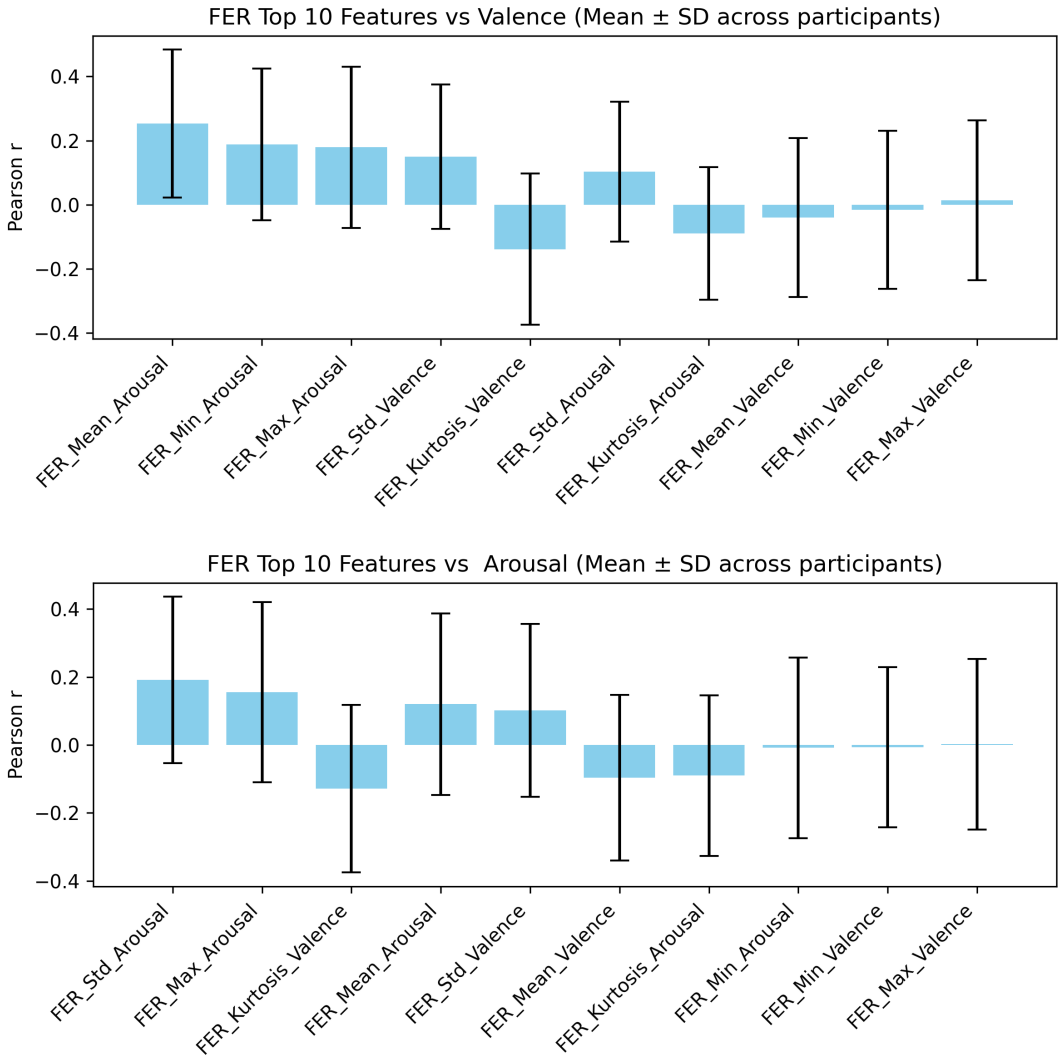


Figure 4.4: Top 10 FER features with the highest mean Pearson correlation (r) with **Valence** (top) and **Arousal** (bottom) across participants. Bars represent the mean correlation for each feature, and error bars indicate the standard deviation across participants. Features related to arousal tend to dominate the Valence correlations, while both arousal- and valence-related features contribute to Arousal correlations.

Key Observations and Next Steps.

- Facial expressions provide better predictive potential for arousal than valence, which can be because of less reactive participants or emotionally unhealthy participants, which can be solved by taking only emotionally healthy participants.
- Individual differences contribute to high variability, limiting the generalisability of the results.
- Further refinement, including multimodal integration (e.g., combining FER with

GSR), may enhance emotion detection performance.

The following section will explore methods to improve these correlations through feature selection, normalisation techniques, and the inclusion of additional physiological signals.

4.1.3 Results of Pupil Size analysis

Our study extracted statistical features from the pupil size data for each stimulus and the baseline (i.e., the grey screen). By normalising the pupil size relative to the baseline measurement, we could effectively isolate the impact of emotional arousal on pupil dilation.

Pearson correlations were calculated between various pupil size features and self-reported arousal and valence to investigate the relationship between pupil size dynamics and emotional states. These features included statistical descriptors (mean, minimum, maximum, skewness, kurtosis, standard deviation) computed separately for the pre- and post-normalisation periods.

Pupil Size Features Correlation with Valence and Arousal

The top 10 pupil size features, ranked by mean Pearson correlation across participants, are shown in Figure 4.5. For **Valence**, the strongest negative correlations were observed for `Pupil_before_mean_normalize` ($r = -0.583 \pm 0.150$), `Pupil_after_mean_normalize` ($r = -0.583 \pm 0.150$), and `Pupil_after_max_normalize` ($r = -0.581 \pm 0.131$), indicating that larger pupil sizes were associated with lower reported valence. Positive correlations were observed for `Pupil_before_kurtosis_normalize` ($r = 0.393 \pm 0.157$) and `Pupil_after_kurtosis_normalize` ($r = 0.393 \pm 0.157$), suggesting that the shape of the pupil distribution (kurtosis) may also carry relevant information about valence.

For **Arousal**, the highest correlations were more moderate. The top features included `Pupil_before_kurtosis_normalize` ($r = 0.153 \pm 0.200$) and `Pupil_after_kurtosis_normalize` ($r = 0.153 \pm 0.200$). Most other features showed weak negative correlations, such as `Pupil_before_mean_normalize` ($r = -0.116 \pm 0.182$), indicating that mean pupil size is slightly inversely associated with arousal. The magnitude of correlations for arousal is smaller than for valence, suggesting that pupil size features are more strongly related to valence than to arousal.

The error bars in Figure 4.5 indicate considerable inter-participant variability. These results imply that pupil size features provide meaningful but variable information about participants' emotional states, particularly for valence.

However, the overall correlations with arousal were weak, and even the valence-

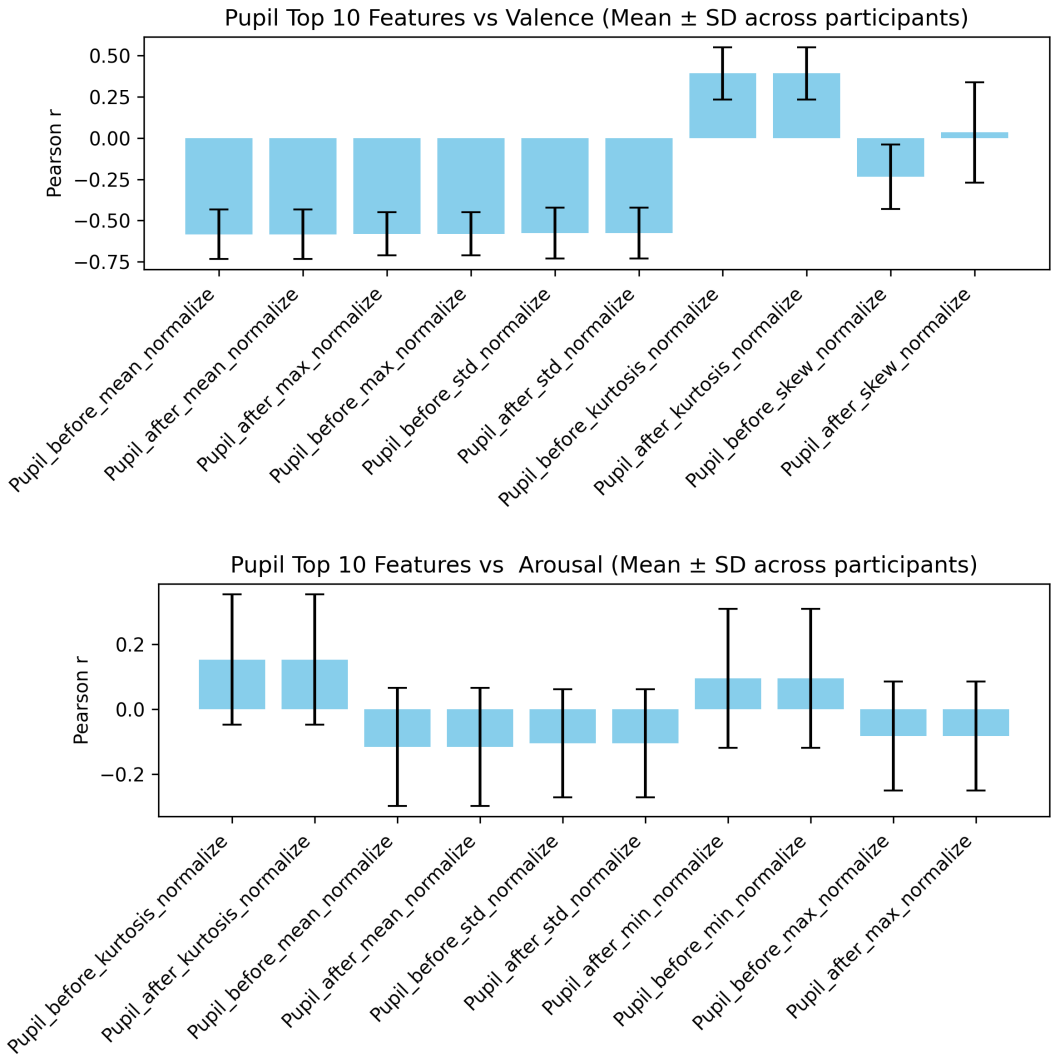


Figure 4.5: Top 10 pupil size features with the highest mean Pearson correlation (r) with **Valence** (top) and **Arousal** (bottom) across participants. Bars represent the mean correlation for each feature, and error bars indicate the standard deviation across participants. Features related to mean and maximum pupil size show strong negative correlations with Valence, while kurtosis-related features show positive correlations. Correlations with Arousal are generally smaller in magnitude.

related associations, though stronger, should be interpreted with caution. It is less likely that these effects are purely due to changes in luminance, as experimental conditions were controlled for light variability. Nonetheless, some residual influence of ambient or stimulus-related luminosity on pupil size cannot be ruled out entirely by subtracting grey screen pupil size. Therefore, it is essential to properly account for the effect of luminosity to understand the impact of emotion on pupil size.

In the following study, we aim to further minimise the confounding effects of lighting by incorporating real-time luminosity tracking. This can help refine the accuracy of pupil-based features in emotion detection models.

4.1.4 Results of GSR analysis

To evaluate the role of GSR in emotion detection, we conducted a series of statistical analyses on the extracted features.

Comparing Stimuli vs. Baseline

We began our analysis with a Wilcoxon signed-rank test to assess the APA of the phasic signal. APA is frequently utilised to measure short-term GSR responses to emotional stimuli across various stimuli in contrast to the baseline condition, which was represented by the grey screen, for all participants. This analysis aimed to determine whether the GSR responses to emotional stimuli differed significantly from those recorded during the neutral baseline condition. The Figure 4.6 shows the bar graphs of the mean APA comparisons of a few stimuli with the same baseline. A statistically significant increase in mean APA values was observed in the high-arousal negative-valence conditions—A_HN and W_HN. For A_HN, the APA mean rose from 0.08 during the baseline to 0.17 during the clip, with a p of 0.0015. Similarly, in the W_HN condition, the mean increased from 0.08 to 0.14 ($p = 0.0018$). These results indicate that these high-arousal, emotionally intense clips elicited a strong physiological response, consistent with the activation of the ANS typically associated with emotional arousal (e.g., fear, anger, stress).

In contrast, the C_LN condition (characterized by low-arousal negative valence such as sadness) and the J_Ne condition (neutral content) showed increases in GSR means—from 0.08 to 0.11 and 0.08 to 0.10, respectively—but these changes were not statistically significant ($p = 0.1172$ for C_LN; $p = 0.2029$ for J_Ne). This suggests that while there may be some physiological engagement, the GSR signal is less responsive to low-arousal or emotionally neutral stimuli.

These findings highlight GSR's sensitivity to high-arousal emotional stimuli, supporting its utility in detecting intense emotional states. However, they also point to a limitation in distinguishing more subtle or low-intensity emotions, indicating that GSR may be most effective when combined with other physiological or behavioural measures

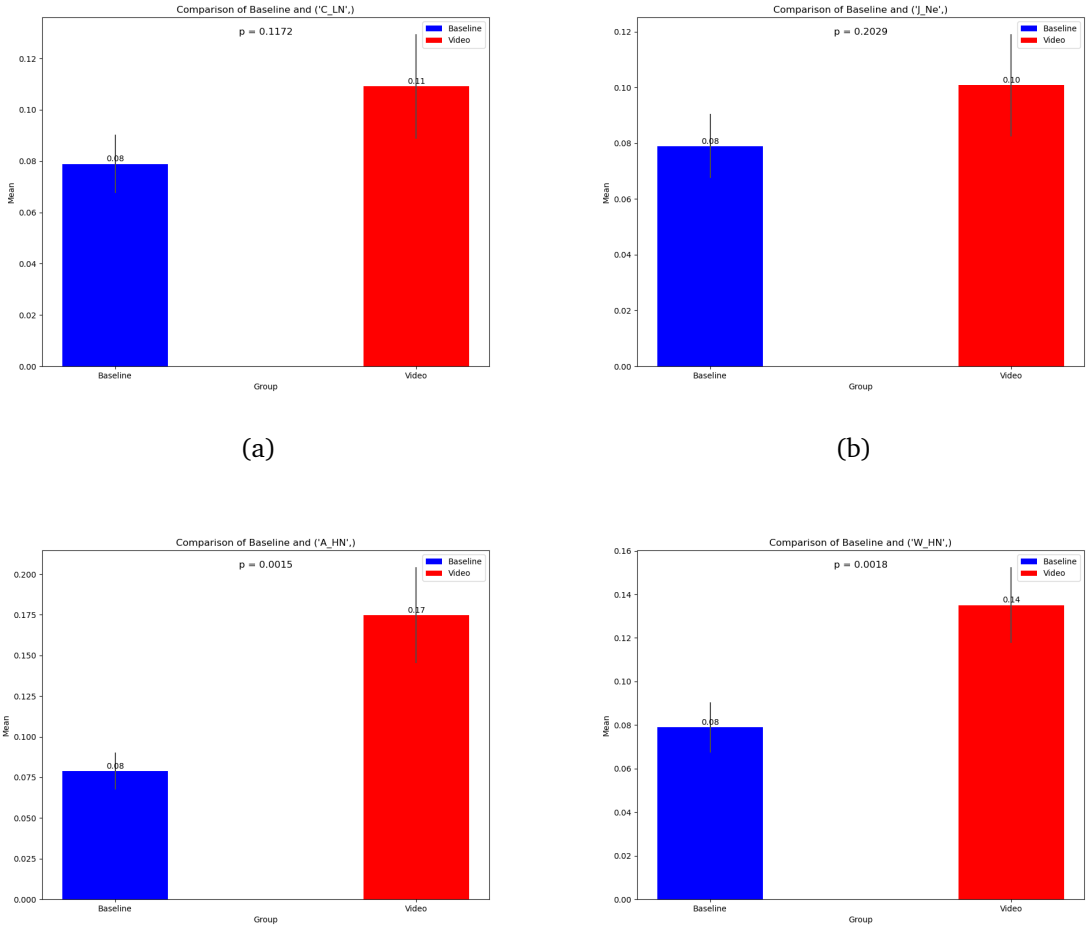


Figure 4.6: GSR Wilcoxon statistics results for few clips across participants.

in emotion detection systems.

Post-hoc Friedman Analysis: High vs. Low Arousal Groups

Next, we ran a post hoc Friedman test to see how the APA stacked between the high-arousal and low-arousal clip groups. We sorted the clips according to their self-reported arousal ratings, and the analysis looked into whether the GSR responses showed any significant differences between these two conditions. Based on the box plot of the statistical analysis for high- and low-arousal groups, shown in Figure 4.7, $p = 0.0311$, which is less than 0.05, suggests a significant difference between the APA values from high- and low-arousal data across all participants.

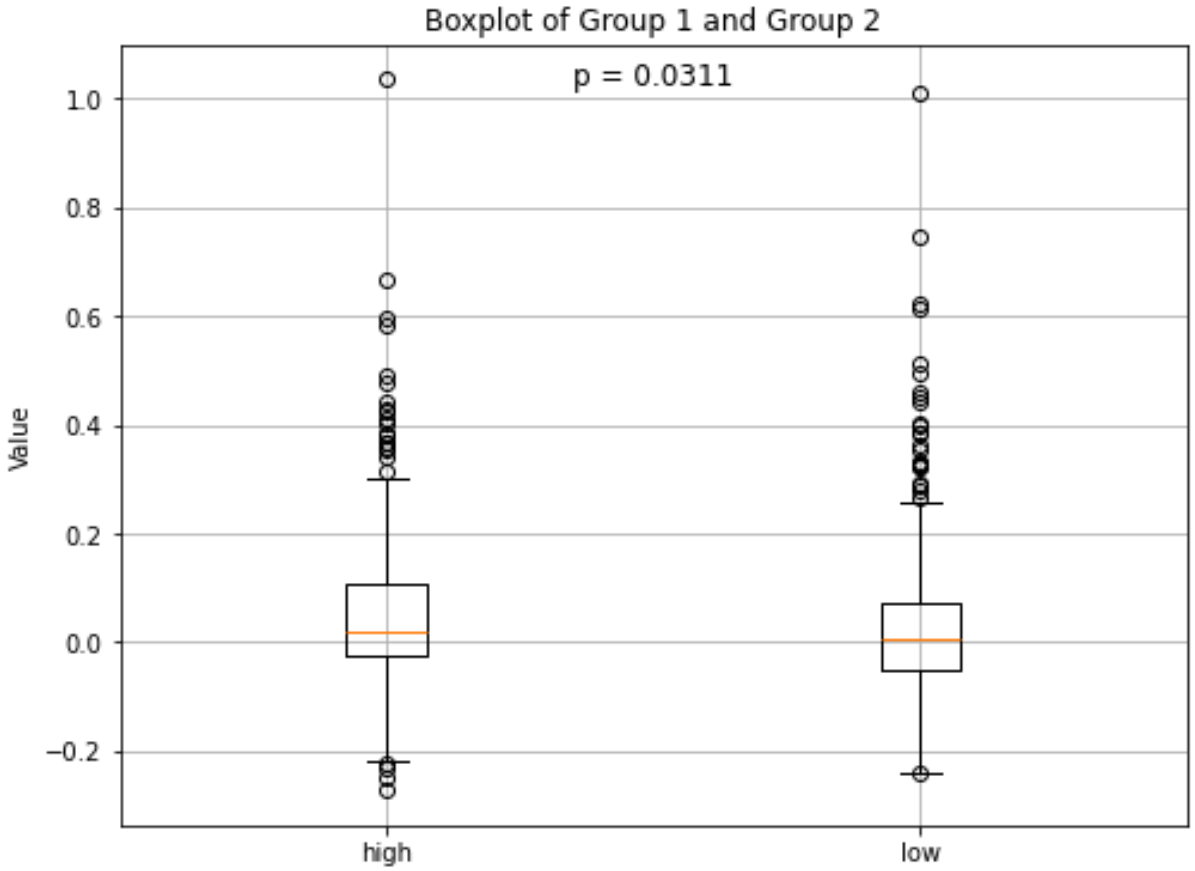


Figure 4.7: One-hoc Friedman statistics results for GSR average peak amplitude for arousal and valence groups.

GSR Features Correlation with Self-Reported Valence and Arousal

After verifying data using statistical analysis, we computed the average Pearson correlation and standard deviation across 45 participants to assess the relationship between GSR features and self-reported emotions. This correlation analysis aimed to understand the linear relationship between features derived from GSR and individuals' emotional responses.

The top 10 galvanic skin response (GSR) features, ranked by mean Pearson correlation across participants, are shown in Figure 4.8. For **Valence**, the strongest negative correlations were observed for `phasic_signal_before_peak_per_min_normalize` ($r = -0.448 \pm 0.083$) and `phasic_signal_after_peak_per_minute_normalize` ($r = -0.448 \pm 0.083$), indicating that higher phasic peak rates are associated with lower valence ratings. Positive correlations were observed for `phasic_signal_before_min_normalize` ($r = 0.124 \pm 0.267$) and `tonic_signal_after_kurtosis_normalize` ($r = 0.094 \pm 0.242$), suggesting that minimum or shape-related features of the GSR signal also carry some information about valence. Overall, the magnitude of cor-

relations varies, with substantial inter-participant variability reflected in the standard deviations.

For **Arousal**, the highest negative correlations were also seen in the phasic peak rate features, `phasic_signal_before_peak_per_min_normalize` ($r = -0.296 \pm 0.031$) and `phasic_signal_after_peak_per_minute_normalize` ($r = -0.296 \pm 0.031$). Other features showed smaller correlations, both positive and negative, with larger standard deviations, indicating weaker and more variable associations with arousal. These results suggest that GSR features are moderately informative for valence and somewhat less so for arousal.

The error bars in Figure 4.8 illustrate the variability across participants, emphasising that individual differences are substantial, particularly for features beyond the peak phasic responses.

Based on the analysis of 45 participants, the results indicate weak and inconsistent correlations between GSR features and self-reported emotions. While some phasic post-stimulus features, particularly peak-related measures, showed stronger associations with arousal, the overall predictive power remains low, with high inter-individual variability. Valence detection exhibited weaker correlations, with no apparent pattern or advantage of specific GSR features. These findings suggest that GSR alone may not be sufficient for robust emotion detection, highlighting the need for multimodal approaches and refined feature selection in affective computing. Moreover, the basic statistical features from GSR are insufficient to provide reliable emotional information. Therefore, more elaborate features are required.

Key Observations and Next Steps.

- **Weak Correlations**: GSR signals alone may not be sufficient for reliable emotion prediction.
- **High Variability Across Participants**: Standard deviations suggest strong individual differences in GSR-emotion relationships.
- **Phasic vs. Tonic Features**: Phasic features generally showed stronger associations with arousal, while valence correlations remained weak.

Given these findings, the next section will focus on improving predictive performance through feature engineering, normalisation strategies, and multimodal data integration to enhance emotion detection accuracy.

Comparison of Pearson, Spearman, and Mutual Information Across Modalities

To investigate the relationships between physiological and facial expression features and emotional dimensions, we computed three different metrics: Pearson correlation,

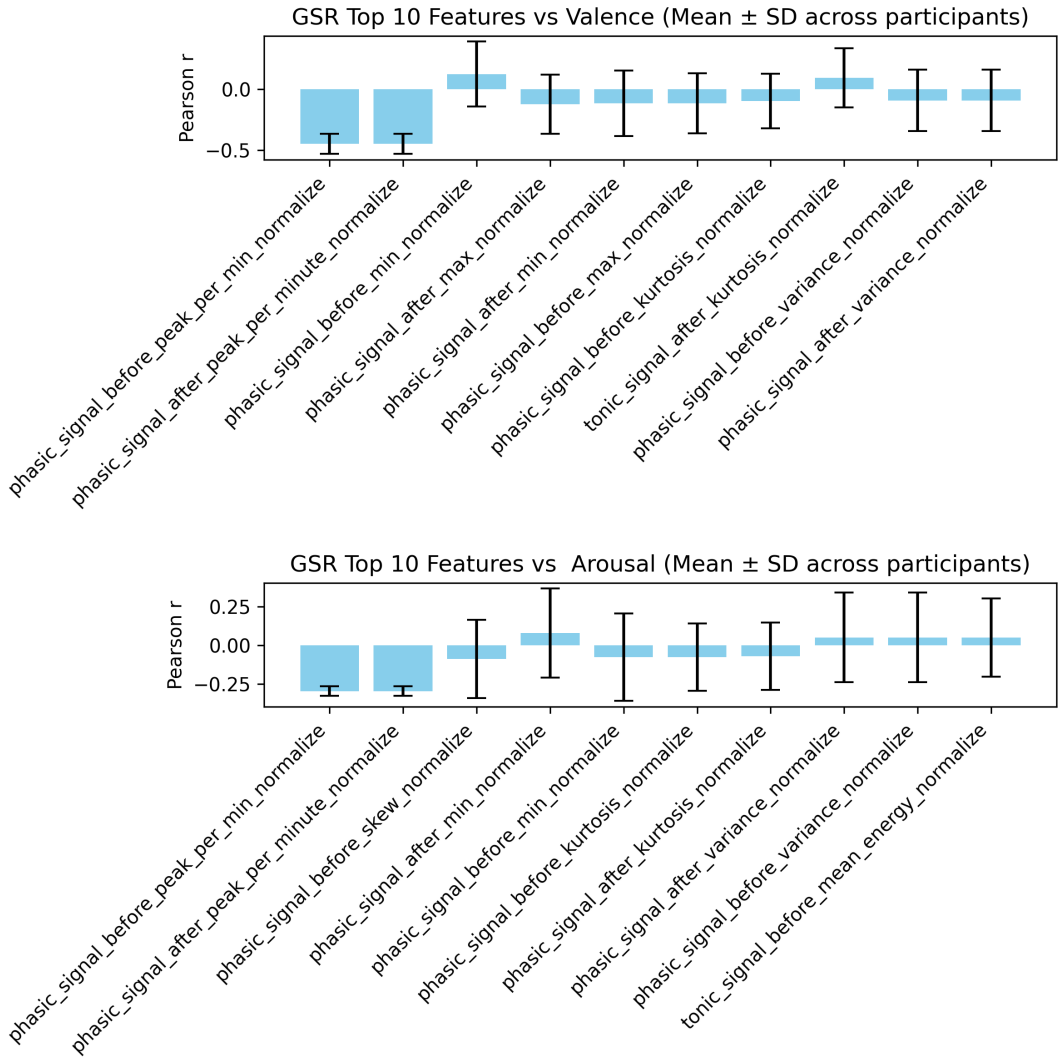


Figure 4.8: Top 10 GSR features with the highest mean Pearson correlation (r) with **Valence** (top) and **Arousal** (bottom) across participants. Bars represent the mean correlation for each feature, and error bars indicate the standard deviation across participants. Phasic peak rate features are strongly negatively correlated with both Valence and Arousal, while other features show weaker and more variable correlations.

Spearman rank correlation, and mutual information. Separate heatmaps were generated for each metric to visualise these associations across modalities (FER, GSR, Pupil) and targets (Valence and Arousal).

FER Features. For FER features, Pearson correlations revealed small but significant associations with Valence (Mean $r = 0.060 \pm 0.267$, $p < 0.001$) and weaker correlations with Arousal ($r = 0.024 \pm 0.273$, $p = 0.076$). Spearman correlations were smaller and mostly non-significant ($r = 0.017 \pm 0.284$ for Valence). Mutual information indicated the presence of non-linear associations (MI = 0.067 for valence, MI = 0.038 for arousal), which can coexist with linear relationships. This suggests that both linear and non-linear dependencies can be present, going beyond what linear or rank-based correlations alone can capture.

GSR Features. GSR features displayed small negative linear correlations with Valence ($r = -0.090 \pm 0.276$, $p < 0.001$) and Arousal ($r = -0.030 \pm 0.278$, $p = 0.001$). Spearman correlations were weaker, with Valence showing a non-significant trend ($r = -0.014 \pm 0.250$) and Arousal showing a weak positive correlation ($r = 0.021 \pm 0.268$, $p = 0.035$). Mutual information values were higher (MI = 0.106 for valence, MI = 0.078 for arousal), indicating the presence of both linear and non-linear associations that were not fully captured by linear correlations.

Pupil Size Features. For Pupil features, Pearson correlations indicated moderate negative associations with Valence ($r = -0.239 \pm 0.418$, $p < 10^{-31}$) and weak correlations with Arousal ($r = -0.015 \pm 0.215$, $p = 0.114$). Spearman correlations were mostly non-significant. Mutual information showed slightly higher values (MI = 0.087 for Valence, MI = 0.062 for Arousal), confirming that pupil size contains information about emotional state, particularly Valence, that may be linear or non-linear.

Pearson and Spearman correlations were generally small, with Spearman correlations being weaker, indicating limited monotonic relationships. Mutual information consistently yielded higher values, underscoring the presence of statistical dependencies between features and self-reported ground truth. These dependencies may include both linear and non-linear components, which would require further analyses to disentangle.

Figures 4.9, 4.10, and 4.11 show the heatmaps for Pearson, Spearman, and mutual information, respectively. These heatmaps illustrate that some features exhibit stronger non-linear relationships that are not captured by Pearson or Spearman correlation alone.

Across all three modalities (FER, GSR, and Pupil), and across multiple correlation metrics (Pearson, Spearman, and mutual information), the associations between in-

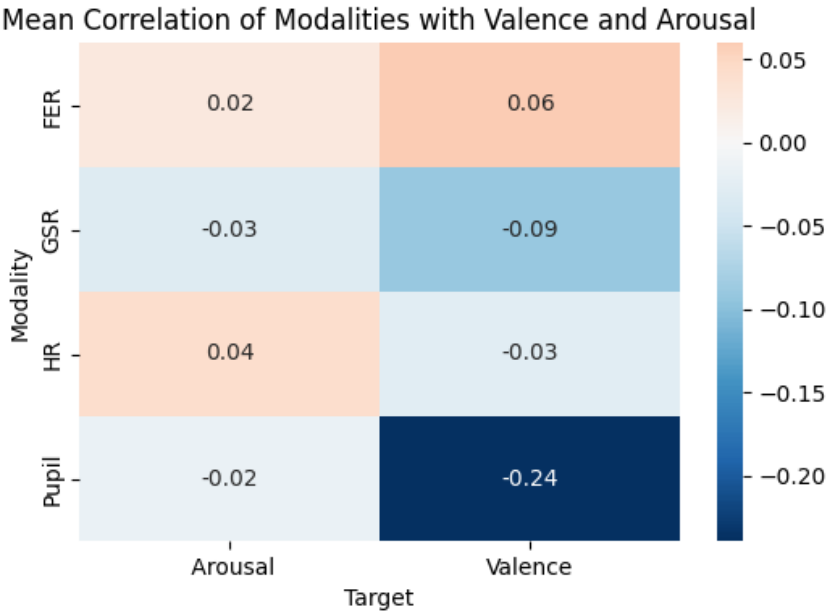


Figure 4.9: Heatmap of Pearson correlation coefficients between features and emotional targets (Valence and Arousal) for FER, GSR, and Pupil. HR is excluded.

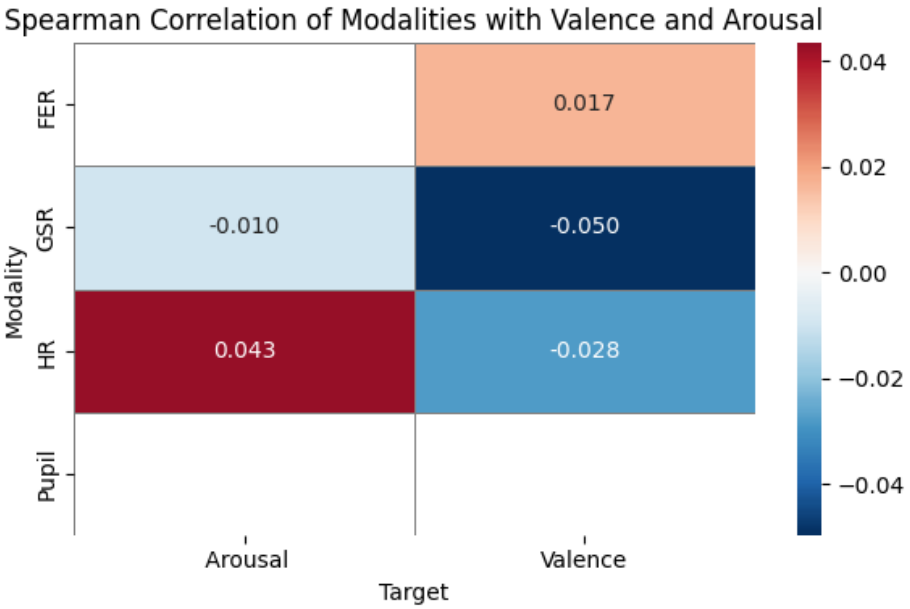


Figure 4.10: Heatmap of Spearman rank correlation coefficients between features and emotional targets (Valence and Arousal) for FER, GSR, and Pupil. HR is excluded.

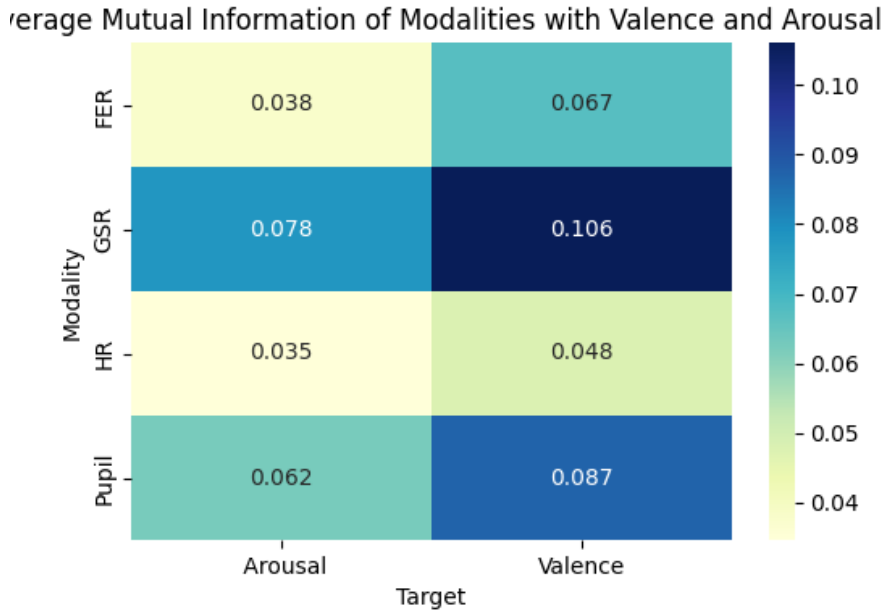


Figure 4.11: Heatmap of mutual information between features and emotional targets (Valence and Arousal) for FER, GSR, and Pupil. HR is excluded.

dividual features and emotional targets (Valence and Arousal) were generally weak or inconsistent across participants. While mutual information revealed moderate dependencies, these did not translate into robust predictive power. This underscores the limitations of relying on single features for emotion estimation and motivated the design of our main study, which integrates multiple features and modalities within a machine learning framework capable of capturing both linear and non-linear relationships between predictors and self-reported affective states.

4.1.5 Lessons Learned from Pilot Study and Modifications for Main Study

Several challenges from the pilot study informed critical methodological improvements for the main study. To ensure accurate emotion detection, participants with psychological conditions like anxiety, depression, or alexithymia were excluded through a screening process, as we noticed high variability between individual emotion labelling. Facial expressions were found unreliable due to inexpressiveness, prompting a shift toward more dependable physiological signals. Pupil size data were corrected for stimulus luminosity using a custom isolation model, ensuring that lighting did not confound emotional changes. Basic statistical features and generic baseline correction proved insufficient for GSR, leading to more refined feature extraction methods. Additionally, we integrated FER, GSR, and pupil size to improve the robustness of emotion detection. These refinements strengthened the foundation of our main study, leading to a more

accurate and reliable emotion detection model.

4.2 Results of the Main Study

This chapter presents the findings from the main study, which aimed to predict self-reported emotional states—arousal and valence—using features extracted from multiple physiological modalities: FER, pupil size, and GSR.

The results are organised to reflect both unimodal and multimodal analysis. We first report the model performance for each modality independently, followed by the outcomes from feature-level fusion, where all extracted features were integrated into a unified representation. Each model was evaluated using LOPO cross-validation combined with five-fold internal validation to ensure robust estimation and generalisability.

Model performance was assessed using standard regression evaluation metrics, including $R2$-score, Pearson correlation coefficient, and $NRMSE$. These results offer insights into the predictive capacity of each modality and the effectiveness of multimodal integration for emotion detection.

4.2.1 Results of Emotion Labelling

As mentioned in the methodology chapter, to categorise emotional responses, each audiovisual clip was mapped into a two-dimensional Valence-Arousal space, where valence indicates the positivity or negativity of the emotion and arousal reflects the intensity (from calm to excited). Based on this mapping, clips were classified into four quadrants:

- High Arousal Positive Valence (HP)
- High Arousal Negative Valence (HN)
- Low Arousal Positive Valence (LP)
- Low Arousal Negative Valence (LN)

This classification is visualised in two comparative plots—one based on INDSCAL and the other using FA.

Comparison Between INDSCAL and FA.

1. INDSCAL – Group Space As shown in the Figure 4.12, INDSCAL provided a more balanced and circular distribution of clips across the Valence-Arousal space. This structure aligns with Russell's Circumplex Model of Affect, which proposes that emotions are arranged circularly around the valence-arousal axes. Each quadrant—HP (yellow dots), HN (red dots), LP (green dots), and LN (blue dots)—is well-represented with a relatively even spread, supporting the intended design of the stimulus set.

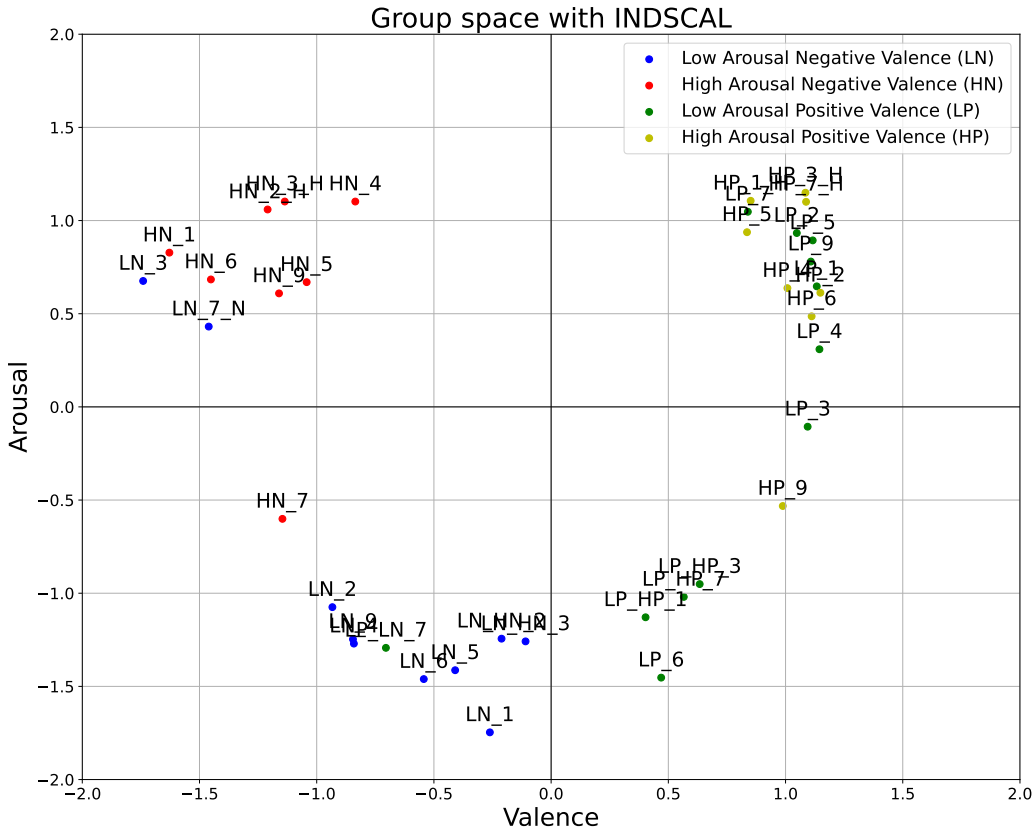


Figure 4.12: Aggregate ground truth responses across all participants using INDSCAL, categorised by stimulus arousal and valence: H = High Arousal, L = Low Arousal, P = Positive Valence, N = Negative Valence. Each point on the graph corresponds to a stimulus, i.e., a video clip. The valence and arousal values are rescaled in the range [-2, 2], where 0 indicates a neutral, average value.

2. FA – Group Space In contrast, the Figure 4.13 shows the distribution derived from FA, which appears more distorted and clustered, particularly around certain regions. The clip separation in terms of arousal and valence was less distinct, and the circular pattern of emotional mapping was less pronounced. This lack of clarity in quadrant separation indicates that FA was less effective for our goal of affective space representation

The comparison between the two methods supports the use of INDSCAL over FA for emotion labelling. INDSCAL not only preserved the individual differences in responses but also resulted in a more interpretable and theoretically sound emotional space, aligning with the circumplex model.

Additionally, to evaluate the effectiveness of INDSCAL versus Factor Analysis (FA) for deriving valence-arousal representations, we performed statistical comparisons on the participant-level data (see Table 4.2). The results indicate that INDSCAL provided

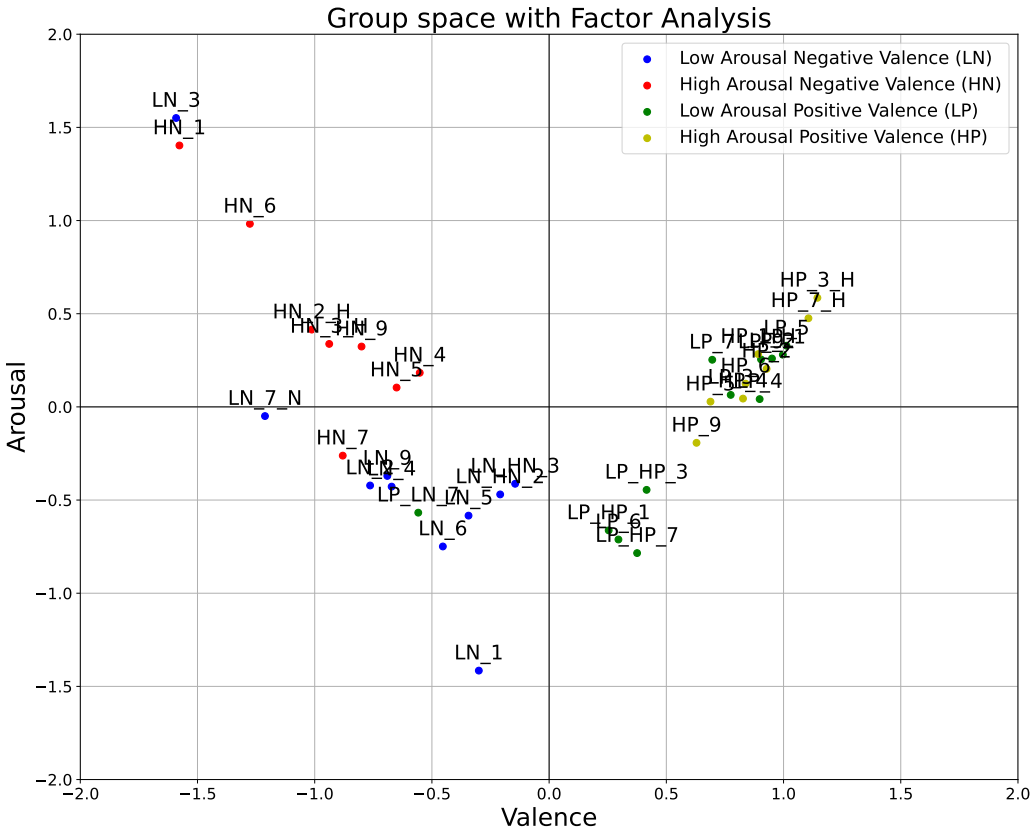


Figure 4.13: Group space for emotional labelling for the main study using FA.

a clearer separation of stimuli into the four quadrants of Russell's circumplex model. Specifically, INDSCAL achieved a higher silhouette score (0.455 vs. 0.402) and a lower Davies-Bouldin index (0.766 vs. 0.802), demonstrating superior cluster separability. Furthermore, INDSCAL yielded substantially lower within-stimulus variance across participants, both for valence (0.021 vs. 0.271) and arousal (0.004 vs. 0.690). This highlights its ability to map stimuli more consistently across individuals. Overall, these findings show that INDSCAL offers a more robust and stable two-dimensional emotional representation compared to FA, justifying its selection as the primary labelling method in this study.

Metric	INDSCAL	FA
Silhouette Score (↑)	0.455	0.402
Davies-Bouldin Index (↓)	0.766	0.802
Valence Variance (↓)	0.021	0.271
Arousal Variance (↓)	0.004	0.690

Table 4.2: Comparison of clustering quality and consistency metrics between INDSCAL and FA. Higher silhouette scores and lower Davies-Bouldin and variance values indicate superior performance.

Following the emotional labelling process, we proceeded with data pre-processing, feature extraction, model training, and evaluation for each modality (FER, pupil size, and GSR). The results of these analyses are presented in the subsequent subsections.

4.2.2 Results of Pupil Size Analysis for the Luminosity Effect Prediction Model (LEPM)

The model illustrated in the section "A. Development of Luminosity Effect Prediction Model" effectively predicts pupil size based on the RGB intensity values of the images on the screen in both dark and bright environments. The calibration procedure described above ensures that the method is flexible and adapted to inter-subjective differences and the settings and type of screen used. As explained in section 3.2.6, we initially used two different approaches, named "grey-based" and "colour-based", and then we used a method consisting of combining both approaches, called the "combined" approach. We assessed each approach using leave-one-image-out cross-validation, i.e., training the model on 27 images, eliminating one image at a time, and predicting the pupil size measured when the eliminated image was shown.

The results obtained from the 18 participants in the dark laboratory are summarised in Table 4.3. These include the mean Pearson correlation coefficient, the mean and maximum p-values (derived from significance tests comparing predicted and actual values), the mean R^2, the mean NRMSE, and the mean percentage error. The mean p represents the average statistical significance across all participants, while the maximum p highlights the least significant case, demonstrating that the results remain highly significant even in the weakest instance.

Table 4.4, on the other hand, shows the results obtained by aggregating the data across 18 participants.

The results for the "combined" method were better than when using only the "grey-based" method or only the "colour-based" method, as shown in tables 4.3 and 4.4, demonstrating that the two methods contribute synergistically to the evaluation of pupil size. Since the best results were achieved with the "combined" method, we used this

Table 4.3: Results of all methods on monochrome images in a dark laboratory - average across participants. (mean p and max p = mean p-value and maximum p-value, respectively.)

Method	**Correlation**	$R2$**-score**	**NRMSE**	**Average Error**
Colour-Based	0.62 $\pm$ 0.124 (mean p = 0.0038, max p = 0.0214)	0.40 $\pm$ 0.153	0.23 $\pm$ 0.048	6.52 $\pm$ 0.221 %
Grey-Based	0.72 $\pm$ 0.150 (mean p = 0.0056, max p = 0.086)	0.55 $\pm$ 0.193	0.21 $\pm$ 0.055	8.76 $\pm$ 0.290 %
Combination of colour and grey-based approach	0.84 $\pm$ 0.070 (mean p = $< 10^{-7}$, max p = 0.00001)	0.70 $\pm$ 0.114	0.13 $\pm$ 0.028	6.75 $\pm$ 0.266 %

Table 4.4: Results of all methods on monochrome images in a dark laboratory - aggregating the data from all the participants

Method	**Correlation**	$R2$**-score**	**NRMSE**	**Average Error**
Colour-Based	0.82 (p $< 10^{-7}$)	0.67	0.10	8.62 $\pm$ 0.79 %
Grey-Based	0.85 (p $< 10^{-7}$)	0.73	0.09	7.89 $\pm$ 0.80 %
Combination of colour and grey-based approach	0.94 (p $< 10^{-7}$)	0.88	0.06	4.62 $\pm$ 0.62 %

method exclusively from that point forward, and all subsequent results were obtained using this approach.

As explained in the paragraph `Testing the Luminosity Effect Prediction Model (LEPM) in a well-lit laboratory and with non-monochrome images`, we finally tested the LEPM and calibration procedure on non-monochrome, non-primary colour images and in both a dark and a lit environment. The test was conducted with 10 participants using a video composed of 27 monochrome images and 46 non-monochrome, emotionally neutral images with different colours and brightness.

Figures 4.14 and 4.16 show the results for the participant `IFL3` in a dark and a well-lit laboratory, respectively. The average results across participants, obtained in the dark and well-lit laboratories, are shown in Table 4.5.

Figures 4.15 and 4.17 and Table 4.6 show the results obtained by aggregating the data from all participants.

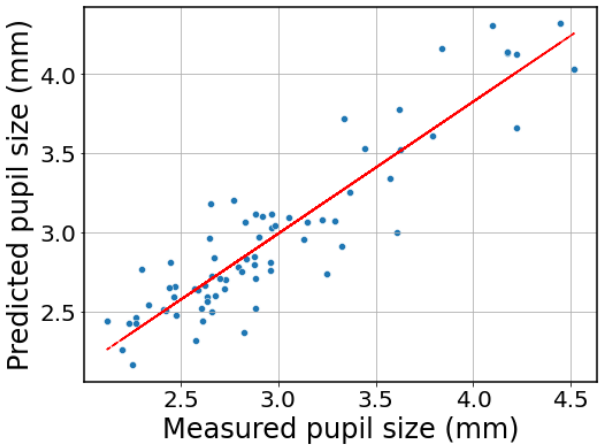


Figure 4.14: Relationship between measured and predicted pupil size in a dark laboratory for Participant `IFL3`, with correlation: 0.91 ($p < 10^{-7}$), $R2$-score: 0.83.

Table 4.5: Validation results of LPEM in dark light and well-light laboratory across all participants.

Method	**Correlation**	$R2$**-score**	**NRMSE**	**Average Error**
Dark light laboratory	0.84 ± 0.061 (mean p $< 10^{-7}$, max p $< 10^{-7}$	0.70 ± 0.101	0.12 ± 0.020	$7.58\% \pm 1.61\%$
Well-light laboratory	0.76 ± 0.045 (mean p $< 10^{-7}$, max p $< 10^{-7}$	0.58 ± 0.680	0.14 ± 0.014	$7.72\% \pm 1.66\%$

The average error between the two conditions is about 0.2%, and the difference in RMSE is about 0.02, highlighting the model's effectiveness across dark and well-lit settings. Additionally, the model performs consistently across various monitor brightness

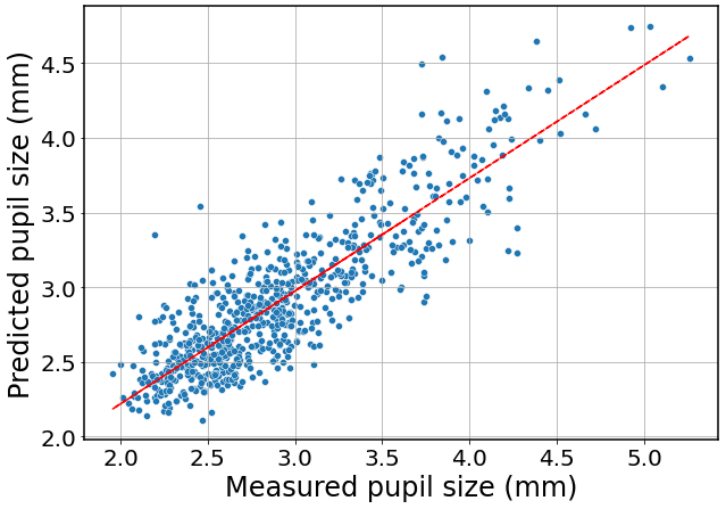


Figure 4.15: Measured and predicted pupil size in a dark laboratory for all the participants, with correlation: 0.87 ($p < 10^{-7}$), $R2$-score: 0.76.

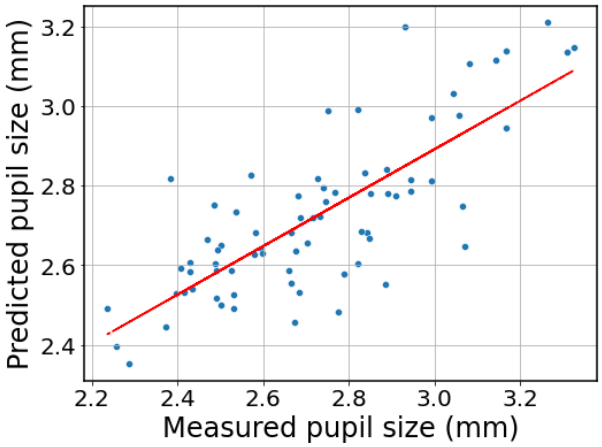


Figure 4.16: Measured and Predicted pupil size in a well-lit laboratory for the Participant `IFL3`, with correlation: 0.78 ($p < 10^{-7}$), $R2$-score: 0.61.

Table 4.6: Validation results of LPEM in dark light and well-light laboratory by aggregating data from all participants.

Method	**Correlation**	$R2$**-score**	**NRMSE**	**Average Error**
Dark light laboratory	0.87 (mean $p < 10^{-7}$, max $p < 10^{-7}$	0.76	0.09	7.28% $\pm$ 0.82%
Well-light laboratory	0.82 (mean $p < 10^{-7}$, max $p < 10^{-7}$	0.68	0.11	7.50% $\pm$ 0.83%

and contrast settings, as well as under different ambient lighting levels. Therefore, there was no need to modify the model to account for changes in environmental luminosity, as our calibration procedure effectively takes this into account. However, we repeated the calibration every 20 minutes for very long experiments.

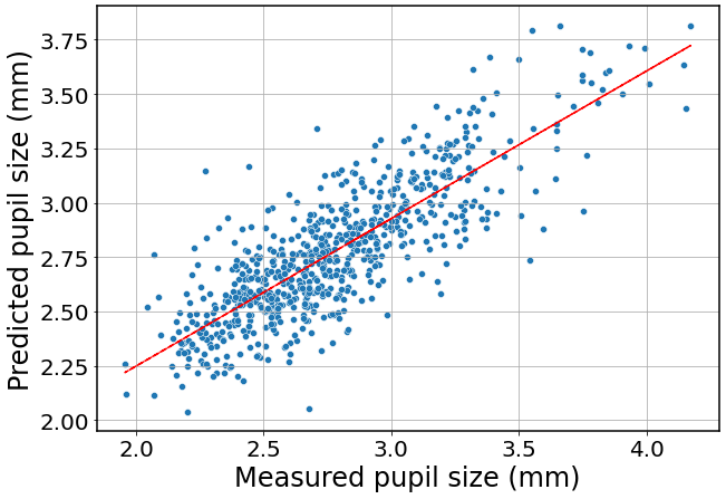


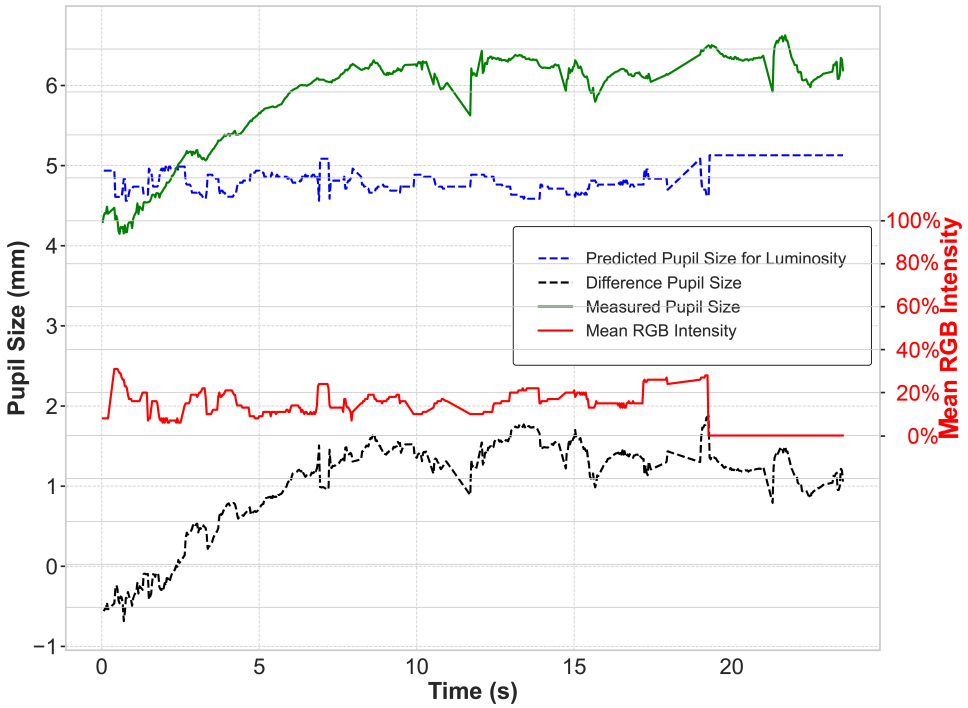
Figure 4.17: Measured and Predicted pupil size in a well-lit laboratory for all the participants, with correlation: 0.82 ($p < 10^{-7}$), $R2$-score: 0.68.

4.2.3 Results of Pupil Size Analysis for the Arousal Detection Model (ADM)

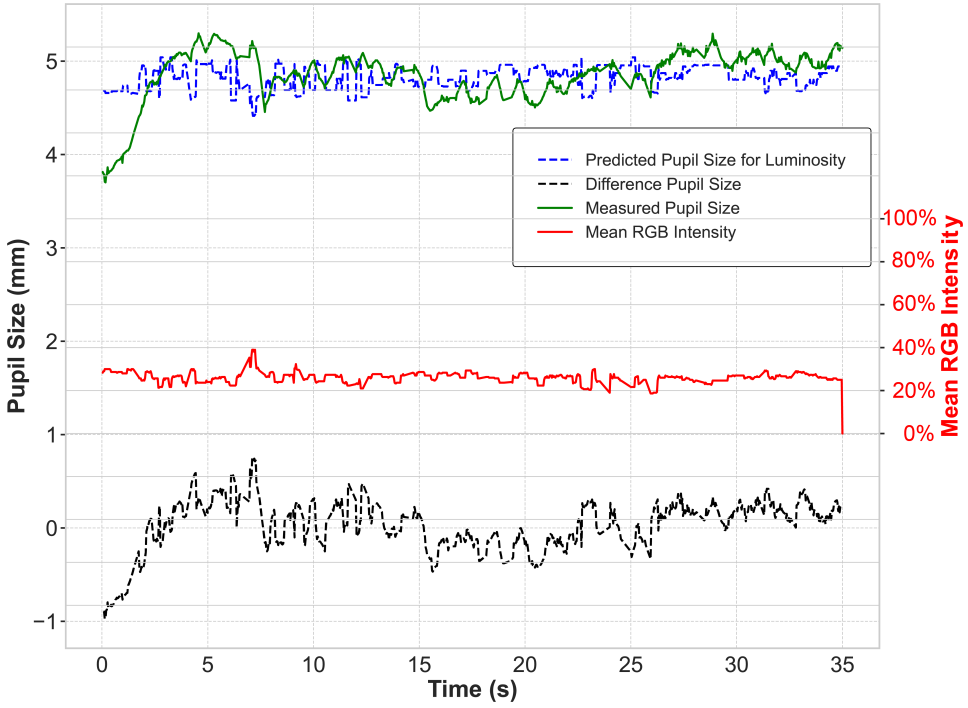
To test the ADM, we predicted the arousal of a participant watching audiovisual clips based on the measured pupil size and compared it with the ground truth. For each participant and each audiovisual clip, we plotted the value of the measured pupil size PS_{measured} throughout the audiovisual clip. This is shown in Figure 4.18 (green line) for a particular participant, and a clip with increasingly high emotional intensity (Figure 4.18(a), and one with low intensity (Figure 4.18(b)). In the figure, we show the average RGB value frame by frame, representing the luminous intensity of the clip (red line). We then calculated the pupil size component due to luminosity $PS_{\text{luminosity}}$ using the LEPM (see Figure 4.18, blue line). Finally, we calculated the pupil size component due to arousal PS_{arousal} by subtracting $PS_{\text{luminosity}}$ from PS_{measured}, according to Equation 3.24 (see the black line in Figure 4.18).

Our analysis revealed a clear distinction in the arousal responses between high-arousal and low-arousal video clips, as shown in the example illustrated in Figure 4.18. In fact, in the left panel (high arousal), the pupil size component due to arousal PS_{arousal} (black line) assumes higher values than in the right panel (low arousal). This difference validates the model's ability to separate the arousal component of pupil size changes from luminosity effects, supporting its application in accurately identifying arousal levels across different emotional stimuli.

We then calculated the average of the PS_{arousal} (black line in Figure 4.18) in the salient intervals, as explained in the paragraph `Model Testing` of section 3.2.6. For example, for the videos represented in Figure 4.18, the salient intervals were [0s, 20s] for video clip (a) and [0s, 10s] for video clip (b). Finally, we plotted the pupil size cor-



(a) High-Arousal



(b) Low-Arousal

Figure 4.18: Plot of participant XPI3pA's response for selected high-arousal (a) and low-arousal (b) videos, showing measured pupil size (green), predicted pupil size (blue), average RGB intensity (red), and arousal-induced pupil size (black).

rected for luminosity PS_{arousal} versus the self-reported arousal $\text{Arousal}_{\text{self-reported}}$ for all video clips and each participant, which is shown in Figure 4.19 for participant XPI3pA,

where the red circles correspond to each video clip (see figure 4.19 for more information). For that participant, we obtained a correlation of 0.71 (p = 1.684e-06); see the red line. For the measured pupil size, non-corrected for luminosity PS_{measured}, the correlation with the ground truth arousal was much worse, e.g., see for participant `XPI3pA` in Figure 4.19, where the blue dots correspond to each video clip. We obtained a correlation of 0.01 (p = 0.971) for that participant (see the blue line in Figure 4.19).

Table 4.7 shows all participants' average results. Correcting for luminosity dramatically increased the correlation compared to the non-corrected pupil size. The most surprising result, however, is that `pupil size had no predictive power for arousal without correcting for luminosity` since its correlation with self-reported arousal was not significantly different from zero (mean p = 0.2283).

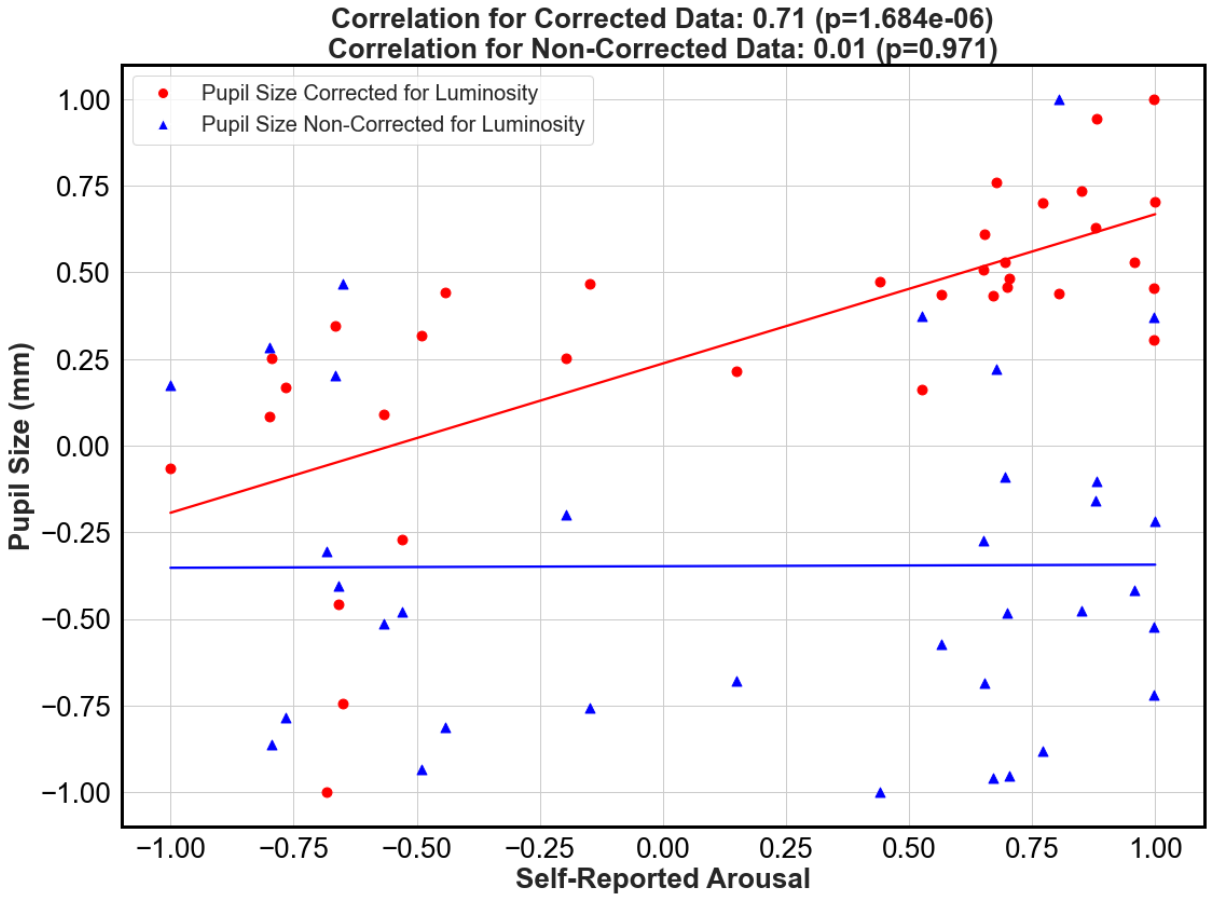


Figure 4.19: Pupil size comparison with and without luminosity correction versus self-reported arousal for Participant `XPI3pA`. The red circles represent pupil size corrected for luminosity with the red linear regression (LR) line, while the blue triangles indicate non-corrected pupil size with the blue LR line.

To further test the model's performance, we computed the predicted arousal $\widehat{\text{Arousal}}_{\text{self-reported}}$ utilizing the equations 3.25 and 3.26 and using a leave-one-participant out cross-validation as explained in the paragraph `ADM Testing`. We plotted the predicted arousal against the ground truth one $\text{Arousal}_{\text{self-reported}}$ both with and without

Table 4.7: Relationship between Self-Reported Arousal and Pupil Size with and without Correction for Luminosity

Metrics	**Corrected for Luminosity**	**Non-Corrected for Luminosity**
Correlation	0.65 ± 0.106 (mean p = 0.0025, max p = 0.096)	0.26 ± 0.150 (mean p = 0.2283, max p = 0.971)
NMRSE	0.27 ± 0.036	0.42 ± 0.054
$R2$	0.436 ± 0.125	0.09 ± 0.089

correction for luminosity, as shown in Figure 4.20 for the participant `XPI3pA`.

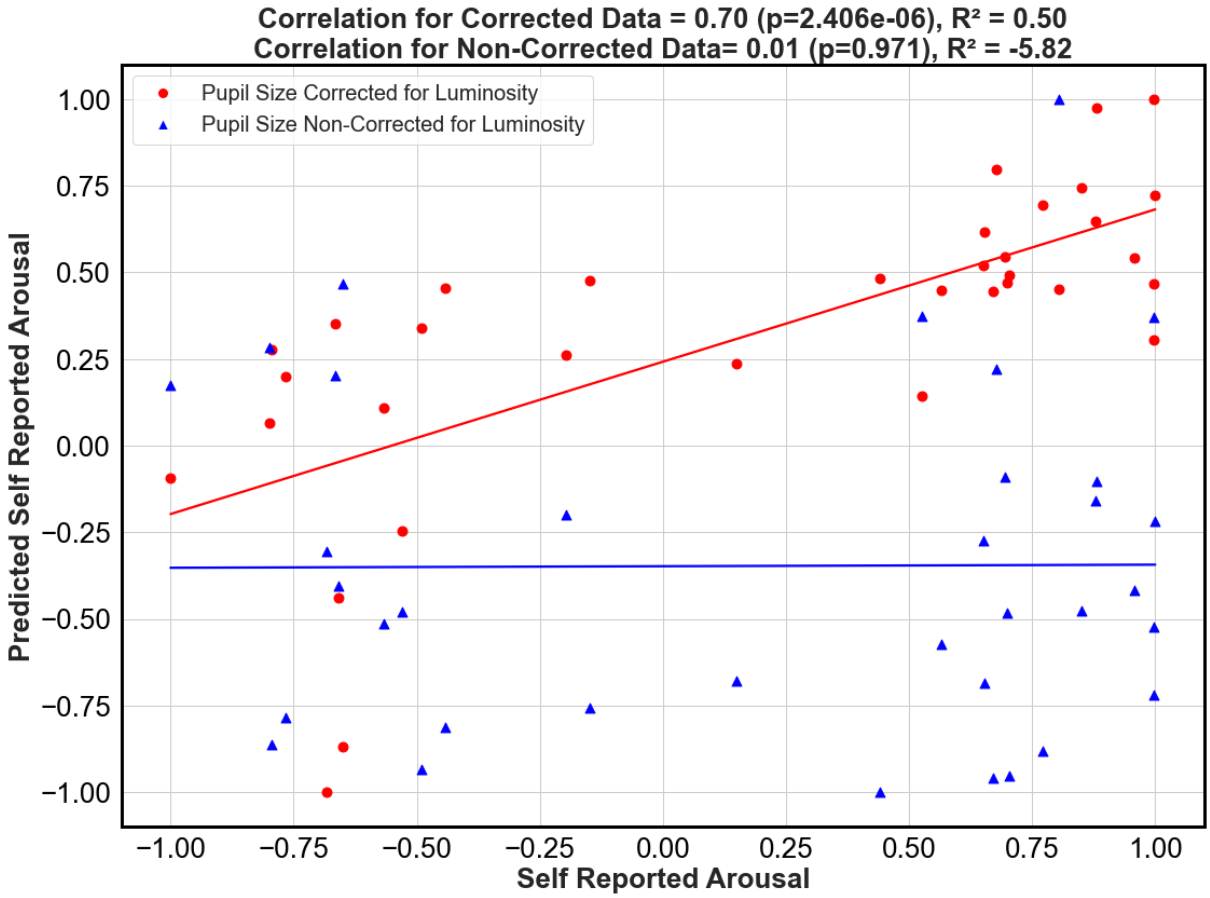


Figure 4.20: Predicted arousal versus self-reported arousal with and without the use of correction for the luminosity for Participant `XPI3pA`.

The average results across all the participants are shown in the left panel in Table 4.8. Again, correcting for luminosity yields a significant increase in correlation compared to the uncorrected pupil size. While the left panel displays the results obtained by calculating self-reported arousal using individual scaling, the right panel presents the results obtained by calculating self-reported arousal using FA. As mentioned in the paragraph `Development of Arousal Detection Model`, INDSCAL works much better than

FA in this case.

Table 4.8: Relationship between predicted and self-reported arousal with and without correction for luminosity using INSCAL and FA.

INDSCAL			**FA**		
Metrics	**Corrected for Luminosity**	**Non-Corrected for Luminosity**	**Metrics**	**Corrected for Luminosity**	**Non-Corrected for Luminosity**
Correlation	0.65 $\pm$ 0.12 (mean p = 0.0025, max p = 0.096)	0.26 $\pm$ 0.15 (mean p = 0.2283, max p = 0.971)	Correlation	0.33 $\pm$ 0.12 (mean p = 0.1397, max p = 0.367)	0.11 $\pm$ 0.15 (mean p = 0.3466, max p = 0.567)
$R2$-score	0.43 $\pm$ 0.12	0.09 $\pm$ 0.089	$R2$-score	0.11 $\pm$ 0.12	0.07 $\pm$ 0.09
NMRSE	0.50 $\pm$ 0.21	2.15 $\pm$ 1.11	NMRSE	1.12 $\pm$ 0.73	1.75 $\pm$ 0.95

Finally, we noticed that, in the case of pupil size corrected for luminosity, the average values of the coefficients of Equation 3.25 were a = 0.3463 $\pm$ 0.0551, b = -0.0126 $\pm$ 0.00038 and the fit was quite good $R2$ = 0.567 $\pm$ 0.073). The variance of the coefficients and $R2$ is since leaving one participant out at a time changes the slope and the intercept of the regression line. However, this variation is minuscule, as the standard deviation is significantly smaller than the mean, indicating that our sample was sufficiently large. In the case of pupil size non-corrected for luminosity, the average values of the coefficients were a = 0.36723 $\pm$ 0.06341, b = 3.8296 $\pm$ 0.0120, and the fit was inferior $R2$ = 0.023 $\pm$ 0.007).

This initial categorisation allowed us to explore a series of emotional responses among the participants, even if not all of them agreed with our initial categorisation. For example, given the subjectivity of emotional responses, some participants may have considered a clip intended to arouse fear boring. What was important was that participants answered honestly so that the machine could learn to associate specific values of the bio signals with the corresponding emotional states.

4.3 Testing models developed by other researchers with our data

4.3.1 Testing Nakayama's Hyperbolic Model with our data

Nakayama and colleagues [274] found a hyperbolic relationship between luminosity and pupil size, which they called standardised pupil size:

$$f(x) = \frac{219.04}{x + 175.34} + 12 \tag{4.1}$$

where x is the luminosity.

The function f should predict the variation in pupil size with luminosity to within a constant multiplicative factor. The authors, therefore, hypothesised that if they use emotionally neutral images, dividing the measured pupil size by the function f should yield a constant factor that they called the "compensated pupil size". If, on the other hand, images with emotional content are used, a variation is observed in the constant value, and this variation is attributed to arousal. We have tested this model on our 73 emotionally neutral test images and obtained a coefficient of variation of 0.09 $\pm$ 0.025. Our exponential model works better, obtaining a coefficient of variation of 0.03 $\pm$ 0.006.

We then used Nakayama's Hyperbolic Model with our emotional video clips. We used the calibration procedure to calculate the only coefficient of the model, consisting of the ratio between the measured pupil size and the standardised pupil size. The comparison with our method is presented in the table 4.9. Nakayama's method worked less well than ours. For example, there is a worsening of the $NRMSE$ of 51.8%.

4.3.2 Testing Linear Models with our data

Both Raiturkar et al. [276] and Asano et al. [277], [278] have developed dynamic linear models that take into account the temporal response of pupil dilation. We did not develop dynamic models because we took the average pupil size over relatively long intervals (at least 5 seconds). However, we wanted to test linear models with our data and obtained very poor results, as shown in the table 4.9.

4.3.3 Testing our model without self-reported arousal

As mentioned, where possible, we applied to our data the methods developed by other researchers to exploit the fact that we had recorded the self-reported arousal for each participant and each video, something that, to the best of our knowledge, no other research on the effect of luminosity on pupil size measurement has done. We investigated the importance of including self-reported arousal. To achieve this, we employed the approach adopted by other researchers: asking 10 independent judges (new participants) to assign an arousal value to each video (e.g., Raiturkar et al. [276]). For each clip, we took the average value. We then used that arousal value in our method, rather than the self-reported one, to observe how much the average arousal of the independent judges could be predicted by pupil size, corrected for the luminosity of our 47 participants. The comparison between the results obtained with our model using the self-reported arousal values of the 47 participants, as described so far, and the results obtained with our model using the arousal values provided by the 10 independent judges, is shown in Table 4.9. The model trained on the arousal reported by the 10

independent judges provides worse results than the one trained on the arousal reported by the 47 participants. For example, a worsening in $NRMSE$ of 188%. This is because when training a model, one must use a ground truth recorded from the same subjects from which the predictors are recorded, in this case, the self-reported arousal and pupil size of the 47 participants in the study.

Table 4.9: Comparison of the relationship between predicted and self-reported arousal in our model versus other researchers' models.

Metric	**Our Model**	**Hyperbolic Model [274]**	**Linear Model [276], [277]**	**Our Model (without self-reported arousal) [276]**
Correlation	0.65 $\pm$ 0.106 (mean p = 0.0025, max p = 0.096)	0.26 $\pm$ 0.150 (mean p = 0.1953, max p = 0.9566)	0.26 $\pm$ 0.146 (mean p = 0.2260, max p = 0.9892)	0.38 $\pm$ 0.074 (mean p = 0.0346, max p = 0.2886)
$NRMSE$	0.27 $\pm$ 0.036	0.41 $\pm$ 0.055	0.42 $\pm$ 0.052	0.78 $\pm$ 0.283
$R2$	0.436 $\pm$ 0.125	0.10 $\pm$ 0.087	0.07 $\pm$ 0.086	0.153 $\pm$ 0.054

After confirming the positive results of the corrected pupil size for luminosity, we utilised that pupil size to extract additional statistical features. We used them to train a machine learning model for emotion detection using only pupil size, GSR, and FER features. We then combined all these features and trained the model. The following section presents the results of all model evaluations.

4.4 Results of Advanced Machine Learning Techniques

This section presents the evaluation results of emotion detection models trained on data from 47 participants in the Emotionally Healthy group. The analysis includes unimodal models based on individual physiological signals—FER, pupil size, and GSR—as well as a multimodal approach that utilises feature fusion to combine all three signals.

4.4.1 Results of Model Training with FER

This section presents the performance of ML models trained using FER features for predicting self-reported arousal and valence. Two types of features were used:

- Interval-Based FER Features: Extracted from emotionally salient segments (EII) identified using EIIS (see Section 3.2.4).

- Entire-Clip FER Features: Extracted from the entire duration of each audiovisual clip.

Table 4.10 shows the comparison of emotion prediction results using interval-based and entire-clip-based FER features.

Table 4.10: Comparison of Emotion Prediction Using FER Interval vs. Full-clip Statistical Features

FER Feature Type	**Target**	$R2$**-Score**	**NRMSE**	**Pearson** r **(p)**
Interval-Based	Arousal	0.056 ± 0.610	0.919 ± 0.073	0.278 ± 0.169 (mean p = 0.218, max p = 0.459)
	Valence	0.052 ± 0.610	0.919 ± 0.073	0.430 ± 0.180 (mean p = 0.088, max p = 0.120)
Entire-Clip	Arousal	0.097 ± 0.090	0.767 ± 0.016	0.353 ± 0.159 (mean p = 0.134, max p = 0.278)
	Valence	0.065 ± 0.274	0.732 ± 0.040	0.370 ± 0.174 (mean p = 0.088, max p = 0.106)

Interval-Based FER Features

Model performance using interval-based FER features was weak. For arousal, the model explained minimal variance ($R2$ = 0.056 $\pm$ 0.610), with a high $NRMSE$ = 0.919 $\pm$ 0.073 and a weak, non-significant correlation (r = 0.278 $\pm$ 0.169 (mean p = 0.218, max p = 0.459)), indicating low predictive ability. For valence, the correlation was slightly stronger (r = 0.430 $\pm$ 0.180 (p = 0.088, mean p = 0.120)), but the $R2$ remained low (0.052 $\pm$ 0.610), and prediction error was consistently high.

This suggests that limited facial dynamics during short intervals may lack sufficient variability to train reliable models. These results are illustrated in the following figures: Figure 4.21 and Figure 4.23: Pearson correlation histograms for arousal and valence prediction.

Figure 4.22 and Figure 4.24: Boxplots of R2-scores across participants.

Entire-Clip FER Features

Using features extracted from the entire clip led to a slight improvement in model performance. For arousal, $R2$ increased to 0.097 $\pm$ 0.090, the correlation rose to 0.353 $\pm$ 0.159 (mean p = 0.134, max p = 0.278), and $NRMSE$ decreased to 0.767 $\pm$ 0.016.

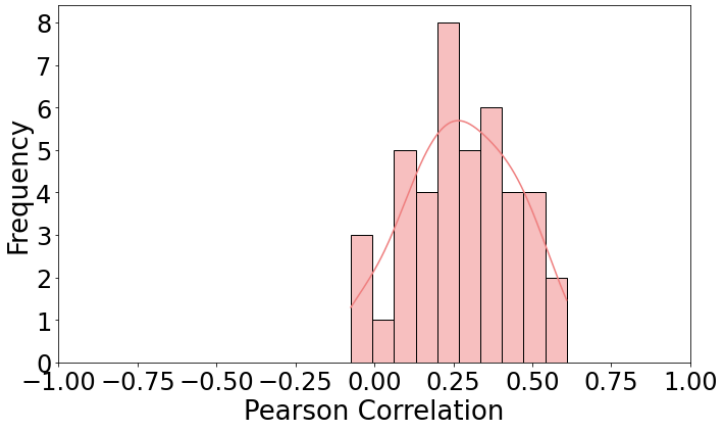


Figure 4.21: Histogram of Pearson Correlation for Arousal Prediction using Interval-Based FER Across all Participants.

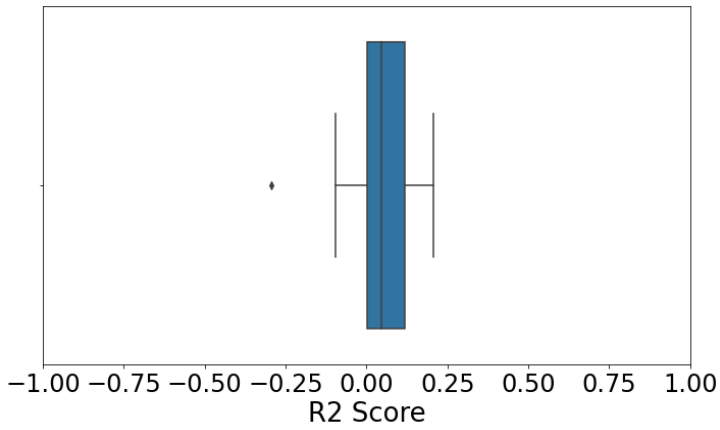


Figure 4.22: Boxplot of R2-Scores for Arousal using Interval-Based FER Across all Participants.

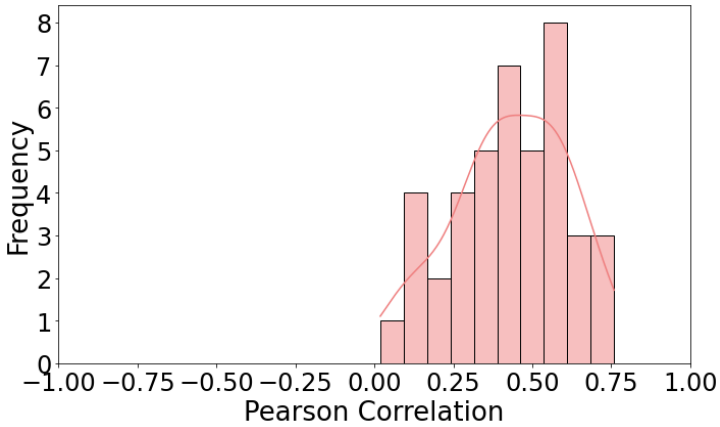


Figure 4.23: Histogram of Pearson Correlation for Valence Prediction using Interval-Based FER Across all Participants.

While still moderate, this represents a noticeable improvement over interval-based features. For valence, performance remained similar, with $R2 = 0.065 \pm 0.274$, correlation $r = 0.370 \pm 0.174$ (p = 0.088, max p = 0.106), and a slightly improved $NRMSE =$

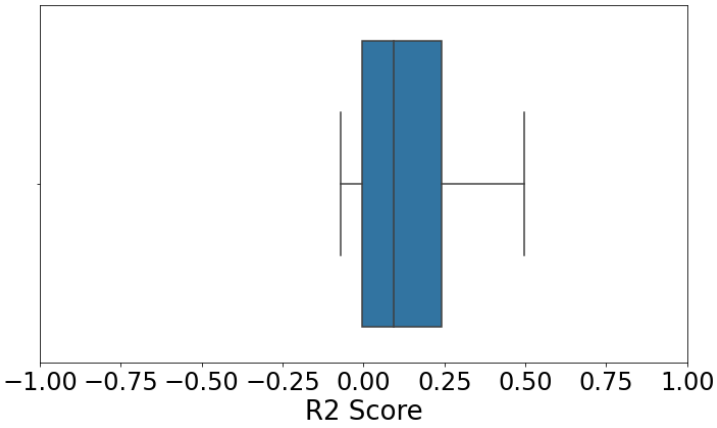


Figure 4.24: Boxplot of R2-Scores for Valence using Interval-Based FER Across all Participants.

0.732 ± 0.040.

These results are illustrated in the following figures:

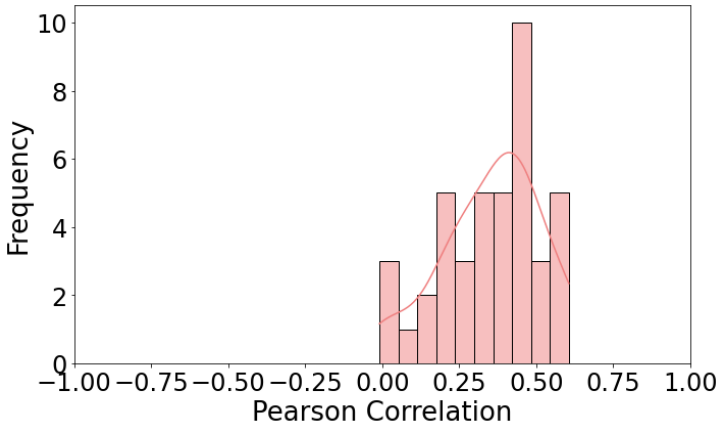


Figure 4.25: Histogram of Pearson Correlation for Arousal Prediction using Entire-Clip FER Across all Participants.

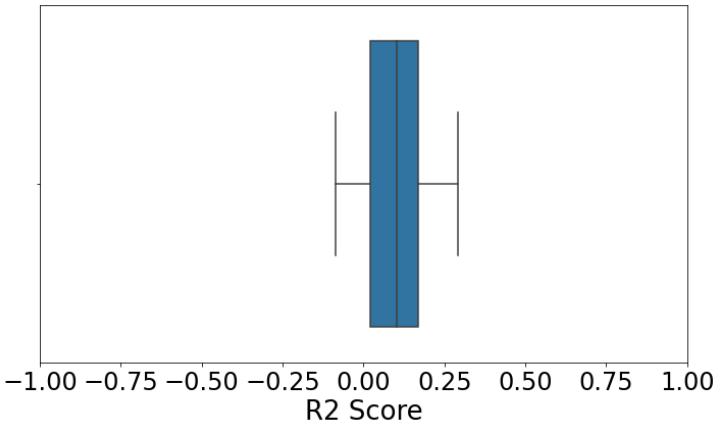


Figure 4.26: Boxplot of R2-Scores for Arousal using Entire-Clip FER Across all Participants.

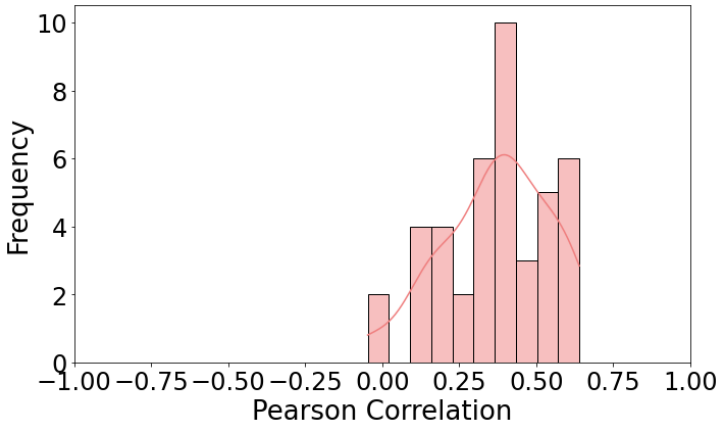


Figure 4.27: Histogram of Pearson Correlation for Valence Prediction using Entire-Clip FER Across all Participants.

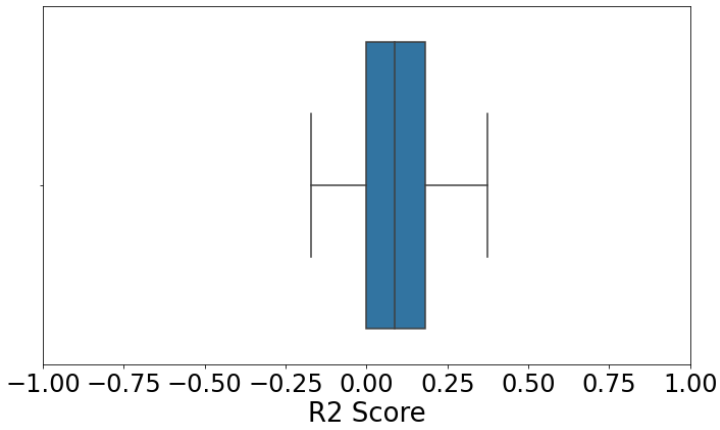


Figure 4.28: Boxplot of R2-Scores for Valence using Entire-Clip FER Across all Participants.

Figure 4.25 and Figure 4.27: Pearson correlation histograms for arousal and valence prediction.

Figure 4.26 and Figure 4.28: Boxplots of R2-scores across participants.

The comparison shows that while emotional intervals were hypothesised to provide stronger emotional signals, their shorter duration and reduced data volume might have limited their utility for regression modelling. Thus, utilising the entire video segment is a more reliable strategy for FER-based emotion detection.

Interpretation. These results show that:

- Entire-clip FER features provide more temporal information, which may explain the moderate gains in arousal prediction.
- Valence prediction does not consistently improve with either approach, likely because facial expressions alone are limited in conveying internal emotional evaluations, especially when subtle.
- The low $R2$-score values and moderate-to-high $NRMSE$s in both methods high-

light that FER alone is not sufficiently robust for accurate emotion prediction and may need to be combined with other modalities (e.g., GSR, pupil size) for more reliable performance.

Therefore, we processed other physiological signals, such as pupil size and GSR, the results of which are mentioned in the following sections.

4.4.2 Results of Model Training with Pupil Size

This section presents the results of training the emotion detection model using pupil size features, comparing the performance between models trained on data corrected for luminance and those using uncorrected features. Table 4.11 summarizes the key evaluation metrics, while Figures 4.29, 4.30, 4.31, and 4.32 visually depict the performance distributions across participants.

Arousal Prediction from pupil size. Correcting pupil size for luminance significantly enhanced model performance. The model trained on corrected features achieved a strong $R2$ of 0.556 $\pm$ 0.085, a high Pearson correlation of 0.765 $\pm$ 0.047, and a low $NRMSE$ of 0.229 $\pm$ 0.022, with mean p $< 10^{-7}$ and max p $< 10^{-7}$ across the participants, indicating high statistical significance. These results suggest that the corrected pupil size features provided a reliable basis for predicting arousal. In contrast, the

Table 4.11: Comparison of Emotion Prediction Performance Using Corrected vs. Non-Corrected Pupil Size Features

Correction	**Target**	$R2$ **Score**	**NRMSE**	**Pearson** r **(p)**
Corrected	Arousal	**0.556** $\pm$ 0.085	**0.229** $\pm$ 0.022	**0.765** $\pm$ 0.047 (mean p $< 10^{-7}$,max p $< 10^{-7}$)
	Valence	**0.259** $\pm$ 0.251	**0.295** $\pm$ 0.041	**0.595** $\pm$ 0.082 (mean p = 0.0012, max p = 0.075)
Non-Corrected	Arousal	0.235 $\pm$ 0.097	0.301 $\pm$ 0.020	0.521 $\pm$ 0.116 (mean p = 0.345, max p = 0.863)
	Valence	0.205 $\pm$ 0.242	0.306 $\pm$ 0.039	0.566 $\pm$ 0.065 (mean p = 0.0020, max p = 0.013)

model trained on non-corrected features showed much weaker performance across all participants: an $R2$ of 0.235 $\pm$ 0.097, a correlation of 0.521 $\pm$ 0.116, and a higher $NRMSE$ of 0.301 $\pm$ 0.020. Moreover, the statistical significance was not met (mean

p = 0.345, max p = 0.863), further emphasising the degradation in model accuracy when luminance is not accounted for.

Figures 4.29 and 4.31 illustrate the improved consistency and higher performance scores across participants when applying luminosity correction.

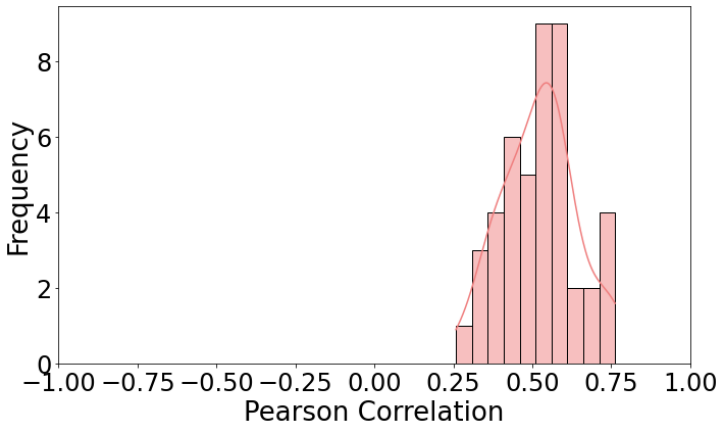


Figure 4.29: Histogram of Pearson Correlation for Arousal Prediction using Luminosity-Corrected Pupil Size Across all Participants.

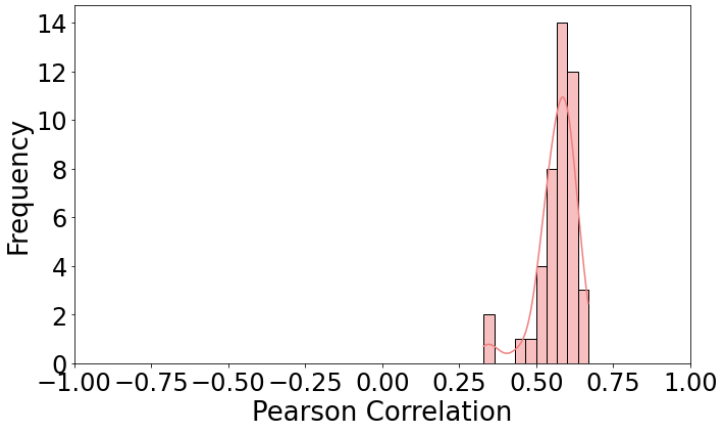


Figure 4.30: Histogram of Pearson Correlation for Valence Prediction using Luminosity-Corrected Pupil Size Across all Participants.

Valence Prediction from pupil size. For valence, the benefit of luminance correction was present but less pronounced. The corrected model achieved an $R2$ of 0.259 $\pm$ 0.251 and a correlation of 0.595 $\pm$ 0.082 (mean p = 0.0012, max p = 0.075) compared to 0.205 $\pm$ 0.242 and 0.566 $\pm$ 0.065 (mean p = 0.0020, max p = 0.013) for the non-corrected model across all participants. The $NRMSE$ also slightly improved from 0.306 $\pm$ 0.039 to 0.295 $\pm$ 0.041.

Figures 4.30 and 4.32 show that the corrected features generally led to better predictions, although the difference was not as substantial as for arousal.

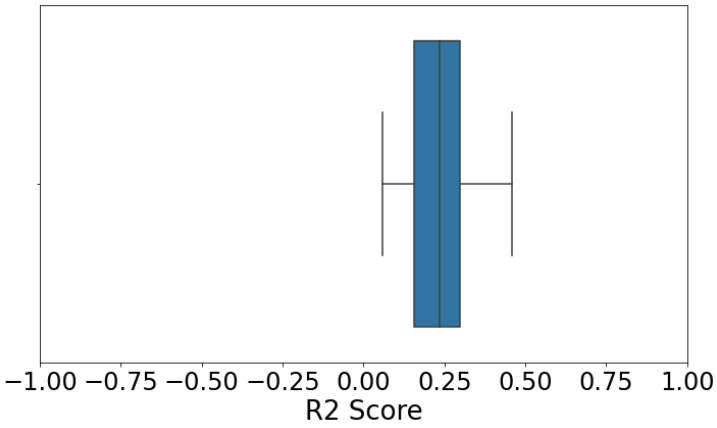


Figure 4.31: Boxplot of R2-scores for Arousal Prediction using Luminosity-Corrected Pupil Size Across all Participants.

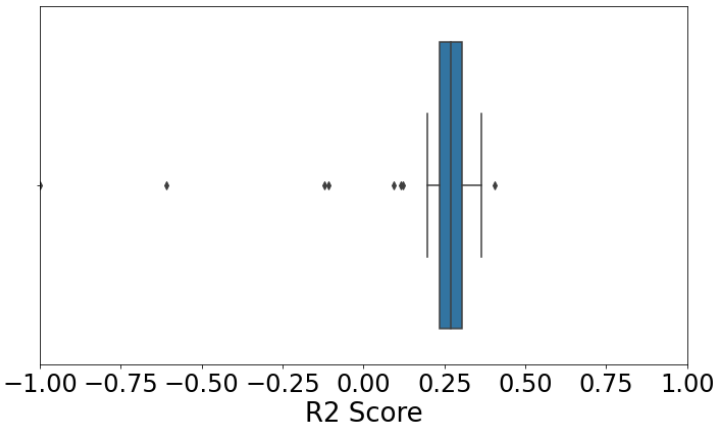


Figure 4.32: Boxplot of R2-scores for Valence Prediction using Luminosity-Corrected Pupil Size Across all Participants.

These findings highlight the importance of accounting for ambient luminance when utilising pupil size as a physiological marker in emotion detection. Failing to consider this factor significantly compromises the model's accuracy, particularly in predicting arousal.

4.4.3 Results of Model Training with GSR

We trained an XGBoost regression model using features extracted from GSR signals to predict self-reported emotional dimensions, specifically arousal and Valence. The model was evaluated using LOPO Cross-Validation, ensuring each participant was used once as a test subject while the others were used for training. Additionally, 5-fold cross-validation was applied within the training data to avoid overfitting and optimise model generalisation.

Arousal Prediction from GSR. The model demonstrated a moderate ability to predict arousal based on GSR features. The average $R2$-score across all participants was 0.469 $\pm$ 0.006, indicating that the model could explain approximately 47% of self-reported arousal variance. The average Pearson correlation between predicted and actual arousal values was 0.720 $\pm$ 0.006, with mean p $< 10^{-7}$ and max p $< 10^{-7}$, confirming strong and statistically significant linear relationships.

Moreover, the $NRMSE$ was 0.251 $\pm$ 0.001, suggesting a reasonably good prediction performance relative to the range of observed arousal values.

Valence Prediction from GSR. For valence prediction, the model performed more effectively. The mean $R2$ reached 0.573 $\pm$ 0.009, indicating a firmer fit than arousal. The correlation coefficient was 0.828 $\pm$ 0.006, again statistically significant with mean p $< 10^{-7}$ and max p $< 10^{-7}$. This reflects a high level of agreement between the predicted and actual valence scores across participants.

The $NRMSE$ for valence was 0.218 $\pm$ 0.002, slightly lower than that for arousal. This further supports the model's higher precision in predicting valence using GSR features.

To better understand how performance varied across participants, we plotted the distribution of Pearson correlation and $R2$-scores:

Figure 4.33 shows that most participants had correlation values between 0.70 and 0.80, with a right-skewed distribution. This suggests that while most participants had high correlation, a few had notably lower values, potentially due to individual variability in physiological responses.

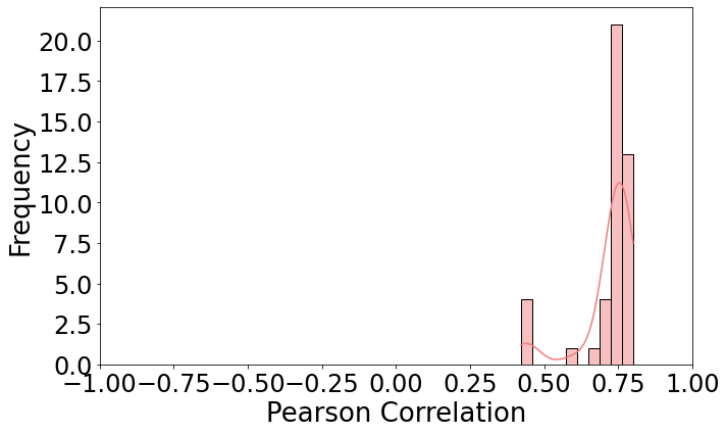


Figure 4.33: Histogram of Pearson Correlation for Arousal Prediction using GSR Across all Participants.

Figure 4.34 illustrates a more potent and more concentrated distribution of correlation values, with most values ranging between 0.80 and 0.90. This confirms the model's superior performance in predicting valence from GSR features compared to

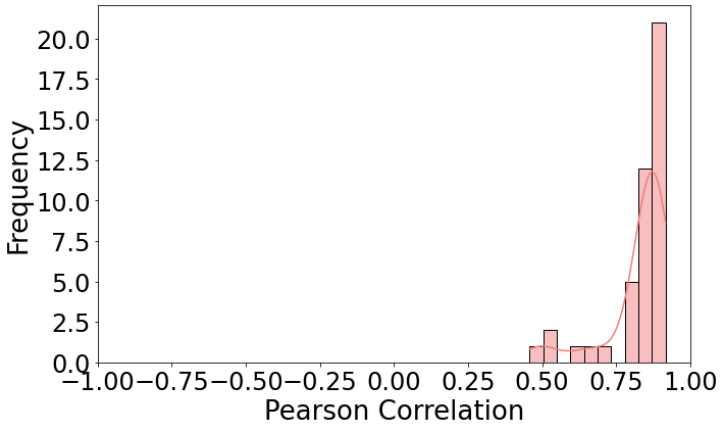


Figure 4.34: Histogram of Pearson Correlation for Valence Prediction using GSR Across all Participants.

arousal. In Figure 4.35, the box plot reveals a moderate spread of $R2$, with a few outliers on the lower end. Despite some participants' low predictive performance, the median $R2$ remains close to 0.50, supporting the model's general efficacy. Figure 4.36

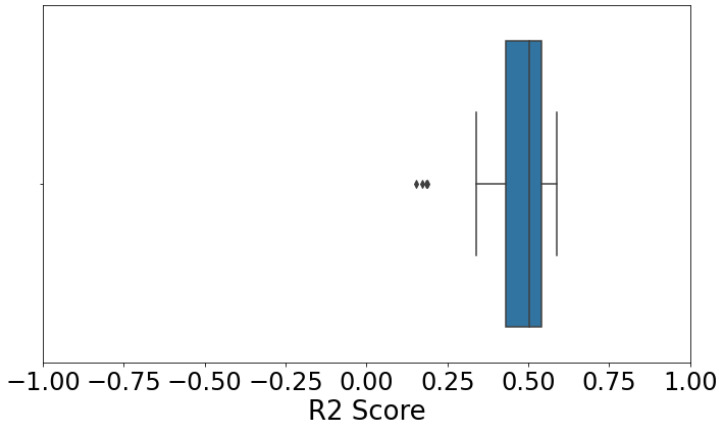


Figure 4.35: Boxplot of R2-Scores for Arousal using GSR Across all Participants.

highlights a tighter and higher $R2$-score distribution for valence predictions, with most participants achieving scores above 0.50 and a median exceeding 0.60. A few outliers on the lower end may reflect cases where GSR signals were less informative due to participant-specific variance or noise.

These findings confirm that GSR signals are valuable predictors of emotional states, especially valence. The model's stronger performance in predicting valence compared to arousal could be attributed to the steadier nature of GSR changes linked to emotional valence, whereas arousal may involve more rapid physiological fluctuations.

In subsequent sections, we compare these findings with results from models trained on pupil size, FER features, and integrated multimodal models.

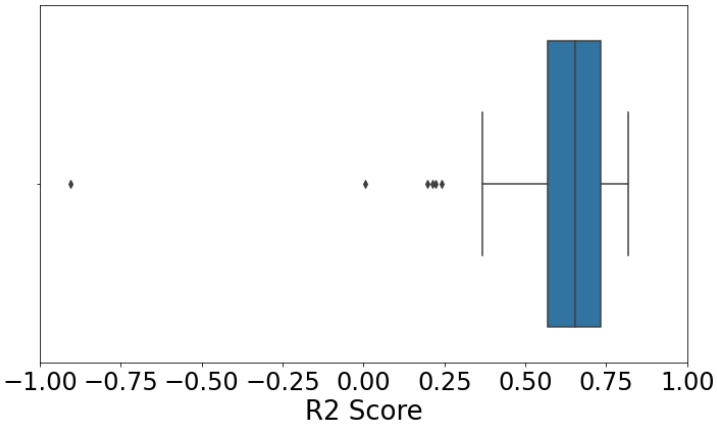


Figure 4.36: Boxplot of R2-Scores for Valence using GSR Across all Participants.

4.4.4 Results of multimodal Feature Fusion: Integrating FER, Pupil Size, and GSR

This section presents the results of ML models trained to predict self-reported arousal and valence using three physiological signal sources—Entire-Clip based FER, pupil size (corrected for luminance), and GSR—evaluated independently and jointly through multimodal feature fusion. Model performance is assessed using the $R2$ score, Pearson correlation r, and $NRMSE$, with all metrics reported as means $\pm$ standard deviations across 47 participants (presented in Table 4.12).

FER. Models trained on Entire-Clip FER features demonstrated limited predictive ability. For arousal prediction, the model yielded a low $R2$ score of 0.097 $\pm$ 0.090, a weak Pearson correlation of$r$ = 0.353 $\pm$ 0.159 (mean p = 0.134, max p = 0.278), and a high $NRMSE$ of 0.767 $\pm$ 0.016. Valence prediction showed slightly better results, with$r$ = 0.370 $\pm$ 0.174 and $R2$ = 0.065 $\pm$ 0.274 (mean p = 0.088, max p = 0.106), though $NRMSE$ remained high at 0.732 $\pm$ 0.040. These outcomes suggest that FER is not sufficiently robust for reliable emotion prediction when used in isolation, likely due to variability in facial expressiveness and context loss in short intervals.

Pupil Size (Corrected for Luminance). Correcting for luminance yielded a substantial performance improvement. For arousal, the model achieved an $R2$ of 0.556 $\pm$ 0.085, a high correlation of r = 0.765 $\pm$ 0.047 (p < 10^{-7}), and a low $NRMSE$ of 0.229 $\pm$ 0.022. Valence prediction improved similarly, with $R2$ = 0.259 $\pm$ 0.251, r = 0.595 $\pm$ 0.082 (p = 0.0012), and $NRMSE$ = 0.295 $\pm$ 0.041. These results confirm that pupil dilation is a reliable emotional indicator when corrected for ambient light conditions.

GSR. GSR-based models demonstrated consistent and high performance for both dimensions. Arousal prediction reached an $R2$ of 0.469 $\pm$ 0.006 and correlation of r

Table 4.12: Comparison of Model Performance Using Different Feature Sets for Emotion Prediction Across all Participants

Feature Set	**Target**	$R2$ **Score**	**NRMSE**	**Pearson's** r **(p)**	**CCC**
FER	Arousal	0.097 ± 0.090	0.767 ± 0.016	0.353 ± 0.159 (mean p = 0.134, max p = 0.278)	0.150 ± 0.102
	Valence	0.065 ± 0.274	0.732 ± 0.040	0.370 ± 0.174 (mean p = 0.088, max p = 0.106)	0.207 ± 0.140
Corrected Pupil Size	Arousal	0.556 ± 0.085	0.229 ± 0.022	0.765 ± 0.047 (mean p $< 10^{-7}$, max p $< 10^{-7}$)	0.758 ± 0.080
	Valence	0.259 ± 0.251	0.295 ± 0.041	0.595 ± 0.082 (mean p = 0.0012, max p = 0.075)	0.600 ± 0.105
GSR	Arousal	0.469 ± 0.006	0.251 ± 0.001	0.720 ± 0.006 (mean p $< 10^{-7}$, max p $< 10^{-7}$)	0.579 ± 0.147
	Valence	0.573 ± 0.009	0.218 ± 0.002	0.828 ± 0.006 (mean p $< 10^{-7}$, max p $< 10^{-7}$)	0.719 ± 0.109
Final Model (All)	Arousal	**0.710** ± 0.098	**0.183** ± 0.030	**0.865** ± 0.061 (mean p $< 10^{-7}$, max p $< 10^{-7}$)	0.814 ± 0.084
	Valence	**0.665** ± 0.359	**0.187** ± 0.070	**0.913** ± 0.041 (mean p $< 10^{-7}$, max p $< 10^{-7}$)	0.822 ± 0.090

= 0.720 ± 0.006 (mean p < 10^{-7}, max p < 10^{-7}), with $NRMSE$ = 0.251 ± 0.001. Valence prediction performed even better: $R2$ = 0.573 ± 0.009, r = 0.828 ± 0.006 (mean p < 10^{-7}, max p < 10^{-7}), and $NRMSE$ = 0.218 ± 0.002. These findings highlight GSR as a strong stand-alone modality, effectively capturing both emotional intensity and evaluative tone.

Multimodal Integration- FER + Pupil Size + GSR. The fusion of all three modalities produced the highest prediction accuracy across all evaluation metrics, reinforcing the value of multimodal emotion recognition. For arousal, the model achieved an $R2$ of 0.710 ± 0.098, a correlation of r = 0.865 ± 0.061 (mean p < 10^{-7}, max p < 10^{-7}), and

an $NRMSE$ of 0.183 $\pm$ 0.030. For valence, performance remained strong, with an $R2$ of 0.665 $\pm$ 0.359, a correlation of r = 0.913 $\pm$ 0.041 (mean p $< 10^{-7}$, max p $< 10^{-7}$), and an $NRMSE$ of 0.187 $\pm$ 0.070.

These outcomes, visualised in Figures 4.37–4.40, illustrate a dramatic reduction in error and near-perfect correlation between predicted and actual values.

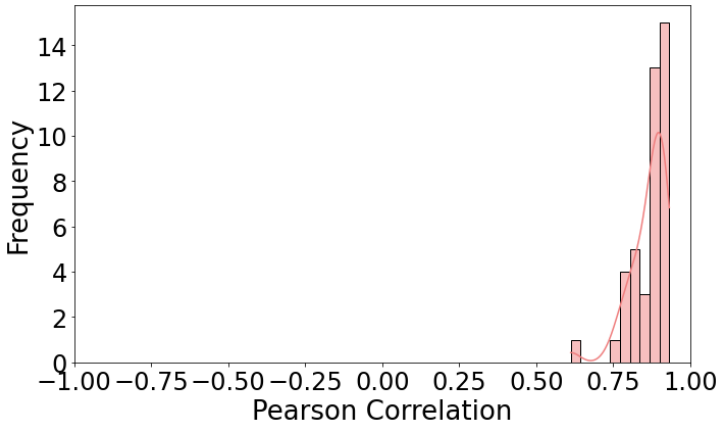


Figure 4.37: Histogram of Pearson Correlation for Arousal (Multimodal Model) Across all Participants.

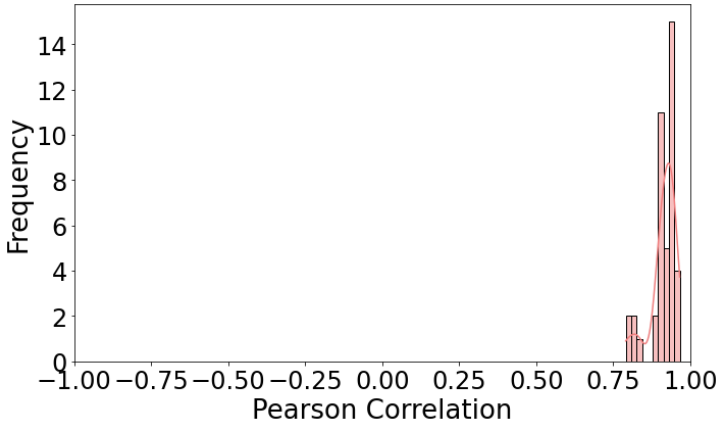


Figure 4.38: Histogram of Pearson Correlation for Valence (Multimodal Model) Across all Participants.

Interpretation. These results collectively highlight the importance of combining complementary physiological signals for accurate emotion prediction:

- Pupil size and GSR are individually strong predictors, particularly for arousal.
- FER, while weak in isolation, enhances performance when integrated with other modalities.
- The multimodal approach captures diverse emotional cues (physiological, cognitive, expressive), offering significant predictive advantages and better generalisation across individuals.

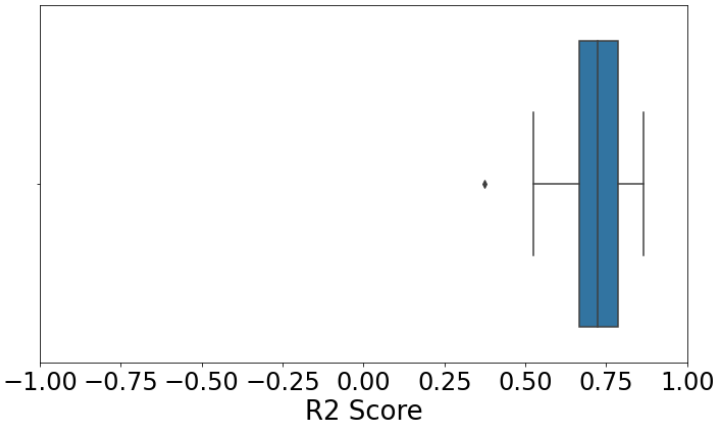


Figure 4.39: Boxplot of $R2$ Scores for Arousal (Multimodal Model) Across all Participants.

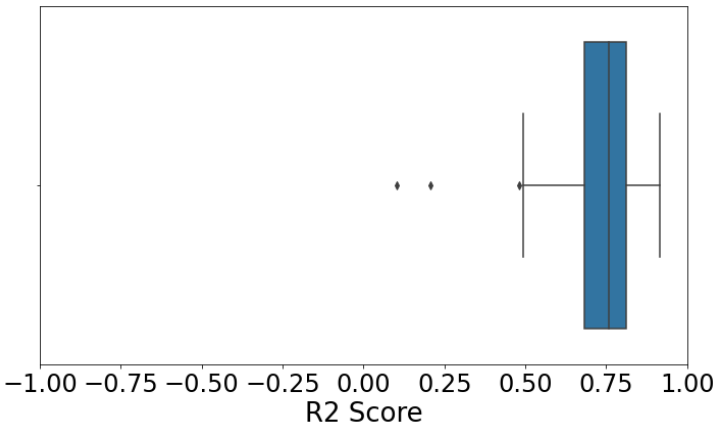


Figure 4.40: Boxplot of $R2$ Scores for Valence (Multimodal Model) Across all Participants.

These findings support the feasibility of real-time, multimodal emotion detection systems for affective computing and personalised human–machine interaction.

4.4.5 Comparative Analysis with Existing Literature

In this subsection, I have done statistic comparison of the existing unimodal literature that has used a regression model to predict continuous emotional states.

Table 4.13 summarises the Fisher-z comparisons between our models and published benchmarks. For FER, our CCCs were substantially lower than those reported by Raju et al. [360], with significant differences for both arousal ($z = -3.30$, $p < 0.01$) and valence ($z = -3.48$, $p < 0.001$), however they used video and audio instead we have only used video without considering audio, therefore the results are not directly comparable. In contrast, our pupil-based models achieved markedly higher performance than O'Dwyer et al. [359], showing significant improvements in both arousal ($z = 4.87$, $p < 0.001$) and valence ($z = 2.78$, $p < 0.01$). Similarly, our GSR-based models outperformed

Table 4.13: Comparison between our CCC results and literature benchmarks. Fisher-z tests were performed only when a literature CCC and a comparable independent sample.

Model/Target	**Our CCC**	**Lit. CCC**	no. of participants	z	p_1	p_2
FER-Arousal	0.150 ± 0.102	0.638 [360]	96	-3.30	0.0005	0.0010
FER-Valence	0.207 ± 0.140	0.689 [360]	96	-3.48	0.0003	0.0005
Pupil-Arousal	0.758 ± 0.080	0.154 [359]	150	4.87	$< 10^{-7}$	$< 10^{-7}$
Pupil-Valence	0.600 ± 0.105	0.212 [359]	150	2.78	0.0027	0.0054
GSR-Arousal	0.579 ± 0.147	0.082 [361]	27	2.28	0.0113	0.0226
GSR-Valence	0.719 ± 0.109	0.177 [361]	27	2.86	0.0021	0.0042

Brady et al. [361], with statistically significant gains in arousal ($z = 2.28$, $p < 0.05$) and valence ($z = 2.86$, $p < 0.01$). Both O'Dwyer et al. [359] and Brady et al. [361] tried to enhance performance by integrating additional modalities such as speech, ECG, video, and audio. However, none of these studies employed the specific combination of corrected pupil size, FER, and GSR used in our fusion model, making their results not directly comparable.

These results highlight that unimodal FER in our dataset still has room for improvement, but confirm strong and statistically significant improvements for pupil and GSR modalities compared to prior studies.

Chapter 5

Discussion and Future Work

This chapter presents the key findings, methodological developments, and implications of the work outlined in this thesis, which focused on developing a multimodal emotion detection model utilising facial expression recognition (FER), pupil size, and galvanic skin response (GSR) signals. The main study introduced refined preprocessing pipelines, improved feature extraction techniques, and regression-based machine learning models, building on a pilot study. A significant contribution was the development of a novel pupil size correction model that removed the influence of luminosity, addressing a longstanding challenge in affective computing.

5.1 Insights and Challenges from the Pilot Study

The pilot study served as a foundational phase, revealing several limitations in data acquisition, preprocessing, and model design that directly informed the refinements introduced in the main study. While the methods used for emotion labelling and FER analysis followed established approaches in the literature, the pilot was crucial in identifying practical challenges that shaped the novelty of our subsequent framework.

A key finding concerned the variability in emotion labels across participants. Around 15 participants provided minimal emotional responses, while others gave overly broad or inconsistent labels. This inconsistency highlighted the difficulty of obtaining stable emotion ground truth without careful participant management and reinforced the importance of developing stricter inclusion criteria in later phases of the project. Although the ground truth construction itself was based on prior work, our pilot confirmed its applicability and clarified the conditions under which it produces reliable outputs.

Another insight was the limited expressiveness observed in FER. Out of the 47 participants, only about 10 displayed clear and consistent facial responses, with the remainder showing weak or flat FER signals. This was partly attributable to the low-resolution facial camera setup used in the pilot. In some cases, participants also ap-

peared overly self-conscious of being recorded, as suggested by FER-derived emotion plots showing little or no variation across all stimuli. These findings directly motivated methodological changes in the main study, including the use of higher-resolution cameras, improved camera placement, and procedural adjustments to reduce participant awareness of recording.

Finally, the pilot demonstrated that FER contributed more strongly to arousal detection than to valence, consistent with prior observations in affective computing. While this confirmed that FER can play a useful role in multimodal emotion recognition, it also highlighted the need to integrate additional physiological signals, such as pupil size and GSR, to improve valence prediction. Thus, although the pilot did not produce novel methodological contributions in itself, it provided critical evidence for the weaknesses of FER as a stand-alone signal and justified the multimodal, methodologically refined approach adopted in the main study.

Pupil size, though physiologically tied to emotional arousal, showed a weak correlation with self-reported emotions in the pilot dataset. Uncontrolled lighting, which introduced noise and masked emotional signals, was blamed for this [207], [274], [297]. Baseline correction using a grey screen was insufficient [109], [205], [292]. Consequently, the pilot study demonstrated that pupil size without luminosity correction is not a reliable stand-alone emotional indicator, prompting the development of a luminosity isolation model for the main study.

In the pilot, only basic statistical features were extracted from GSR signals, limiting emotional resolution. Additionally, using a fixed baseline across participants was found to be inappropriate, given natural individual variation. The signal-to-noise ratio (SNR) was challenging to establish without an actual ground truth. This highlighted the need for more prosperous feature extraction (e.g., phasic response latency, rise time) and individualised baseline correction, which we incorporated into the main study.

These challenges highlighted areas for improvement and motivated a shift towards a more sophisticated modelling approach, specifically, the adoption of regression techniques to handle the increased complexity of emotion detection in real-world settings.

5.1.1 Modelling Challenges and Motivation for Regression

During the pilot study, we did not apply classification models. However, our observations and initial data analysis revealed several challenges commonly associated with classification, which informed the design of the main study. These included the limited ability of rigid class labels to capture continuous emotional states, poor generalisation across participants, and issues with class imbalance.

- **Rigid Class Labels:** Predefined categories (e.g., high vs. low arousal) failed to capture the nuanced and continuous nature of emotions. To overcome this, we

adopted regression-based modelling, which represents emotional responses on a continuous scale and provides greater granularity.

- **Generalisation Issues:** Using Leave-One-Participant-Out (LOPO) cross-validation, we found that models trained with categorical labels often failed to generalise to unseen participants due to inter-individual variability. Regression helps mitigate this by focusing on continuous emotional dimensions rather than discrete classes, offering more consistent predictions across users.
- **Class Imbalance:** The pilot study dataset showed an unequal distribution of emotional states (e.g., fewer high-arousal or negative-valence samples). While we addressed this by selecting balanced stimuli using Russell's Circumplex Model [32], regression further reduces reliance on categorical boundaries, minimising bias towards dominant states.

These limitations highlighted the need for a more flexible modelling framework. Regression approaches are particularly well-suited for emotion detection because they: (i) capture the full spectrum of emotional fluctuations, (ii) offer better generalisation across participants, and (iii) support real-time tracking of emotions, an essential requirement in practical applications such as affective computing, clinical monitoring, and neurorehabilitation.

This transition to regression thus provided a more sophisticated, accurate, and context-appropriate solution for emotion detection in the main study.

5.1.2 Methodological Refinements in the Main Study

Several refinements have been made based on the findings of the pilot:

- Psychological screening excluded participants with conditions like anxiety or alexithymia, improving emotional data consistency.
- Luminosity correction in pupil size removed ambient confounds, enhancing signal fidelity.
- Expanded GSR features, then just using basic statistical features.
- Regression modelling replaced classification, allowing smoother prediction and better generalisation.
- The LOPO Cross-Validation ensured robustness across individuals.
- Feature fusion enabled effective multimodal integration.

5.2 Main Study: Improvements and Findings

This section presents the key findings derived from the methodologies employed in this study, including luminosity correction for pupil size, the extraction of a comprehensive set of GSR features, the development of a FER-to-vector mapping method,

and multimodal integration of physiological and behavioural signals. It also highlights how our approach addresses limitations identified in prior research related to pupil size confounding by luminosity, GSR signal variability, and the limited use of feature-level fusion in emotion detection.

We began by collecting multimodal physiological data while ensuring that participants did not have underlying mental health conditions that could affect emotional responses. As part of the selection process, we applied psychological screening to exclude individuals with alexithymia, anxiety, depression, or personality disorders. This ensured that the dataset reflected emotionally healthy, non-clinical populations and reduced potential variability and noise in both physiological signals and ground-truth labels. Prior literature has shown that such conditions affect emotional awareness, perception, and self-report accuracy [16]–[18]. By excluding these cases, we improved the consistency and interpretability of the dataset, providing a stronger foundation for training emotion detection models.

A central contribution of this work is the generation of individualised ground truth, rather than relying solely on aggregated averages across participants. Given that each individual may perceive stimuli differently due to cognitive biases, emotional sensitivity, or personal experience, averaging can obscure meaningful variations in perception. To address this, we employed the INDSCAL multidimensional scaling technique, which enabled us to model a shared two-dimensional emotional space while simultaneously capturing individual perceptual weights. This approach preserved the circular structure of emotions central to Russell's Circumplex Model [32], ensuring that both common trends and individual differences in emotion perception were retained. Although we explored factor analysis as an alternative [394], it failed to reproduce this circular structure, confirming INDSCAL as the more appropriate choice. In doing so, we established a psychologically meaningful and mathematically coherent framework for emotion labelling that directly responds to the challenges of defining ground truth in affective computing.

Beyond pupillometry, our contributions extend to other modalities. For FER, we introduced a novel vectorial mapping approach that projects the seven basic emotions detected by iMotions' AFFDEX toolkit into continuous two-dimensional valence–arousal coordinates. This transformation allowed FER to be meaningfully integrated with other modalities in a unified affective space. For GSR, we extracted 40 features spanning time-domain, frequency-domain, and time–frequency characteristics, following the framework outlined in Shukla et al. [305]. These features provided a rich representation of emotional arousal and valence dynamics.

Most importantly, we integrated features from GSR, corrected pupil size, and FER within a unified machine learning framework. To the best of our knowledge, no prior study has combined exactly these modalities at the feature level for training emotion

detection models. This multimodal approach yielded superior predictive performance compared to single-modality models, while also maintaining practicality and ease of use for clinical applications, where lightweight and non-invasive measurement systems are essential.

5.2.1 FER: challenges and Novelty

Our study builds on this literature, introducing a distinctive and interpretable approach. Unlike prior work that either (i) relies on normative mappings of basic emotions to valence-arousal coordinates [262] or (ii) trains black-box regressors from FER embeddings to continuous affect labels [263], we implemented a vectorial transformation that integrates both the direction (circumplex angle) and magnitude (AU-derived intensity) of each basic emotion. In this research, our goal wasn't to compute emotions from facial recording; our clear goal was to map the emotions recorded by iMotions affectiva (see chapter 3, section 3.1.7) into 2-dimensional space based on Russell's circumplex model, so that we can use it aa a comparable feature with self-reported arousal and valence to train the machine learning.

By treating emotion categories as vectors in the circumplex, our method allows multiple emotions to co-occur and combine into a single continuous valence-arousal coordinate. This contrasts with categorical FER pipelines that force a single dominant label, thereby losing information about mixed states.

The results from our main study confirmed that this transformation yields psychologically interpretable emotional trajectories, with FER-derived valence and arousal points clustering in expected quadrants. More importantly, when combined with GSR and corrected pupil size, FER added complementary behavioural information that improved multimodal performance despite being the weakest unimodal predictor. This demonstrates that our FER pipeline is not merely a stand-alone classifier but a bridge between categorical recognition and dimensional modelling. This contribution fills a methodological gap in the literature.

In summary, our approach to FER is novel in two key respects:

- Vectorial mapping of basic emotion intensities into circumplex space - preserving co-occurrence and intensity rather than collapsing expressions into single categories.
- Integration into a multimodal framework with corrected pupil size and GSR - an unexplored signal combination that leverages FER's behavioural cues alongside autonomic and ocular measures.

Together, these contributions extend prior FER research by demonstrating how discrete facial signals can be systematically transformed into dimensional affective coordinates, enabling richer and more interpretable multimodal emotion recognition.

5.2.2 Pupil Size and Luminosity Correction: A Novel Contribution

One of this thesis's most critical methodological contributions was developing a novel mathematical model to isolate emotional arousal from pupil size data by removing the effects of stimulus and ambient luminosity [207], [274], [297].

Pupil dilation is known to reflect emotional arousal, but is highly sensitive to lighting conditions, which can confound interpretation. Traditional approaches, such as linear regression, isoluminant stimuli, and baseline correction, have attempted to address this issue, albeit with limited success in terms of universality and accuracy [205], [275], [276], [278].

We developed an exponential calibration model to calibrate pupil response to luminosity across subjects. Adjusts for individual differences in pupillary reactivity and removes luminosity effects at ambient and stimulus levels.

Compared to Nakayama's isoluminant model and Raiturkar's linear model [274], [276], our model showed lower variance (coefficient of variation: 0.03 vs. 0.09) and higher emotion prediction accuracy (R2 increased from 0.10 to 0.436 for arousal). Unlike prior work, our model generalises across participants and varying ambient luminosity conditions through a calibration process that adapts to both individual baselines and environmental lighting. Empirically, we observed a difference of less than 3% in prediction accuracy (difference between correlation of the prediction and ground truth) when comparing dark and well-lit laboratory conditions, demonstrating robustness to real-world variability. This makes the approach suitable for deployment beyond controlled environments [203], [292], [297], [367], [368].

The model's first stage, which involves correction, utilises mathematical calibration; the second stage, which requires emotion prediction, employs simple linear regression. This balance of rigour and usability means it can be deployed in systems without requiring ML or psychophysiology expertise.

Our evaluation predicted self-reported arousal with substantial accuracy, without requiring complex feature extraction, by directly using pupil size. When tested using arousal labels from independent judges, prediction performance dropped substantially, highlighting the importance of subjective emotional self-reports in modelling. This evaluation confirms that individual perception must be considered in conjunction with physiological responses.

In future work, we can incorporate the temporal dynamics of the pupil response, for example, by using sequence modelling approaches such as recurrent neural networks (RNNs) or temporal convolutional networks. While our current model relies on aggregated or static features, it does not capture how pupil size evolves in response to emotional stimuli. This temporal information may contain valuable cues about the onset, intensity, and duration of emotional reactions, which are essential for distinguishing

between subtle affective states. Modelling these dynamics could enhance the model's sensitivity to complex or mixed emotions, providing a more nuanced understanding of emotion processing over time.

In future studies, we will make our model dynamic, extend it to extreme lighting conditions, such as those encountered when wearing eye-tracker glasses outside a laboratory on a sunny day, and adapt it to experiments without eye-trackers and with regular webcams in online settings.

Due to its ease of implementation, our model has practical applications, particularly in mental health monitoring and consumer research. It requires only existing human insights software, such as iMotions, and basic eye-tracking hardware [395]. No advanced skills are required to detect emotional arousal from pupil size, even without using more complex bio signals to analyse, such as GSR, EEG, and ECG in the mental health field, where a shortage of trained counsellors persists [396]. Our model offers an accessible AI-driven solution for emotion detection, benefitting underserved populations. In consumer research, pupillometry has been used to assess emotional responses to advertisements; however, it has not gained much popularity among marketers due to its low reliability [397]. By eliminating luminosity effects, our model enables marketers to gauge the emotional arousal elicited by their advertisements with greater accuracy. Extensive trials in real-world settings will be essential to improve generalisability, and models should be tested across diverse populations and stimuli. Building on this, the following section examines the role of GSR features in modelling emotional arousal, further contributing to the multimodal framework employed in this study.

5.2.3 GSR-Based Emotion Detection

The use of the GSR for emotion detection has received considerable attention due to its sensitivity to autonomic arousal, a key physiological marker of emotional intensity. However, several limitations affect the stand-alone reliability of GSR in practical applications, such as signal variability and susceptibility to noise. We extended our feature extraction approach beyond the pilot study to address these challenges. Specifically, we included broader features from the time, frequency and time-frequency domains. This enriched feature space enables the model to more effectively capture the complex and dynamic patterns of GSR signals, including both short-term fluctuations and longer-term trends. As a result, our approach improves the model's ability to detect emotional responses more accurately and robustly from GSR alone.

Traditional GSR-based systems typically classify emotions into discrete categories, such as happy or sad. However, emotions are more accurately represented along continuous valence and arousal dimensions. To tackle this, our model was crafted using GSR through a regression-based method, which attains an 83% correlation with

valence and a 72% correlation with arousal. This provides a nuanced and authentic interpretation of emotional states, as opposed to the rigidity of discrete classification systems [365].

In our model, GSR demonstrated the highest predictive power, particularly for valence ($R2 = 0.573$), possibly due to its strong association with sustained autonomic arousal. GSR is directly influenced by SNS activity and reflects changes in sweat gland activity, particularly in response to emotionally salient or arousing stimuli [212]. While GSR has traditionally been associated with arousal rather than valence, some studies suggest that the duration and context of the emotional stimulus may lead to stronger correlations with valence, especially when valence is operationalised along an affective arousal continuum [398]. Thus, the high predictive power of the GSR in this context may reflect its sensitivity not only to momentary spikes in arousal but also to more sustained, affectively charged states that tend to co-vary with valence in naturalistic or continuous emotional experiences (e.g., positive high-arousal vs. negative high-arousal states).

Furthermore, a standard limitation in GSR research is its focus on static features, which fail to capture the dynamic and evolving nature of emotional responses. Our approach improves upon this by analysing the entire stimulus period, allowing the model to track the gradual and dynamic shifts in arousal that occur over time [369].

GSR is also typically limited in distinguishing between emotions that share similar arousal levels but differ in valence, such as fear and excitement. To overcome this, we integrated FER and pupil size data, which offer complementary insights into emotional valence. This multimodal fusion significantly improves the system's ability to differentiate emotional states with higher precision [370].

External factors, such as motion distortions and environmental conditions like temperature, can distort GSR signals, making real-world applications challenging. We mitigated these issues using robust preprocessing techniques and standardised data collection protocols to ensure consistency and reliability across participants [302].

Processing multimodal data in real time is often computationally demanding. Our model utilises the XGBoost regression algorithm to ensure scalability and efficiency, enabling real-time emotion prediction on mobile or wearable platforms with minimal computational overhead [252].

Future work should focus on refining dynamic GSR feature extraction, incorporating contextual information such as environmental and social cues, and establishing standardised protocols to improve comparability and generalisability across studies.

While GSR remains a valuable signal for arousal detection, its predictive capabilities are significantly enhanced when combined with other modalities. The following section examines how multimodal integration can improve the accuracy and reliability of emotion detection systems.

5.2.4 Multimodal Modelling Enhances Emotion Prediction

Each unimodal signal offered distinct contributions to emotion prediction. GSR demonstrated the highest predictive power, particularly for valence (R^2 = 0.573). This likely reflects the fact that skin conductance captures both short-term arousal spikes and sustained autonomic states that often correspond to how pleasant or unpleasant an experience feels. Corrected pupil size effectively tracked rapid fluctuations in arousal (R^2 = 0.556) once luminosity effects were removed, highlighting the value of our calibration procedure in isolating genuine emotional responses from environmental noise. FER, although the weakest stand-alone predictor, provided behavioural information that was not available in the physiological channels and therefore carried complementary value in multimodal integration.

The multimodal fusion of FER, GSR, and pupil size substantially outperformed any individual modality, with average $R^2 = 0.710$ and $r = 0.865$ for arousal, and average $R^2 = 0.665$ and $r = 0.913$ for valence (see Table 4.12). These results confirm the complementary nature of the three modalities, but more importantly, they highlight the novelty of our framework. To the best of our knowledge, this is the first study to integrate corrected pupil size, GSR, and vectorially mapped FER into a unified regression-based model of continuous valence and arousal. Prior multimodal work has typically focused on combinations such as audio-video or ECG-EDA [339], [362], whereas our configuration explores a new space of ocular, autonomic, and facial behavioural signals. This unique signal mix, combined with feature-level fusion and participant-independent cross-validation, allowed us to achieve robust and interpretable predictions that set a new benchmark for non-invasive, camera- and sensor-based emotion recognition.

5.2.5 Comparative Analysis with Existing Literature

While previous studies used classification methods using XGBoost and hybrid CNN-XGBoost architectures, for binary or multi-class sentiment and valence/arousal detection [399], [400], our approach focuses on a *regression* problem, aiming to predict continuous emotional states rather than discrete categories. Regression works on continuous scales rather than classes, so performance cannot be directly compared with classification studies. In addition, our dataset was multimodal, incorporating FER, pupil size (corrected for screen luminosity), and GSR features, in contrast to previous works that often rely on high-density EEG or hybrid deep learning pipelines.

A one-to-one statistical comparison was performed between the present study's results and the best-reported literature benchmarks for each modality, using the concordance correlation coefficient (CCC) as the primary evaluation metric. Where a literature paper reported both a CCC and a comparable sample count for independent recordings,

Our multimodal (FER + GSR + corrected pupil) results were:

$$\text{Valence } = 0.913 \pm 0.041$$

$$\text{Arousal } = 0.865 \pm 0.061$$

For comparison, Patania et al. [362], using RECOLA with late fusion on the evaluation set, reported substantially lower CCCs (valence = 0.424 $\pm$ 0.203, arousal = 0.585 $\pm$ 0.114). Joudeh et al. [339] reported extremely high physiological CCCs on RECOLA (valence = 0.996, arousal = 0.998), but these results were obtained from a restricted subset of 18 recordings and with a different bio-signals integration. Therefore, while their values are not directly comparable to our pipeline, they highlight the methodological diversity in the field.

Additionally, the comparative analysis of unimodal performance against existing literature highlights important patterns. FER alone underperformed relative to the benchmarks reported by Raju et al. [360], likely because their study combined video and audio modalities, whereas our approach relied solely on facial video data. In contrast, both pupil-based and GSR-based models achieved substantially higher CCCs than those previously reported in similar unimodal studies by O'Dwyer et al. [359] and Brady et al. [361], with differences confirmed as statistically significant. This suggests that physiological measures provide more stable and reliable indicators of affective states than facial expressions alone, which are often influenced by individual variability and cultural modulation. Importantly, these findings validate our decision to integrate FER, corrected pupil size and GSR into the multimodal framework, as their robustness complements the weaker performance of FER. By combining these modalities within a unified machine learning pipeline, our approach not only addresses the limitations of unimodal FER but also establishes a novel configuration that yields interpretable, generalisable, and state-of-the-art performance in continuous valence–arousal prediction.

Moreover, our system design incorporates practical considerations to address key limitations in multimodal emotion recognition research. We selected audiovisual stimuli based on Russell's Circumplex model of affect [32], ensuring an equal distribution of videos across all four quadrants (high/low arousal and positive/negative valence) to mitigate the data imbalance problem discussed earlier. The video clips were prescreened through a survey, which helped to guide the selection process by establishing a general emotional label for each clip. However, our aim was not to elicit a specific target emotion but rather to identify the actual emotional response evoked in each participant, which may differ from the intended emotion of the clip (e.g., a fearful video may elicit happiness in some individuals, or a happy video may elicit sadness). To account for this subjectivity, we used INDSCAL, which provides an individual-level scaling of emotional perception. At the same time, the survey served as a general baseline for the

expected emotional category of each clip. To enhance robustness against inter-subject variability, we applied LOPO Cross-Validation. We relied on physiological signals, such as pupil size and GSR, which are less influenced by cultural factors, thereby improving generalizability, as discussed in "Modelling Challenges and Motivation for Regression" [36], [258]. Instead of using heuristic fusion approaches, we adopted XGBoost for feature-level fusion, allowing the model to effectively manage feature interactions, redundancy, and importance [250], [401]. The computational efficiency of XGBoost also ensures suitability for real-time deployment on mobile or wearable devices, supporting seamless integration into practical applications [252], [259]. We implemented frame-synchronised data acquisition to reduce cross-modal misalignment and treated pre- and post-stimulus intervals separately, maintaining temporal consistency [107], [253].

Our study also advances reproducibility in emotion detection research by implementing a validated emotion elicitation protocol supported by participant self-reports and conducting data collection within a controlled, multimodal framework. These practices directly address persistent concerns in affective computing related to data consistency and replicability [192], [366].

Future work should explore hybrid models that combine structured feature learners (e.g., XGBoost) with temporal or deep learning architectures (e.g., LSTMs or CNNs), allowing for efficient feature selection and dynamic sequence modeling [402]. Extending validation to ecologically valid, real-world environments (e.g., everyday settings, mobile or wearable devices) is essential for assessing the generalisability and robustness of affective models outside controlled laboratory conditions. These directions aim to enhance the accuracy and practical utility of emotion recognition systems.

Developing diverse datasets and addressing challenges such as signal synchronisation and noise is crucial to building universal and explainable affective computing systems for real-time use, such as in clinical settings. Different modalities (e.g., GSR, FER, pupil size) often operate on various time scales, making alignment difficult. In addition, real-world environments introduce significant noise from motion, lighting, and sensor variability. Future systems must be designed to address these issues through improved pre-processing and robust model architectures that perform reliably in uncontrolled laboratory conditions.

5.2.6 Exploring the Impact of Multimodal Emotion Detection in Clinical Settings

The integrated emotion detection model, which combines FER, pupil size, and GSR, presents a promising solution for monitoring emotional states in clinical and rehabilitation settings. Each of these modalities offers unique insights that, when combined, provide a more complete and accurate understanding of a patient's emotional

experience during therapy [403]. The integration of these systems enables continuous, real-time observation of physiological responses, thereby enhancing clinical assessment without interrupting the therapeutic process.

The real-time data collected from FER, pupil size, and GSR provides an objective, data-driven basis for understanding emotional responses that may not be readily expressed through verbal or visible cues alone. For instance, FER helps identify subtle facial expressions that may reveal suppressed emotions. At the same time, pupil dilation and GSR provide insights into emotional arousal and stress that patients may find difficult to articulate. These tools are particularly valuable in psychotherapy, where patients might mask their true feelings or experience difficulties in emotional expression, as seen in conditions like Post-Traumatic Stress Disorder (PTSD), depression, and autism spectrum disorders [343], [351].

The multimodal approach also shows great promise for neurorehabilitation, especially in tracking emotional recovery following events like stroke or TBI [354]. Therapists can better assess progress and adjust real-time interventions by monitoring emotional reactivity. Furthermore, the model's passive nature ensures minimal patient compliance burden, making it well-suited for routine monitoring and long-term emotional tracking [357].

As the technology evolves, its applications could expand beyond in-clinic settings. For example, remote emotional monitoring via telehealth platforms could help patients manage their emotional states during therapy sessions. In the future, integrating this emotion-aware system with AI-driven therapeutic assistants could allow for adaptive interventions based on real-time emotional data, offering personalised therapy adjustments or biofeedback [101], [404]. Moreover, this system could be integrated into neurofeedback-based rehabilitation programs, further improving the personalisation of treatment plans.

Our model also has some limitations, listed in the following section, which can be addressed to improve it in the future.

5.3 Limitations and Future Work

Although this research addresses several key challenges, some areas can be further improved to enhance the model's performance and applicability.

Adding More Physiological Signals like PPG-Based HR and EEG. Incorporating additional physiological signals, such as PPG-based HR and EEG, is a promising direction to further enhance the robustness and accuracy of the multimodal emotion detection model. PPG can provide valuable information on HRV, which reflects sympathetic and parasympathetic nervous system activity, closely linked to emotional arousal. HRV met-

rics such as Root Mean Square of Successive Differences (RMSSD), Low Frequency (LF)/High Frequency (HF) ratio, and Standard Deviation of Normal-to-Normal intervals (SDNN) can help distinguish between stress, calmness, and other affective states. We performed a preliminary analysis of PPG-based heart rate (HR) by decomposing the signal into three frequency bands-very low frequency (VLF: 0-0.04 Hz), low frequency (LF: 0.04-0.15 Hz), and high frequency (HF: 0.15-0.5 Hz)-for each stimulus and participant. Following Rakshit et al. [72], who demonstrated that LF and HF components are indicative of sympathetic and parasympathetic nervous system activity during emotional responses, we examined patterns of HR variability to assess their relationship with different emotional states. Detailed results are presented in the Appendices chapter 6. Although preliminary PPG-based HR analysis has been conducted, further work is needed to extract relevant emotion-related features from the PPG signal and integrate them into the existing multimodal framework for improved emotion detection. Meanwhile, EEG captures cortical brain activity and can provide real-time insights into emotional valence and arousal through frequency-domain features, such as alpha asymmetry and frontal theta power. Integrating these modalities with signals such as GSR, FER, and pupil size may improve model generalisability and resilience to noise or signal loss from any source. Additionally, combining central (EEG) and peripheral (GSR, PPG) physiological signals enables a more comprehensive understanding of the emotional response, making the system more suitable for sensitive applications such as clinical diagnostics, therapy monitoring, and adaptive HCI.

Enhancing GSR Feature Extraction for Emotion Detection. To improve the contribution of GSR signals in multimodal emotion detection, several advanced signal processing and feature extraction techniques can be explored in future work. Instead of relying solely on preprocessed or downsampled GSR signals, raw GSR data can be utilised alongside custom noise filtering methods tailored to preserve emotionally relevant signal components [405]. A custom approach to isolating the phasic component of the GSR signal can improve temporal precision, particularly for detecting stimulus-locked arousal changes [406].

Further enhancement could improve event-related GSR feature extraction by implementing customised peak detection algorithms that adapt to individual variability through dynamic, data-driven thresholds. These adaptive methods promise to identify genuine arousal-related responses more accurately while reducing false detections caused by noise or baseline fluctuations.

Moreover, incorporating time–frequency analysis on stimulus-locked windows, especially around emotionally salient peaks, may help uncover transient spectral patterns associated with autonomic arousal [407], [408]. An auspicious direction is the computation of phasic signal power within the 0–30 Hz range, segmented into 4 Hz bands.

Although this spectral decomposition is more commonly used in EEG-based emotion recognition [154], [192], its adaptation to GSR could provide new insights into frequency-specific signatures of sympathetic nervous system activity. This approach may reveal nuanced temporal features that are often overlooked by traditional time-domain analysis, thereby enhancing the model's ability to detect subtle patterns of emotional arousal.

Given the delayed nature of GSR responses, it is also essential to account for carry-over effects by extending the analysis window a few seconds beyond the stimulus offset [212], [258], [392]. This adjustment helps to capture late GSR responses, particularly in cases where emotional peaks occur toward the end of a clip.

Beyond time-domain and frequency-domain features, additional feature types, such as Hjorth parameters, higher-order crossings, spectral power features, and signal energy, can offer richer representations of GSR dynamics [305]. These features may provide greater sensitivity to subtle emotional shifts, thereby enhancing model performance in distinguishing between arousal levels.

These enhancements could maximise the utility of GSR signals by capturing both global and transient emotional patterns, improving the temporal resolution of arousal detection and enabling more robust multimodal fusion in emotion recognition systems.

Exploring Hybrid Fusion Models for Enhanced Emotion Detection. While XGBoost-based models have demonstrated strong performance in emotion detection due to their efficiency, scalability, and ability to handle non-linear relationships, they may still fail to capture complex temporal dynamics and higher-order feature interactions in physiological signals [409]. Future work could focus on developing hybrid fusion models that combine the structured decision-making power of XGBoost with the deep feature representation capabilities of neural networks, such as LSTM networks or CNNs. These models are well-suited for time-series data and can learn temporal dependencies in emotional responses, which is particularly significant for signals such as GSR or pupil dynamics. Additionally, integrating transfer learning techniques can improve generalisation by leveraging pre-trained models on similar affective datasets, reducing the need for large-scale emotion-specific data collection. Such hybrid models could significantly enhance multimodal emotion detection systems' robustness, accuracy, and adaptability, especially in real-time or clinical applications where subtle emotional cues must be detected reliably.

Enhancing FER Feature Extraction for Emotion Detection. Although FER did not significantly contribute to our current analysis, incorporating a broader range of emotional indicators beyond basic emotions may enhance performance. In this study, we used seven basic emotions—anger, fear, disgust, sadness, contempt, joy, and surprise—extracted via iMotions software. However, iMotions also provides access to de-

tailed facial landmark data, including metrics such as lip movement, eyebrow position, and eye openness. These granular features can compute additional affective indicators or construct continuous, vector-based representations of facial expressions. Incorporating enriched feature sets could provide more nuanced information for emotion detection and potentially improve model accuracy in future work.

Checking Applicability in Clinical Settings. Testing the model with `Clinical group` participants will help us understand the behaviour of people with mental health problems regarding emotion recognition. Evaluating the model in naturalistic settings will further validate its real-world applicability, particularly in mental health screening and monitoring, clinical neurorehabilitation, and the development of protocols for emotional regulation.

Chapter 6

Conclusion

This thesis presents the development of a highly accurate, multimodal regression model for continuous emotion detection, integrating FER, pupil size, and GSR to enhance the robustness of emotion arousal detection. Supported by methodologically rigorous preprocessing and signal correction strategies, this approach overcomes several existing limitations in emotion detection. A significant innovation in this work is the introduction of a novel, user-friendly pupil size correction model, which outperforms existing solutions and demonstrates considerable potential for widespread application in real-world systems.

Our model's performance demonstrates strong accuracy, with a 91% correlation for valence and 78% for arousal. This validates the effectiveness of our multimodal approach, which captures a broader range of emotional responses than traditional unimodal systems. This integration addresses the limitations of modality-specific approaches, providing a more nuanced and reliable emotion detection solution. Our approach has established a solid foundation for more accurate and versatile emotion detection systems. The ability to separate emotional signals from extraneous factors, such as lighting changes, and to integrate diverse modalities offers exciting potential for real-world applications.

The potential for further refinement and expansion of this integrated technology is vast. As our understanding of human emotions continues to deepen, the applications of such emotion-aware systems are poised to extend into fields like mental health monitoring, HCI, adaptive technologies, and neurorehabilitation. This technology offers a valuable tool for real-time emotional monitoring in clinical settings, supporting personalised treatments for emotional disorders and aiding in more accurate assessments. By combining FER, pupil size measurement, and GSR, the system functions as a passive, non-invasive clinical assistant during psychotherapy and other therapeutic sessions. It unobtrusively tracks emotional responses, capturing subtle facial expressions, shifts in arousal via pupil dilation, and physiological changes through GSR. After each session, therapists receive a comprehensive visual timeline that maps the patient's emotional

intensity and fluctuations, helping identify key moments of suppressed emotions, emotional disengagement, or stress spikes. This data-driven approach can significantly enhance diagnosis accuracy, provide deeper insights into emotional states, and enable tailored therapeutic interventions, ultimately improving patient care and outcomes. As the system evolves, integrating AI-driven therapeutic tools and remote monitoring platforms will enhance its potential for personalised, adaptive care. With minimal patient compliance required, it promises to revolutionise both psychological therapy and neurorehabilitation, offering new avenues for precision psychiatry and emotion-aware rehabilitation.

By addressing both the technical and human-centred challenges in emotion detection, this research lays the groundwork for the next generation of emotion-aware systems that are highly accurate but also user-friendly, accessible, and adaptable to a wide range of use cases. The promising results of this work hold great potential for future advancements in emotion detection, particularly in neurorehabilitation, clinical applications, and other domains of affective computing.

Appendices

6.1 Emotion Survey Questionnaire

The survey administered after each audiovisual clip consisted of 12 questions, each designed to measure the self-reported intensity of a specific emotion. The exact wording used in the study is listed below. Participants responded on a 10-point Likert-type scale ranging from 0 (not at all) to 9 (extremely).

1. To what extent did you feel **positive** while watching the clip?
2. To what extent did you feel **excited**?
3. To what extent did you feel **happy**?
4. To what extent did you feel **amused**?
5. To what extent did you feel **calm**?
6. To what extent did you feel **content**?
7. To what extent did you feel **negative**?
8. To what extent did you feel **angry**?
9. To what extent did you feel **afraid**?
10. To what extent did you feel **anxious**?
11. To what extent did you feel **sad**?
12. To what extent did you feel **bored**?

Each response was recorded on a 0–9 scale using the iMotions interface, where 0 indicated *not at all* and 9 indicated *extremely*.

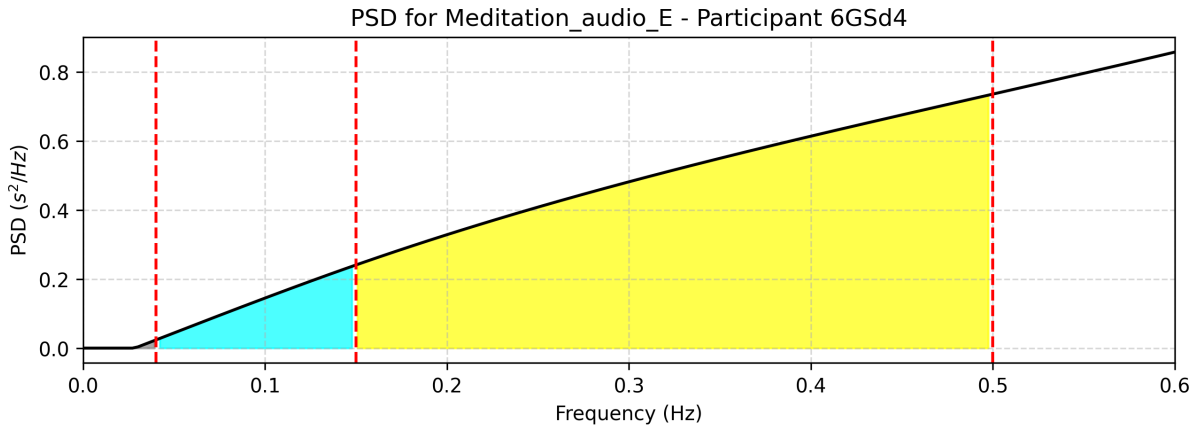
PPG-Based HR Frequency Band Patterns

To explore the relationship between emotional arousal and heart rate variability, we analysed PPG-based HR signals for each participant and each stimulus. The HR data were decomposed into three standard frequency bands: VLF (0-0.04 Hz), LF (0.04-0.15 Hz), and HF (0.15-0.5 Hz).

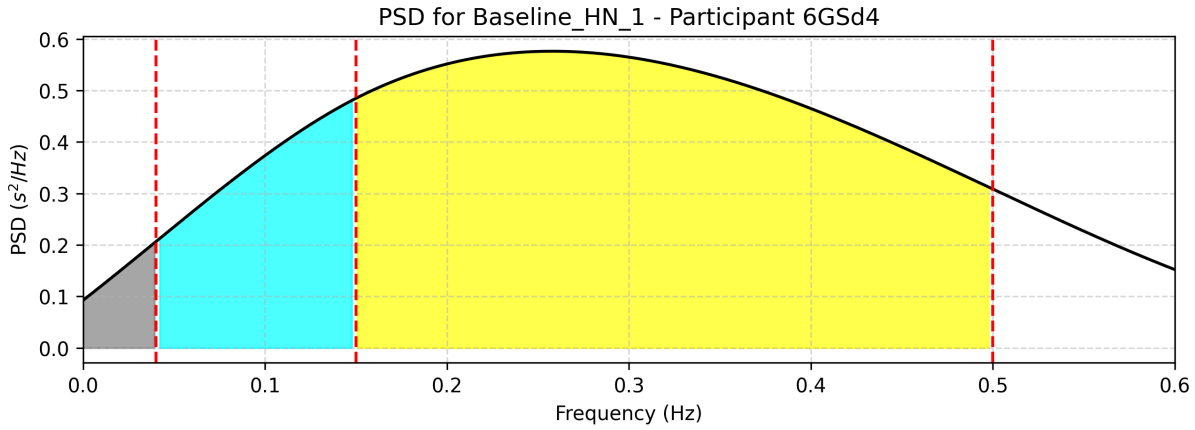
We plotted the PSD across these bands for all participants and stimuli. Below, we present a selection of representative plots from a few participants that clearly illustrate the differences between high- and low-arousal stimuli.

To begin the analysis, first, we check the relaxation pattern in the meditation audio and the 5-second neutral grey screen stimulus (baseline) presented before each audiovisual clip.

Figure 6.1 shows that, in the baseline stimulus (see Figure 6.1b), the LF band exhibits moderate power, reflecting baseline sympathetic activity. In contrast, the HF band is dominant, indicating strong parasympathetic influence typical of a resting state. The VLF band shows a slight but noticeable contribution, consistent with long-term physiolo-



(a) Meditation Audio



(b) 5-Second Neutral Grey Screen

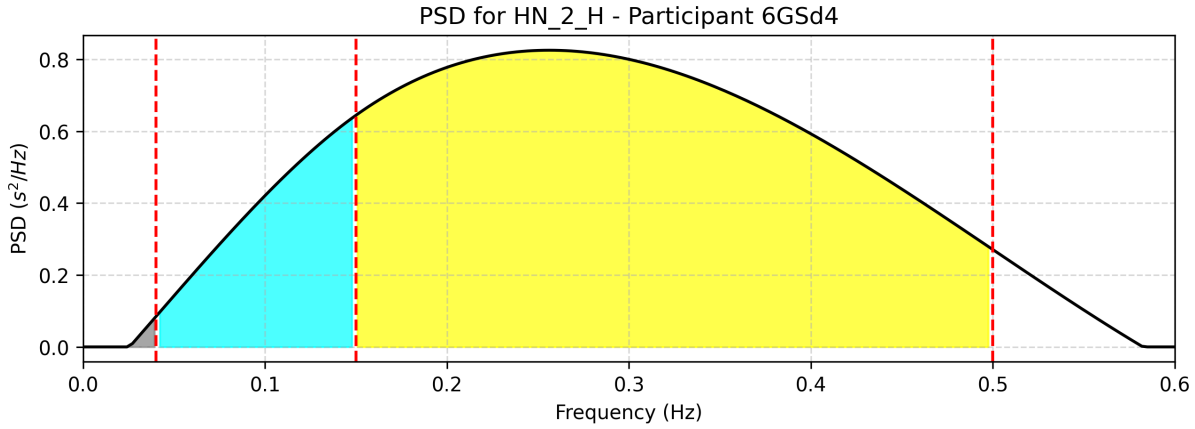
Figure 6.1: The PSD plots for PPG-based heart rate data from Participant `6GSd4` under meditation audio stimulus (a) and the 5-second neutral grey screen (presented before one emotional stimulus) (b) conditions reveal distinct differences in ANS activity across the standard frequency bands: VLF, LF, and HF, indicated by red dashed lines.

gical regulation under non-stimulated conditions.

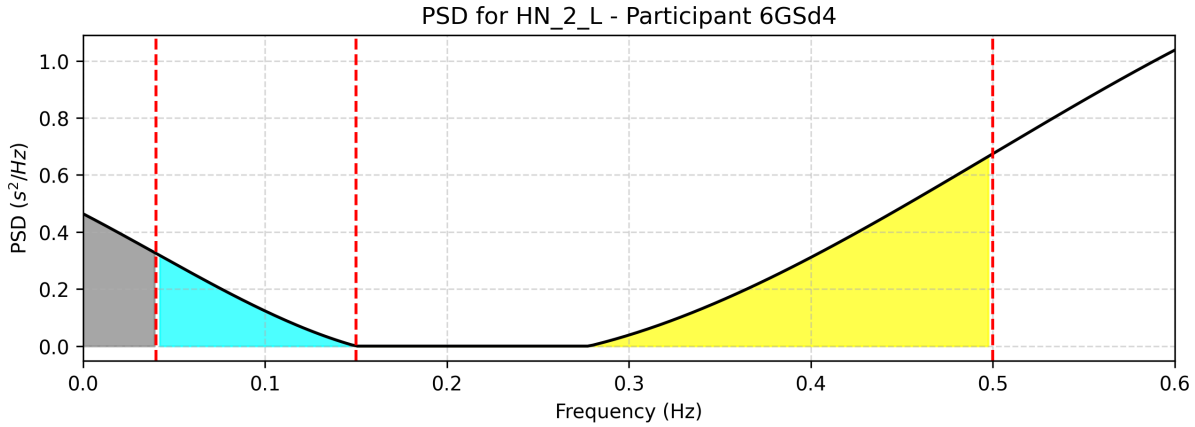
In contrast, the Meditation Audio stimulus (see figure 6.1a) exhibits reduced power in the LF band and an increase in the HF band, suggesting a shift toward enhanced parasympathetic activity and reduced sympathetic drive. The VLF band is nearly absent, which is expected in short-term recordings during relaxation-focused tasks. The gradual rise in power across the HF range further supports a calming physiological response associated with meditative audio.

The observed shift in spectral power from LF to HF between the baseline and meditation conditions suggests increased relaxation and parasympathetic dominance during meditation, consistent with the expected physiological outcomes of mindfulness practice. The meditation audio appears to induce greater relaxation than the baseline, likely due to its calming nature and the relatively short duration of the baseline, which may not have allowed participants to reach an actual resting state. This highlights the im-

portance of selecting appropriate baseline conditions when using them to normalise emotional data. If the baseline does not accurately reflect a resting physiological state, due to factors such as insufficient duration or anticipatory effects, it may distort comparisons and affect interpretation. Future studies should evaluate different baseline types (e.g., prolonged rest, neutral content, or guided relaxation) to determine which most reliably represents an individual's physiological baseline, thereby increasing the validity of emotion recognition models.



(a) High Arousal Stimulus



(b) Low Arousal Stimulus

Figure 6.2: The PSD plots for PPG-based heart rate data from Participant `6GSd4` under high arousal Stimulus (a) and low arousal stimulus (b) reveal distinct differences across the standard frequency bands: VLF, LF, and HF, as indicated by the red dashed lines.

The Figure 6.2 shows that, in the high arousal condition, the Figure 6.2b, the LF band (shaded cyan) shows a substantial increase in power, suggesting heightened sympathetic nervous system activity typically associated with increased arousal. The HF band (shaded yellow) also contributes significantly but less dominantly than LF, indicating some parasympathetic modulation. The VLF band (shaded grey) shows minimal contribution, as is common in short-duration recordings.

In contrast, the low arousal condition, the Figure 6.2b presents a different spectral

profile. The LF band exhibits significantly lower power compared to the high arousal condition, while the HF band maintains a relatively stronger presence, indicating increased parasympathetic influence. Interestingly, the VLF band exhibits higher power in the low arousal condition, which may reflect baseline regulatory processes that are more prominent during rest or low engagement.

The shift in spectral power distribution between these two conditions aligns with expected physiological responses to different arousal states. High arousal is characterised by dominant LF activity, while low arousal shows a more balanced or parasympathetic-dominant pattern with elevated VLF contribution. This supports the effectiveness of frequency domain analysis of PPG signals for distinguishing emotional arousal levels.

Similarly, we extended this analysis to data from 47 participants, and the results consistently showed a clear distinction between low- and high-arousal stimuli in terms of ANS activity, particularly as reflected in PPG-based frequency domain features. These consistent patterns support the reliability of using physiological signals to differentiate emotional arousal levels. Building on these findings, the next step is to systematically extract the most relevant and discriminative features from these physiological signals across time, frequency, and time-frequency domains to train a robust emotion detection model. Emphasis will also be placed on selecting features that generalise well across participants while maintaining sensitivity to individual differences in emotional responses.

6.1.1 Result of Ground Truth computation using Euclidean and Manhattan Distance Metrics

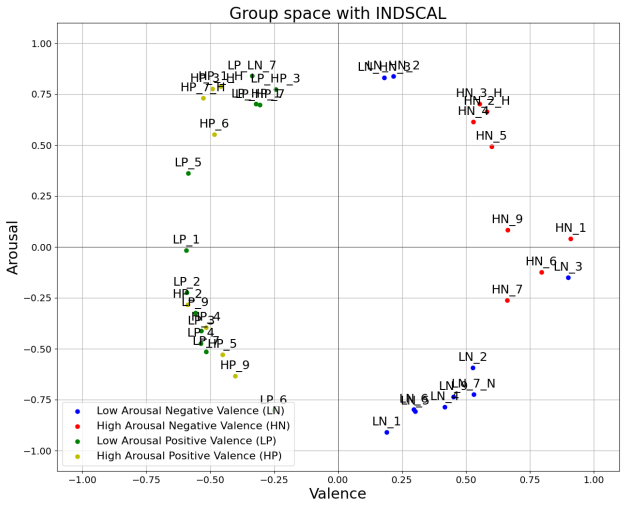


Figure 6.3: Ground Truth using Manhattan.

When comparing the two INDSCAL group space plots, the version in Figure 6.4 provides a more reliable ground truth for emotion recognition than the version in Figure 6.3. In the plot with Euclidean distance, the stimuli are more evenly distributed across the four quadrants of Russell's Circumplex Model, with High Arousal Positive Valence (HP), High Arousal Negative Valence (HN), Low Arousal Positive Valence (LP), and Low Arousal Negative Valence (LN) clearly separated. The wider range of valence and arousal values (approximately –2 to +2) enhances the bipolar structure of the model and reduces overlap between categories. By contrast, the plot with Manhattan distance compresses the axes (–1 to +1), causing stimuli to cluster near the centre and along the upper band, which blurs the quadrant boundaries and increases category overlap. Consequently, the PDF representation aligns more closely with the theoretical circular structure of the circumplex, providing cleaner quadrant separation and a stronger ground truth for training emotion recognition models.

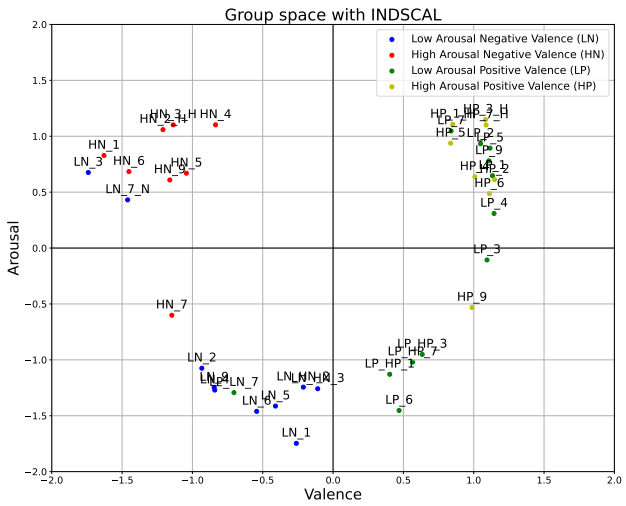


Figure 6.4: Ground Truth using Euclidean Distance Metrics.

Bibliography

[1] J. Esbjörnsson, *Emo-a computational emotional state module: Emotions and their influence on the behaviour of autonomous agents*, 2007.

[2] J. Storbeck and G. L. Clore, "On the interdependence of cognition and emotion," *Cognition and emotion*, vol. 21, no. 6, pp. 1212–1237, 2007.

[3] J. Luo and R. Yu, "Follow the heart or the head? the interactive influence model of emotion and cognition," *Frontiers in psychology*, vol. 6, p. 573, 2015.

[4] T. Brosch, K. Scherer, D. Grandjean and D. Sander, "The impact of emotion on perception, attention, memory, and decision-making," *Swiss medical weekly*, vol. 143, no. 1920, w13786–w13786, 2013.

[5] C. M. Tyng, H. U. Amin, M. N. Saad and A. S. Malik, "The influences of emotion on learning and memory," *Frontiers in psychology*, vol. 8, p. 235 933, 2017.

[6] E. Sadler-Smith and E. Shefy, "The intuitive executive: Understanding and applying 'gut feel'in decision-making," *Academy of Management Perspectives*, vol. 18, no. 4, pp. 76–91, 2004.

[7] A. ROJO, "20 the role of emotions," *The Handbook of Translation and Cognition*, p. 369, 2017.

[8] S. B. Osman *et al.*, "Analysis of teacher job satisfaction through spiritual intelligence and emotional intelligence in palu, central sulawesi, indonesia," *International Journal of Social Science and Business*, vol. 8, no. 2, pp. 328–336, 2024.

[9] A. Kumari and S. Gupta, "A study of emotional intelligence and frustration tolerance among adolescent," *Advance research journal of social science*, vol. 6, 2015.

[10] J. Agbeniga, K. O. Ayodele, A. O. Adeoye and O. Oyerinde, "Self-efficacy, emotional intelligence and achievement motivation as predictors of impulsive behaviour among secondary school students," *Editorial Board*, p. 45, 2015.

[11] R. Brooks, S. Brooks and S. Goldstein, "The power of mindsets: Nurturing engagement, motivation, and resilience in students," in *Handbook of research on student engagement*, Springer, 2012, pp. 541–562.

[12] D. Scott Ridley, "Reflective self-awareness: A basic motivational process," *The Journal of Experimental Education*, vol. 60, no. 1, pp. 31–48, 1991.

[13] W. Xiao, L. Quanliang, Y. Huanli and S. Banerjee, "The role of emotional intelligence in leadership and team dynamics," *International Journal For Multidisciplinary Research (IJFMR)*, vol. 5, pp. 1–2, 2023.

[14] S. H. Zarit and A. B. Edwards, "Family caregiving: Research and clinical intervention," *Handbook of the clinical psychology of ageing*, pp. 331–368, 1996.

[15] U.-S. Donges, A. Kersting and T. Suslow, "Alexithymia and perception of emotional information: A review of experimental psychological findings," *Universitas Psychologica*, vol. 13, no. 2, pp. 745–756, 2014.

[16] A. N. Da Silva, A. B. Vasco and J. C. Watson, "Alexithymia and emotional processing: A longitudinal mixed methods research," *Research in Psychotherapy: Psychopathology, Process, and Outcome*, vol. 21, no. 1, p. 292, 2018.

[17] V. Gray, K. M. Douglas and R. J. Porter, "Emotion processing in depression and anxiety disorders in older adults: Systematic review," *BJPsych Open*, vol. 7, no. 1, e7, 2021.

[18] M. E. Kret and A. Ploeger, "Emotion processing deficits: A liability spectrum providing insight into comorbidity of mental disorders," *Neuroscience & Biobehavioral Reviews*, vol. 52, pp. 153–171, 2015.

[19] F. Kılıç, A. Demirdaş, Ü. Işık, M. Akkuş, I. M. Atay and D. Kuzugüdenlioğlu, "Empathy, alexithymia, and theory of mind in borderline personality disorder," *The Journal of Nervous and Mental Disease*, vol. 208, no. 9, pp. 736–741, 2020.

[20] E. Onur, T. Alkın, M. J. Sheridan and T. N. Wise, "Alexithymia and emotional intelligence in patients with panic disorder, generalized anxiety disorder and major depressive disorder," *Psychiatric Quarterly*, vol. 84, pp. 303–311, 2013.

[21] K. W. Grant, "The solitary spectrum: Insights into schizoid personality disorder.,"

[22] S. L. Bistricky, R. E. Ingram and R. A. Atchley, "Facial affect processing and depression susceptibility: Cognitive biases and cognitive neuroscience.," *Psychological bulletin*, vol. 137, no. 6, p. 998, 2011.

[23] R. D. Lane, C. Subic-Wrana, L. Greenberg and I. Yovel, "The role of enhanced emotional awareness in promoting change across psychotherapy modalities.," *Journal of Psychotherapy Integration*, vol. 32, no. 2, p. 131, 2022.

[24] N. S. Schutte, J. M. Malouff and E. B. Thorsteinsson, "Increasing emotional intelligence through training: Current status and future directions," 2013.

[25] D. Sharma, "Empathetic technology: Integrating emotional intelligence into assistive devices for aging adults and individuals with disabilities," in *Assistive Technology Solutions for Aging Adults and Individuals With Disabilities*, IGI Global Scientific Publishing, 2025, pp. 73–88.

[26] J. I. Montana, M. Matamala-Gomez, M. Maisto *et al.*, "The benefits of emotion regulation interventions in virtual reality for the improvement of wellbeing in

adults and older adults: A systematic review," *Journal of clinical medicine*, vol. 9, no. 2, p. 500, 2020.

[27] P. Slovák and G. Fitzpatrick, "Teaching and developing social and emotional skills with technology," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 22, no. 4, pp. 1–34, 2015.

[28] T. Brosch, G. Pourtois and D. Sander, "The perception and categorisation of emotional stimuli: A review," *Cognition and emotion*, vol. 24, no. 3, pp. 377–400, 2010.

[29] E. Mower, M. J. Matarić and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1057–1070, 2010.

[30] N. J. Shoumy, L.-M. Ang, K. P. Seng, D. M. Rahaman and T. Zia, "Multimodal big data affective analytics: A comprehensive survey using text, audio, visual and physiological signals," *Journal of Network and Computer Applications*, vol. 149, p. 102 447, 2020.

[31] S. Kumar G. S., A. Arun, N. Sampathila and R. Vinoth, "Machine learning models for classification of human emotions using multivariate brain signals," *Computers*, vol. 11, no. 10, 2022, ISSN: 2073-431X. DOI: 10.3390/computers11100152. [Online]. Available: https://www.mdpi.com/2073-431X/11/10/152.

[32] J. A. Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[33] S. Nowak and S. Rüger, "How reliable are annotations via crowdsourcing: A study about inter-annotator agreement for multi-label image annotation," in *Proceedings of the international conference on Multimedia information retrieval*, 2010, pp. 557–566.

[34] J. R. Fontaine, K. R. Scherer, E. B. Roesch and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological science*, vol. 18, no. 12, pp. 1050–1057, 2007.

[35] J. Zhang, Z. Yin, P. Chen and S. Nichele, "Emotion recognition using multimodal data and machine learning techniques: A tutorial and review," *Information Fusion*, vol. 59, pp. 103–126, 2020.

[36] M. Egger, M. Ley and S. Hanke, "Emotion recognition from physiological signal analysis: A review," *Electronic Notes in Theoretical Computer Science*, vol. 343, pp. 35–55, 2019.

[37] K. H. Kim, S. W. Bang and S. R. Kim, "Emotion recognition system using short-term monitoring of physiological signals," *Medical and biological engineering and computing*, vol. 42, pp. 419–427, 2004.

[38] J. Tejedor, C. A. García, D. G. Márquez, R. Raya and A. Otero, "Multiple physiological signals fusion techniques for improving heartbeat detection: A review," *Sensors*, vol. 19, no. 21, p. 4708, 2019.

[39] J. O. Johnson, "Autonomic nervous system: Physiology," in *Pharmacology and physiology for anesthesia*, Elsevier, 2019, pp. 270–281.

[40] G. Manion and T. Stokkermans, "The effect of pupil size on visual resolution," *StatPearls*, 2024.

[41] M. Guillon, K. Dumbleton, P. Theodoratos, M. Gobbe, C. B. Wooley and K. Moody, "The effects of age, refractive status, and luminance on pupil size," *Optometry and vision science*, vol. 93, no. 9, pp. 1093–1100, 2016.

[42] F. Maqsood, "Effects of varying light conditions and refractive error on pupil size," *Cogent Medicine*, vol. 4, no. 1, p. 1 338 824, 2017.

[43] W. Einhäuser, "The pupil as marker of cognitive processes," *Computational and cognitive neuroscience of vision*, pp. 141–169, 2017.

[44] R. F. Stanners, M. Coulter, A. W. Sweet and P. Murphy, "The pupillary response as an indicator of arousal and cognition," *Motivation and Emotion*, vol. 3, pp. 319–340, 1979.

[45] V. L. Kinner, L. Kuchinke, A. M. Dierolf, C. J. Merz, T. Otto and O. T. Wolf, "What our eyes tell us about feelings: Tracking pupillary responses during emotion regulation processes," *Psychophysiology*, vol. 54, no. 4, pp. 508–518, 2017.

[46] P. Jerčić, C. Sennersten and C. Lindley, "Modeling cognitive load and physiological arousal through pupil diameter and heart rate," *Multimedia Tools and Applications*, vol. 79, no. 5, pp. 3145–3159, 2020.

[47] T. Kosch, J. Karolus, J. Zagermann, H. Reiterer, A. Schmidt and P. W. Woźniak, "A survey on measuring cognitive workload in human-computer interaction," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–39, 2023.

[48] C. A. Hall and R. P. Chilcott, "Eyeing up the future of the pupillary light reflex in neurodiagnostics," *Diagnostics*, vol. 8, no. 1, p. 19, 2018.

[49] A. H. Firth, "The pupil and its reflexes in insanity," *Journal of Mental Science*, vol. 60, no. 249, pp. 224–277, 1914.

[50] B. Mahanama, Y. Jayawardana, S. Rengarajan *et al.*, "Eye movement and pupil measures: A review," *frontiers in Computer Science*, vol. 3, p. 733 531, 2022.

[51] H. Singh and J. Singh, "Human eye tracking and related issues: A review," *International Journal of Scientific and Research Publications*, vol. 2, no. 9, pp. 1–9, 2012.

[52] D. C. Niehorster, R. Andersson and M. Nyström, "Titta: A toolbox for creating psychtoolbox and psychopy experiments with tobii eye trackers," *Behavior research methods*, vol. 52, pp. 1970–1979, 2020.

[53] L. Zhang and H. Cui, "Reliability of muse 2 and tobii pro nano at capturing mobile application users' real-time cognitive workload changes," *Frontiers in Neuroscience*, vol. 16, p. 1 011 475, 2022.

[54] S. Chen and J. Epps, "Automatic classification of eye activity for cognitive load measurement with emotion interference," *Computer methods and programs in biomedicine*, vol. 110, no. 2, pp. 111–124, 2013.

[55] C. Kosel, S. Michel, T. Seidel and M. Foerster, "Exploring the dynamic interplay of cognitive load and emotional arousal by using multimodal measurements: Correlation of pupil diameter and emotional arousal in emotionally engaging tasks," *arXiv preprint arXiv:2403.00366*, 2024.

[56] P. Van der Wel and H. Van Steenbergen, "Pupil dilation as an index of effort in cognitive control tasks: A review," *Psychonomic bulletin & review*, vol. 25, pp. 2005–2015, 2018.

[57] V. Skaramagkas, G. Giannakakis, E. Ktistakis *et al.*, "Review of eye tracking metrics involved in emotional and cognitive processes," *IEEE Reviews in Biomedical Engineering*, vol. 16, pp. 260–277, 2021.

[58] J. W. Chen, Z. J. Gombart, S. Rogers, S. K. Gardiner, S. Cecil and R. M. Bullock, "Pupillary reactivity as an early indicator of increased intracranial pressure: The introduction of the neurological pupil index," *Surgical neurology international*, vol. 2, 2011.

[59] J. Halszka, K. Holmqvist and H. Gruber, "Eye tracking in educational science: Theoretical frameworks and research agendas.," *Journal of eye movement research*, vol. 10, no. 1, 2017.

[60] J. L. Rosch and J. J. Vogel-Walcutt, "A review of eye-tracking applications as tools for training," *Cognition, technology & work*, vol. 15, pp. 313–327, 2013.

[61] J. Widacki, "Discoverers of the galvanic skin response," *European Polygraph*, vol. 9, no. 4 (34), pp. 209–221, 2015.

[62] C. Tronstad, M. Amini, D. R. Bach and Ø. G. Martinsen, "Current trends and opportunities in the methodology of electrodermal activity measurement," *Physiological measurement*, vol. 43, no. 2, 02TR01, 2022.

[63] J. Montagu and E. M. Coles, "Mechanism and measurement of the galvanic skin response.," *Psychological Bulletin*, vol. 65, no. 5, p. 261, 1966.

[64] J. L. McGaugh, *Emotions and bodily responses: A psychophysiological approach*. Academic Press, 2013.

[65] C. Latulipe, E. A. Carroll and D. Lottridge, "Love, hate, arousal and engagement: Exploring audience responses to performing arts," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2011, pp. 1845–1854.

[66] C. P. Richter, "Physiological factors involved in the electrical resistance of the skin," *American Journal of Physiology-Legacy Content*, vol. 88, no. 4, pp. 596–615, 1929.

[67] M. Memar and A. Mokaribolhassan, "Stress level classification using statistical analysis of skin conductance signal while driving," *SN Applied Sciences*, vol. 3, no. 1, p. 64, 2021.

[68] A. Greco, G. Valenza, L. Citi and E. P. Scilingo, "Arousal and valence recognition of affective sounds based on electrodermal activity," *IEEE Sensors Journal*, vol. 17, no. 3, pp. 716–725, 2016.

[69] R. Kavya, N. Nayana, K. B. Karangale, H. Madhura and S. Sheela, "Photoplethysmography—a modern approach and applications," in *2020 International Conference for Emerging Technology (INCET)*, IEEE, 2020, pp. 1–4.

[70] M. Nitzan and Z. Ovadia-Blechman, "Physical and physiological interpretations of the ppg signal," in *Photoplethysmography*, Elsevier, 2022, pp. 319–340.

[71] J. Přibil, A. Přibilová and I. Frollo, "Comparative measurement of the ppg signal on different human body positions by sensors working in reflection and transmission modes," *Engineering proceedings*, vol. 2, no. 1, p. 69, 2020.

[72] R. Rakshit, V. R. Reddy and P. Deshpande, "Emotion detection and recognition using hrv features derived from photoplethysmogram signals," in *Proceedings of the 2nd workshop on Emotion Representations and Modelling for Companion Systems*, 2016, pp. 1–6.

[73] D. P. Tobón Vallejo and A. El Saddik, "Emotional states detection approaches based on physiological signals for healthcare applications: A review," *Connected Health in Smart Cities*, pp. 47–74, 2020.

[74] M.-A. Savard, R. Merlo, A. Samithamby, A. Paas and E. B. Coffey, "Approaches to studying emotion using physiological responses to spoken narratives: A scoping review," *Psychophysiology*, e14642, 2024.

[75] J. Fine, K. L. Branan, A. J. Rodriguez *et al.*, "Sources of inaccuracy in photoplethysmography for continuous cardiovascular monitoring," *Biosensors*, vol. 11, no. 4, p. 126, 2021.

[76] P. Ekman, "Facial expression and emotion.," *American psychologist*, vol. 48, no. 4, p. 384, 1993.

[77] iMotions, *Facial expression analysis module - imotions lab*, Accessed: 2025-02-04, 2025. [Online]. Available: `https://imotions.com/products/imotions-lab/modules/fea-facial-expression-analysis/`.

[78] B. C. Ko, "A brief review of facial emotion recognition based on visual information," *sensors*, vol. 18, no. 2, p. 401, 2018.

[79] P. Giannopoulos, I. Perikos and I. Hatzilygeroudis, "Deep learning approaches for facial emotion recognition: A case study on fer-2013," *Advances in hybridization of intelligent methods: Models, systems and applications*, pp. 1–16, 2018.

[80] A. Ryan, J. F. Cohn, S. Lucey *et al.*, "Automated facial expression recognition system," in *43rd annual 2009 international Carnahan conference on security technology*, IEEE, 2009, pp. 172–177.

[81] M. V. Cruz, S. Jamal and S. C. Sethuraman, "A comprehensive survey of brain-computer interface technology in healthcare: Research perspectives," 2024.

[82] A. Jain and A. Jain, "Ai-based emotion detection system in healthcare for patient," *Generative Artificial Intelligence for Biomedical and Smart Health Informatics*, pp. 455–470, 2025.

[83] C. Halkiopoulos, E. Gkintoni, A. Aroutzidis and H. Antonopoulou, "Advances in neuroimaging and deep learning for emotion detection: A systematic review of cognitive neuroscience and algorithmic innovations," *Diagnostics*, vol. 15, no. 4, p. 456, 2025.

[84] K. V. Keefer, "Self-report assessments of emotional competencies: A critical look at methods and meanings," *Journal of Psychoeducational Assessment*, vol. 33, no. 1, pp. 3–23, 2015.

[85] M. B. Solhan, T. J. Trull, S. Jahng and P. K. Wood, "Clinical assessment of affective instability: Comparing ema indices, questionnaire reports, and retrospective recall.," *Psychological assessment*, vol. 21, no. 3, p. 425, 2009.

[86] P. Bota, C. Wang, A. Fred and H. P. D. Silva, "A review, current challenges, and future possibilities on emotion recognition using machine learning and physiological signals," *IEEE Access*, vol. 7, pp. 140 990–141 020, 2019. DOI: 10.1109/ACCESS.2019.2944001.

[87] A. D'Agostino, S. Covanti, M. Rossi Monti and V. Starcevic, "Reconsidering emotion dysregulation," *Psychiatric Quarterly*, vol. 88, pp. 807–825, 2017.

[88] V. Vine, "The transdiagnostic, context-sensitive role of emotion identification in emotion regulation and psychological health.," Ph.D. dissertation, Yale University, 2016.

[89] R. Argent, A. Bevilacqua, A. Keogh, A. Daly and B. Caulfield, "The importance of real-world validation of machine learning systems in wearable exercise biofeedback platforms: A case study," *Sensors*, vol. 21, no. 7, p. 2346, 2021.

[90] C. M. Ilyas, "Facial emotion recognition for citizens with traumatic brain injury for therapeutic robot interaction," 2021.

[91] N. Nigar, "Speech emotion recognition using cnn and its use case in digital healthcare," *arXiv preprint arXiv:2406.10741*, 2024.

[92] I. El Naqa and M. J. Murphy, *What is machine learning?* Springer, 2015.

[93] G. S. Nadella, S. Satish, K. Meduri and S. S. Meduri, "A systematic literature review of advancements, challenges and future directions of ai and ml in healthcare," *International Journal of Machine Learning for Sustainable Development*, vol. 5, no. 3, pp. 115–130, 2023.

[94] A. Kilic, "Artificial intelligence and machine learning in cardiovascular health care," *The Annals of thoracic surgery*, vol. 109, no. 5, pp. 1323–1329, 2020.

[95] L. Zhou, S. Pan, J. Wang and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," *Neurocomputing*, vol. 237, pp. 350–361, 2017.

[96] I. H. Sarker, "Ai-based modeling: Techniques, applications and research issues towards automation, intelligent and smart systems," *SN Computer Science*, vol. 3, no. 2, p. 158, 2022.

[97] F. Schwenker and E. Trentin, "Pattern classification and clustering: A review of partially supervised learning approaches," *Pattern Recognition Letters*, vol. 37, pp. 4–14, 2014.

[98] V. Nasteski, "An overview of the supervised machine learning methods," *Horizons. b*, vol. 4, no. 51-62, p. 56, 2017.

[99] S. Graham, C. Depp, E. E. Lee *et al.*, "Artificial intelligence for mental health and mental illnesses: An overview," *Current psychiatry reports*, vol. 21, pp. 1–18, 2019.

[100] D. Ayata, Y. Yaslan and M. E. Kamasak, "Emotion recognition from multimodal physiological signals for emotion aware healthcare systems," *Journal of Medical and Biological Engineering*, vol. 40, pp. 149–157, 2020.

[101] F. Doctor, C. Karyotis, R. Iqbal and A. James, "An intelligent framework for emotion aware e-healthcare support systems," in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, IEEE, 2016, pp. 1–8.

[102] G. Orrù, A. Gemignani, R. Ciacchini, L. Bazzichi and C. Conversano, "Machine learning increases diagnosticity in psychometric evaluation of alexithymia in fibromyalgia," *Frontiers in medicine*, vol. 6, p. 319, 2020.

[103] S. McDonald and H. Genova, "The effect of severe traumatic brain injury on social cognition, emotion regulation, and mood," *Handbook of clinical neurology*, vol. 183, pp. 235–260, 2021.

[104] D. T. Cordaro, R. Sun, D. Keltner, S. Kamble, N. Huddar and G. McNeil, "Universals and cultural variations in 22 emotional expressions across five cultures.," *Emotion*, vol. 18, no. 1, p. 75, 2018.

[105] B. Mesquita and R. Walker, "Cultural differences in emotions: A context for interpreting emotional experiences," *Behaviour research and therapy*, vol. 41, no. 7, pp. 777–793, 2003.

[106] Z. Kövecses, *Emotion concepts*. Springer Science & Business Media, 2012.

[107] G. Udahemuka, K. Djouani and A. M. Kurien, "Multimodal emotion recognition using visual, vocal and physiological signals: A review," *Applied Sciences*, vol. 14, no. 17, p. 8071, 2024.

[108] T. Chutia and N. Baruah, "A review on emotion detection by using deep learning techniques," *Artificial Intelligence Review*, vol. 57, no. 8, p. 203, 2024.

[109] Z. Pansara, G. Navyte, H. Gillmeister, C. Cinel and V. De Feo, "Towards an accurate measure of emotional pupil dilation responses: A model for removing the effect of luminosity," in *2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*, IEEE, 2024, pp. 1–6.

[110] Y.-K. Lee, O.-W. Kwon, H. S. Shin, J. Jo and Y. Lee, "Noise reduction of ppg signals using a particle filter for robust emotion recognition," in *2011 IEEE International Conference on Consumer Electronics-Berlin (ICCE-Berlin)*, IEEE, 2011, pp. 202–205.

[111] M. K. Ahirwal, A. Kumar and G. K. Singh, "Eeg/erp adaptive noise canceller design with controlled search space (css) approach in cuckoo and other optimization algorithms," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 10, no. 6, pp. 1491–1504, 2013.

[112] Y. Huang, J. Xiao, K. Tian, A. Wu and G. Zhang, "Research on robustness of emotion recognition under environmental noise conditions," *IEEE Access*, vol. 7, pp. 142 009–142 021, 2019.

[113] F. Noroozi, M. Marjanovic, A. Njegus, S. Escalera and G. Anbarjafari, "Audio-visual emotion recognition in video clips," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 60–75, 2017.

[114] R. Joshi and M. Jadeja, "The synergy of clinical psychology and affective computing: Advancements in emotion recognition and therapy," in *Affective Computing for Social Good: Enhancing Well-being, Empathy, and Equity*, Springer, 2024, pp. 21–45.

[115] O. Higgins, B. L. Short, S. K. Chalup and R. L. Wilson, "Artificial intelligence (ai) and machine learning (ml) based decision support systems in mental health: An integrative review," *International Journal of Mental Health Nursing*, vol. 32, no. 4, pp. 966–978, 2023.

[116] C. Whitney, H. Preis and A. R. Vargas, "Anticipatory moral distress in machine learning-based clinical decision support tool development: A qualitative analysis," *SSM-Qualitative Research in Health*, p. 100 540, 2025.

[117] K. D. Kannan, S. K. Jagatheesaperumal, R. N. Kandala, M. Lotfaliany, R. Alizadehsanid and M. Mohebbi, "Advancements in machine learning and deep learning for early detection and management of mental health disorder," *arXiv preprint arXiv:2412.06147*, 2024.

[118] A. Katirai, "Ethical considerations in emotion recognition technologies: A review of the literature," *AI and Ethics*, vol. 4, no. 4, pp. 927–948, 2024.

[119] J. Sousa, S. Santos, L. André and J. Ferreira, "Converging affective computing and ethical challenges: The quest for universal access in human-machine cooperation," in *International Conference on Human-Computer Interaction*, Springer, 2024, pp. 106–121.

[120] C. E. Izard, *Human emotions*. Springer Science & Business Media, 2013.

[121] R. S. Lazarus, "Thoughts on the relations between emotion and cognition.," *American psychologist*, vol. 37, no. 9, p. 1019, 1982.

[122] A. S. Fox, R. C. Lapate, A. J. Shackman and R. J. Davidson, *The nature of emotion: Fundamental questions*. Oxford University Press, 2018.

[123] A. Buijs and A. Lawrence, "Emotional conflicts in rational forestry: Towards a research agenda for understanding emotions in environmental conflicts," *Forest Policy and Economics*, vol. 33, pp. 104–111, 2013.

[124] C. P. Sobel and P. Li, *The cognitive sciences: An interdisciplinary approach*. Sage Publications, 2013.

[125] E. Scharaga, "Cognitive-affective," *EPPP Step One Exam Review: Comprehensive Review, PLUS 450 Questions Based on the Latest Exam Blueprint*, p. 43, 2024.

[126] C. Lutz and G. M. White, "The anthropology of emotions," *Annual review of anthropology*, pp. 405–436, 1986.

[127] Oxford University Press, *Oxford english dictionary*, in *Oxford English Dictionary*, 2023. [Online]. Available: https://doi.org/10.1093/OED/2262908073.

[128] Cambridge University Press, *Cambridge dictionary: Emotion*, Online; accessed 8 July 2024, Definition of "emotion", 2024. [Online]. Available: https://dictionary.cambridge.org/dictionary/english/emotion.

[129] C. Darwin, *The Expression of Emotions in Man and Animals*. John Murray, 1872.

[130] O. E. Dror, "Deconstructing the "two factors": The historical origins of the schachter–singer theory of emotions," *Emotion Review*, vol. 9, no. 1, pp. 7–16, 2017.

[131] D. Hopkins, *Theorizing emotions: Sociological explorations and applications*. Campus Verlag, 2009.

[132] D. S. Massey, "A brief history of human society: The origin and role of emotion in social life," *American sociological review*, vol. 67, no. 1, pp. 1–29, 2002.

[133] K. Richards, C. Campenni and J. Muse-Burke, "Self-care and well-being in mental health professionals: The mediating effects of self-awareness and mindfulness," *Journal of mental health counseling*, vol. 32, no. 3, pp. 247–264, 2010.

[134] M. Zeidner, G. Matthews and R. D. Roberts, *What we know about emotional intelligence: How it affects learning, work, relationships, and our mental health*. MIT press, 2012.

[135] D. Kozlowski, M. Hutchinson, J. Hurley, J. Rowley and J. Sutherland, "The role of emotion in clinical decision making: An integrative literature review," *BMC medical education*, vol. 17, pp. 1–13, 2017.

[136] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis *et al.*, "Emotion recognition in human-computer interaction," *IEEE Signal processing magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[137] D. J. Siegel, *The developing mind: How relationships and the brain interact to shape who we are*. Guilford Publications, 2020.

[138] P. Ekman, "Are there basic emotions?," 1992.

[139] R. Plutchik, "What is an emotion?" *The Journal of psychology*, vol. 61, no. 2, pp. 295–303, 1965.

[140] H. Lövheim, "A new three-dimensional model for emotions and monoamine neurotransmitters," *Medical hypotheses*, vol. 78, no. 2, pp. 341–348, 2012.

[141] A. Cowen, D. Sauter, J. L. Tracy and D. Keltner, "Mapping the passions: Toward a high-dimensional taxonomy of emotional experience and expression," *Psychological Science in the Public Interest*, vol. 20, no. 1, pp. 69–90, 2019.

[142] S. Liang, "Enhancing multimodal emotional information extraction in film and television through adaptive feature fusion with densene, transformer, and 3d cnn models," *Applied Artificial Intelligence*, vol. 38, no. 1, p. 2 419 609, 2024.

[143] J. S. Rahman, T. Gedeon, S. Caldwell, R. Jones and Z. Jin, "Towards effective music therapy for mental health care using machine learning tools: Human affective reasoning and music genres," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 11, no. 1, pp. 5–20, 2021.

[144] S. Ismail, N. A. A. Aziz, S. Z. Ibrahim, C. T. Khan and M. A. Rahman, "Selecting video stimuli for emotion elicitation via online survey," *Human-Centric Computing and Information Sciences*, vol. 11, no. 36, pp. 1–18, 2021.

[145] R. Schleicher and J.-N. Antons, "Evoking emotions and evaluating emotional impact," in *Quality of Experience: Advanced Concepts, Applications and Methods*, Springer, 2014, pp. 121–132.

[146] F. Pan, L. Zhang, Y. Ou and X. Zhang, "The audio-visual integration effect on music emotion: Behavioral and physiological evidence," *PloS one*, vol. 14, no. 5, e0217040, 2019.

[147] M. K. Uhrig, N. Trautmann, U. Baumgärtner *et al.*, "Emotion elicitation: A comparison of pictures and films," *Frontiers in psychology*, vol. 7, p. 180, 2016.

[148] J. J. McGinley and B. H. Friedman, "Autonomic specificity in emotion: The induction method matters," *International Journal of Psychophysiology*, vol. 118, pp. 48–57, 2017.

[149] R. Westermann, K. Spies, G. Stahl and F. W. Hesse, "Relative effectiveness and validity of mood induction procedures: A meta-analysis," *European Journal of social psychology*, vol. 26, no. 4, pp. 557–580, 1996.

[150] Z. Romeo, F. Fusina, L. Semenzato, M. Bonato, A. Angrilli and C. Spironelli, "Comparison of slides and video clips as different methods for inducing emotions: An electroencephalographic alpha modulation study," *Frontiers in Human Neuroscience*, vol. 16, p. 901 422, 2022.

[151] M. Laroche, R. Li, M.-O. Richard and M. Zhou, "An investigation into online atmospherics: The effect of animated images on emotions, cognition, and purchase intentions," *Journal of Retailing and Consumer Services*, vol. 64, p. 102 845, 2022.

[152] E. E. Bartolini, "Eliciting emotion with film: Development of a stimulus set," 2011.

[153] D. Duan, W. Zhong, S. Ran, L. Ye and Q. Zhang, "A standardized database of chinese emotional short videos based on age and gender differences," *PloS One*, vol. 18, no. 3, e0283573, 2023.

[154] Y. Zhang, G. Zhao, Y. Shu *et al.*, "Cped: A chinese positive emotion database for emotion elicitation and analysis," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1417–1430, 2021.

[155] M. Martín and M. D. Valiña, "Heuristics, biases and the psychology of reasoning: State of the art," *Psychology*, vol. 14, no. 2, pp. 264–294, 2023.

[156] M. G. Haselton, D. Nettle and P. W. Andrews, "The evolution of cognitive bias," *The handbook of evolutionary psychology*, pp. 724–746, 2015.

[157] D. Watson, L. A. Clark and A. Tellegen, "Development and validation of brief measures of positive and negative affect: The panas scales.," *Journal of personality and social psychology*, vol. 54, no. 6, p. 1063, 1988.

[158] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[159] J. A. Russell and A. Mehrabian, "Distinguishing anger and anxiety in terms of emotional response factors.," *Journal of consulting and clinical psychology*, vol. 42, no. 1, p. 79, 1974.

[160] N. Garnefski and V. Kraaij, "The cognitive emotion regulation questionnaire," *European journal of psychological assessment*, vol. 23, no. 3, pp. 141–149, 2007.

[161] C. E. Izard, "Differential emotions theory," in *Human emotions*, Springer, 1977, pp. 43–66.

[162] K. A. Winter and N. A. Kuiper, "Individual differences in the experience of emotions," *Clinical psychology review*, vol. 17, no. 7, pp. 791–821, 1997.

[163] E. D. Klonsky, S. E. Victor, A. S. Hibbert and G. Hajcak, "The multidimensional emotion questionnaire (meq): Rationale and initial psychometric properties," *Journal of Psychopathology and Behavioral Assessment*, vol. 41, pp. 409–424, 2019.

[164] J. Kim, J. Wang, D. H. Wedell and S. V. Shinkareva, "Identifying core affect in individuals from fmri responses to dynamic naturalistic audiovisual stimuli," *PloS one*, vol. 11, no. 9, e0161589, 2016.

[165] W. Szwoch, "Emotion recognition using physiological signals," 15:1–15:8, 2015. DOI: 10.1145/2814464.2814479.

[166] W. Wei, Q. Jia, Y. Feng and G. Chen, "Emotion recognition based on weighted fusion strategy of multichannel physiological signals," *Computational Intelligence and Neuroscience*, vol. 2018, 2018. DOI: 10.1155/2018/5296523.

[167] D. Bierman and D. Radin, "Conscious and non-conscious emotional processes: A reversal of the arrow of time?,"

[168] C. Kauschke, D. Bahn, M. Vesker and G. Schwarzer, "The role of emotional valence for the processing of facial and verbal stimuli—positivity or negativity bias?" *Frontiers in psychology*, vol. 10, p. 1654, 2019.

[169] F. Galvão, S. M. Alarcão and M. J. Fonseca, "Predicting exact valence and arousal values from eeg," *Sensors*, vol. 21, no. 10, p. 3414, 2021.

[170] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE transactions on affective computing*, vol. 13, no. 3, pp. 1195–1215, 2020.

[171] A. A. Pise, M. A. Alqahtani, P. Verma, P. K, D. A. Karras and A. Halifa, "Methods for facial expression recognition with applications in challenging situations," *Computational intelligence and neuroscience*, vol. 2022, no. 1, p. 9 261 438, 2022.

[172] European Data Protection Supervisor (EDPS), "Facial emotion recognition," European Data Protection Supervisor (EDPS), Tech. Rep., May 2021, TechDispatch Issue, May 2021. DOI: 10.2804/519064. [Online]. Available: https://www.edps.europa.eu/system/files/2021-05/21-05-26_techdispatch-facial-emotion-recognition_ref_en.pdf.

[173] Aratek Biometrics. "How does facial emotion recognition express your feelings?" Accessed: 2025-02-04. (2025), [Online]. Available: https://www.aratek.co/news/how-does-facial-emotion-recognition-express-your-feelings.

[174] Softweb Solutions. "Benefits of facial recognition." Accessed: 2025-02-04. (2025), [Online]. Available: https://www.softwebsolutions.com/resources/benefits-of-facial-recognition.html.

[175] Y. Yang, B. Vuksanovic and H. Ma, "The performance analysis of facial expression recognition system using local regions and features," *Journal of Image and Graphics (United Kingdom)*, vol. 11, no. 2, pp. 104–114, 2023.

[176] G. Sandbach, S. Zafeiriou, M. Pantic and D. Rueckert, "A dynamic approach to the recognition of 3d facial expressions and their temporal models," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, IEEE, 2011, pp. 406–413.

[177] H. Fang, N. Mac Parthaláin, A. J. Aubrey *et al.*, "Facial expression recognition in dynamic sequences: An integrated approach," *Pattern Recognition*, vol. 47, no. 3, pp. 1271–1281, 2014.

[178] M. A. Rahim, M. N. Hossain, T. Wahid and M. S. Azam, "Face recognition using local binary patterns (lbp)," *Global Journal of Computer Science and Technology*, vol. 13, no. 4, pp. 1–8, 2013.

[179] D. Feng and F. Ren, "Dynamic facial expression recognition based on two-stream-cnn with lbp-top," in *2018 5th IEEE International conference on cloud computing and intelligence systems (CCIS)*, IEEE, 2018, pp. 355–359.

[180] J. Zhou and T. Wang, "Fer based on the improved convex nonnegative matrix factorization feature," *Multimedia Tools and Applications*, vol. 79, pp. 26 305–26 325, 2020.

[181] W. Xie, X. Jia, L. Shen and M. Yang, "Sparse deep feature learning for facial expression recognition," *Pattern Recognition*, vol. 96, p. 106 966, 2019.

[182] H. Arabian, V. Wagner-Hartl and K. Moeller, "Traditional versus neural network classification methods for facial emotion recognition," *Current Directions in Biomedical Engineering*, vol. 7, no. 2, pp. 203–206, 2021.

[183] Y. Khaireddin and Z. Chen, "Facial emotion recognition: State of the art performance on fer2013," *arXiv preprint arXiv:2105.03588*, 2021.

[184] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi and T. Gedeon, "Video and image based emotion recognition challenges in the wild: Emotiw 2015," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*, 2015, pp. 423–426.

[185] Papers with Code, *Facial expression recognition datasets*, Accessed: 2025-02-04, 2025. [Online]. Available: `https://paperswithcode.com/datasets?task=facial-expression-recognition&mod=images`.

[186] Z. U. Ningbo Innovation Center and Z. University, "Emotion recognition with multi-modal peripheral physiological signals," *Frontiers in Computer Science*, Dec. 2023.

[187] Z. Ahmad and N. Khan, "A survey on physiological signal-based emotion recognition," *Bioengineering*, vol. 9, no. 11, p. 688, 2022.

[188] M. D. Tooley, D. Carmel, A. Chapman and G. M. Grimshaw, "Dissociating the physiological components of unconscious emotional responses," *Neuroscience of Consciousness*, vol. 2017, no. 1, nix021, 2017.

[189] Z. Fu, B. Zhang, X. He, Y. Li, H. Wang and J. Huang, "Emotion recognition based on multi-modal physiological signals and transfer learning," *Frontiers in Neuroscience*, vol. 16, p. 1 000 716, 2022.

[190] E.-G. Han, T.-K. Kang and M.-T. Lim, "Physiological signal-based real-time emotion recognition based on exploiting mutual information with physiologically common features," *Electronics*, vol. 12, no. 13, p. 2933, 2023.

[191] S. Jerritta, M. Murugappan, R. Nagarajan and K. Wan, "Physiological signals based human emotion recognition: A review," *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*, pp. 410–415, 2011. DOI: 10.1109/CSPA.2011.5759912.

[192] L. Shu, J. Xie, M. Yang *et al.*, "A review of emotion recognition using physiological signals," *Sensors*, vol. 18, no. 7, p. 2074, 2018.

[193] F. Larradet, R. Niewiadomski, G. Barresi, D. G. Caldwell and L. S. Mattos, "Toward emotion recognition from physiological signals in the wild: Approaching the methodological issues in real-life data collection," *Frontiers in psychology*, vol. 11, p. 1111, 2020.

[194] D. R. Seshadri, R. T. Li, J. E. Voos *et al.*, "Wearable sensors for monitoring the internal and external workload of the athlete," *npj Digital Medicine*, vol. 2, no. 1, p. 71, Jul. 2019. DOI: 10.1038/s41746-019-0149-2. [Online]. Available: https://doi.org/10.1038/s41746-019-0149-2.

[195] O. Lowenstein and I. E. Loewenfeld, "Role of sympathetic and parasympathetic systems in reflex dilatation of the pupil: Pupillographic studies," *Archives of Neurology & Psychiatry*, vol. 64, no. 3, pp. 313–340, 1950.

[196] T. Partala and V. Surakka, "Pupil size variation as an indication of affective processing," *International Journal of Human-Computer Studies*, Applications of Affective Computing in Human-Computer Interaction, vol. 59, no. 1, pp. 185–198, 1st Jul. 2003, ISSN: 1071-5819. DOI: 10.1016/S1071-5819(03)00017-X. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S107158190300017X (visited on 06/05/2024).

[197] J. Z. Lim, J. Mountstephens and J. Teo, "Eye-tracking feature extraction for biometric machine learning," *Frontiers in neurorobotics*, vol. 15, p. 796 895, 2022.

[198] S. Sirois and J. Brisson, "Pupillometry," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 5, no. 6, pp. 679–692, 2014.

[199] C. Aracena, S. Basterrech, V. Snáel and J. Velásquez, "Neural networks for emotion recognition based on eye tracking data," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, IEEE, 2015, pp. 2632–2637.

[200] T. Partala and V. Surakka, "Pupil size variation as an indication of affective processing," *International journal of human-computer studies*, vol. 59, no. 1-2, pp. 185–198, 2003.

[201] F. Wu, Y. Zhao and H. Zhang, "Ocular autonomic nervous system: An update from anatomy to physiological functions," *Vision*, vol. 6, no. 1, p. 6, 2022.

[202] R. Henderson and J. Spies, "Autonomic nervous system disorders," in *Hankey's Clinical Neurology*, CRC Press, 2020, pp. 749–756.

[203] A. Babiker, I. Faye, K. Prehn and A. Malik, "Machine learning to differentiate between positive and negative emotions using pupil diameter," *Frontiers in psychology*, vol. 6, p. 1921, 2015.

[204] A. Unknown, "Investigating the relationship between background luminance and emotional arousal on pupillary responses," *Scandinavian Journal of Psychology*, 2022, Accessed: 2025-02-04. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0001691822000476.

[205] M. M. Bradley, L. Miccoli, M. A. Escrig and P. J. Lang, "The pupil as a measure of emotional arousal and autonomic activation," *Psychophysiology*, vol. 45, no. 4, pp. 602–607, 2008.

[206] N. Grujic, R. Polania and D. Burdakov, "Neurobehavioral meaning of pupil size," *Neuron*, vol. 112, no. 20, pp. 3381–3395, 2024.

[207] Y.-G. Cherng, T. Baird, J.-T. Chen and C.-A. Wang, "Background luminance effects on pupil size associated with emotion and saccade preparation," *Scientific reports*, vol. 10, no. 1, p. 15 718, 2020.

[208] F. for Young Minds, "Pupils: A window into the mind," *Frontiers for Young Minds*, 2019. [Online]. Available: https://kids.frontiersin.org/articles/10.3389/frym.2019.00003.

[209] J. Pan, X. Sun, E. Park *et al.*, "The effects of emotional arousal on pupil size depend on luminance," *Scientific Reports*, vol. 14, no. 1, p. 21 895, 2024.

[210] Y. Zhang, S. Li, J. Wang *et al.*, "Pupil size estimation based on spatially weighted corneal flux density," *IEEE Photonics Journal*, vol. 11, no. 6, pp. 1–9, Dec. 2019, Conference Name: IEEE Photonics Journal, ISSN: 1943-0655. DOI: 10.1109/JPHOT.2019.2948223. [Online]. Available: https://ieeexplore.ieee.org/document/8876637 (visited on 29/03/2024).

[211] A. E. Urai, A. Braun and T. H. Donner, "Pupil-linked arousal is driven by decision uncertainty and alters serial choice bias," *Nature communications*, vol. 8, no. 1, p. 14 637, 2017.

[212] R. A. McCleary, "The nature of the galvanic skin response.," *Psychological Bulletin*, vol. 47, no. 2, p. 97, 1950.

[213] J. J. Braithwaite, D. G. Watson, R. Jones and M. Rowe, "A guide for analysing electrodermal activity (eda) & skin conductance responses (scrs) for psychological experiments," *Psychophysiology*, vol. 49, no. 1, pp. 1017–1034, 2013.

[214] Z. M. Cohen, D. Carmichael, J. Winston and M. Richardson, "Non-invasive acute neuromodulation and measurement of epileptogenicity," 2025.

[215] A. K. E. Al-sabaawi, "A pilot study on the suitability of the galvanic skin response (gsr) as a measure of emotional state," 2015.

[216] G. I. Christopoulos, M. A. Uy and W. J. Yap, "The body and the brain: Measuring skin conductance responses to understand the emotional experience," *Organizational Research Methods*, vol. 22, no. 1, pp. 394–420, 2019.

[217] J. Gohumpu, M. Xue and Y. Bao, "Emotion recognition with multi-modal peripheral physiological signals," *Frontiers in Computer Science*, vol. 5, p. 1 264 713, 2023.

[218] S. Rinella, S. Massimino, P. G. Fallica *et al.*, "Emotion recognition: Photoplethysmography and electrocardiography in comparison," *Biosensors*, vol. 12, no. 10, p. 811, 2022.

[219] M. S. Lee, Y. K. Lee, D. S. Pae, M. T. Lim, D. W. Kim and T. K. Kang, "Fast emotion recognition based on single pulse ppg signal with convolutional neural network," *Applied Sciences*, vol. 9, no. 16, p. 3355, 2019.

[220] E. Mejía-Mejía, J. M. May, R. Torres and P. A. Kyriacou, "Pulse rate variability in cardiovascular health: A review on its applications and relationship with heart rate variability," *Physiological Measurement*, vol. 41, no. 7, 07TR01, 2020.

[221] J. J. Allen, H. L. Urry, S. K. Hitt and J. A. Coan, "The stability of resting frontal electroencephalographic asymmetry in depression," *Psychophysiology*, vol. 41, no. 2, pp. 269–280, 2004.

[222] R. J. Davidson, "What does the prefrontal cortex "do" in affect: Perspectives on frontal eeg asymmetry research," *Biological psychology*, vol. 67, no. 1-2, pp. 219–234, 2004.

[223] R. Vempati and L. D. Sharma, "A systematic review on automated human emotion recognition using electroencephalogram signals and artificial intelligence," *Results in Engineering*, vol. 18, p. 101 027, 2023.

[224] K. Keuper, P. Zwitserlood, M. A. Rehbein *et al.*, "Early prefrontal brain responses to the hedonic quality of emotional words–a simultaneous eeg and meg study," *PLoS One*, vol. 8, no. 8, e70788, 2013.

[225] S. Moratti, G. Rubio, P. Campo, A. Keil and T. Ortiz, "Hypofunction of right temporoparietal cortex during emotional arousal in depression," *Archives of general psychiatry*, vol. 65, no. 5, pp. 532–541, 2008.

[226] I. Daly, "Neural decoding of music from the eeg," *Scientific Reports*, vol. 13, no. 1, p. 624, 2023.

[227] P. A. Gable, D. L. Adams and G. H. Proudfit, "Transient tasks and enduring emotions: The impacts of affective content, task relevance, and picture duration on the sustained late positive potential," *Cognitive, Affective, & Behavioral Neuroscience*, vol. 15, no. 1, pp. 45–54, 2015.

[228] 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, 2019.

[229] E. M. Younis, S. Mohsen, E. H. Houssein and O. A. S. Ibrahim, "Machine learning for human emotion recognition: A comprehensive review," *Neural Computing and Applications*, pp. 1–47, 2024.

[230] T. Vu, V. T. Huynh and S.-H. Kim, "Multi-scale transformer-based network for emotion recognition from multi physiological signals," *arXiv preprint arXiv:2305.00769*, 2023.

[231] N. Gahlan and D. Sethia, "Federated learning in emotion recognition systems based on physiological signals for privacy preservation: A review," *Multimedia Tools and Applications*, pp. 1–69, 2024.

[232] T. Seehapoch and S. Wongthanavasu, "Speech emotion recognition using support vector machines," in *2013 5th international conference on Knowledge and smart technology (KST)*, IEEE, 2013, pp. 86–91.

[233] M. Li, H. Xu, X. Liu and S. Lu, "Emotion recognition from multichannel eeg signals using k-nearest neighbor classification," *Technology and health care*, vol. 26, no. S1, pp. 509–519, 2018.

[234] Y. Zheng, J. Ding, F. Liu and D. Wang, "Adaptive neural decision tree for eeg based emotion recognition," *Information Sciences*, vol. 643, p. 119 160, 2023.

[235] X. Zhang, C. Yan, C. Gao, B. A. Malin and Y. Chen, "Predicting missing values in medical data via xgboost regression," *Journal of healthcare informatics research*, vol. 4, pp. 383–394, 2020.

[236] S. Gharsalli, B. Emile, H. Laurent, X. Desquesnes and D. Vivet, "Random forest-based feature selection for emotion recognition," in *2015 International Conference on Image Processing Theory, Tools and Applications (IPTA)*, IEEE, 2015, pp. 268–272.

[237] S. Koelstra, C. Muhl, M. Soleymani *et al.*, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[238] N. Mehendale, "Facial emotion recognition using convolutional neural networks (ferc)," *SN Applied Sciences*, vol. 2, no. 3, p. 446, 2020.

[239] L. Chao, J. Tao, M. Yang, Y. Li and Z. Wen, "Long short term memory recurrent neural network based multimodal dimensional emotion recognition," in

Proceedings of the 5th international workshop on audio/visual emotion challenge, 2015, pp. 65–72.

[240] I. Shahin, A. B. Nassif and S. Hamsa, "Emotion recognition using hybrid gaussian mixture model and deep neural network," *IEEE access*, vol. 7, pp. 26 777–26 787, 2019.

[241] A. W. Awan, I. Taj, S. Khalid, S. M. Usman, A. S. Imran and M. U. Akram, "Advancing emotional health assessments: A hybrid deep learning approach using physiological signals for robust emotion recognition," *IEEE Access*, 2024.

[242] L. Li and J.-h. Chen, "Emotion recognition using physiological signals," in *International conference on artificial reality and telexistence*, Springer, 2006, pp. 437–446.

[243] S. Saganowski, B. Perz, A. G. Polak and P. Kazienko, "Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1876–1897, 2022.

[244] J. A. Domínguez-Jiménez, K. C. Campo-Landines, J. C. Martínez-Santos, E. J. Delahoz and S. H. Contreras-Ortiz, "A machine learning model for emotion recognition from physiological signals," *Biomedical Signal Processing and Control*, vol. 55, p. 101 646, 1st Jan. 2020, ISSN: 1746-8094. DOI: 10.1016/j.bspc.2019.101646. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809419302277 (visited on 17/04/2024).

[245] Q. Li, Y. Liu, F. Yan, Q. Zhang and C. Liu, "Emotion recognition based on multiple physiological signals," *Zhongguo yi liao qi xie za zhi = Chinese journal of medical instrumentation*, vol. 44 4, pp. 283–287, 2020. DOI: 10.3969/j.issn.1671-7104.2020.04.001.

[246] Z. He, Z. Li, F. Yang *et al.*, "Advances in multimodal emotion recognition based on brain–computer interfaces," *Brain sciences*, vol. 10, no. 10, p. 687, 2020.

[247] M. A. Ramadan, N. M. Salem, L. N. Mahmoud and I. Sadek, "Multimodal machine learning approach for emotion recognition using physiological signals," *Biomedical Signal Processing and Control*, vol. 96, p. 106 553, 2024.

[248] K. Ross, P. Hungler and A. Etemad, "Unsupervised multi-modal representation learning for affective computing with multi-corpus wearable data," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, pp. 1–26, Oct. 2021. DOI: 10.1007/s12652-021-03462-9.

[249] N. Sebe, I. Cohen and T. S. Huang, "Multimodal emotion recognition," in *Handbook of pattern recognition and computer vision*, World Scientific, 2005, pp. 387–409.

[250] M. Soleymani, M. Pantic and T. Pun, "Multimodal emotion recognition in response to videos," *IEEE transactions on affective computing*, vol. 3, no. 2, pp. 211–223, 2011.

[251] O. Ghita, D. E. Ilea and P. F. Whelan, "Adaptive noise removal approach for restoration of digital images corrupted by multimodal noise," *IET image processing*, vol. 6, no. 8, pp. 1148–1160, 2012.

[252] T. Zhang, A. El Ali, C. Wang, X. Zhu and P. Cesar, "Corrfeat: Correlation-based feature extraction algorithm using skin conductance and pupil diameter for emotion recognition," in *2019 International Conference on Multimodal Interaction*, 2019, pp. 404–408.

[253] W.-L. Zheng, B.-N. Dong and B.-L. Lu, "Multimodal emotion recognition using eeg and eye tracking data," in *2014 36th annual international conference of the IEEE engineering in medicine and biology society*, IEEE, 2014, pp. 5040–5043.

[254] P. Iacono and N. Khan, "Multi-modal emotion recognition using eeg and eye tracking features," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2024, pp. 1–5.

[255] A. F. Bulagang, N. G. Weng, J. Mountstephens and J. Teo, "A review of recent approaches for emotion classification using electrocardiography and electrodermography signals," *Informatics in Medicine Unlocked*, vol. 20, p. 100 363, 2020.

[256] T. Kumar, R. C. Singh and R. Kumar, "Emotions recognition based on physiological signals using machine learning techniques," *2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, pp. 823–827, 2023. DOI: 10.1109/ICTACS59847.2023.10390266.

[257] A. Alam, S. Urooj and A. Q. Ansari, "Human emotion recognition models using machine learning techniques," *2023 International Conference on Recent Advances in Electrical, Electronics & Digital Healthcare Technologies (REEDCON)*, pp. 329–334, 2023. DOI: 10.1109/REEDCON57544.2023.10151406.

[258] C.-A. Wang, T. Baird, J. Huang, J. D. Coutinho, D. C. Brien and D. P. Munoz, "Arousal effects on pupil size, heart rate, and skin conductance in an emotional face task," *Frontiers in neurology*, vol. 9, p. 1029, 2018.

[259] M. D. Hssayeni and B. Ghoraani, "Multi-modal physiological data fusion for affect estimation using deep learning," *IEEE Access*, vol. 9, pp. 21 642–21 652, 2021.

[260] G. Udovičić, J. Đerek, M. Russo and M. Sikora, "Wearable emotion recognition system based on gsr and ppg signals," in *Proceedings of the 2nd international workshop on multimedia for personal health and health care*, 2017, pp. 53–59.

[261] T. Thanapattheerakul, K. Mao, J. Amoranto and J. H. Chan, "Emotion in a century: A review of emotion recognition," in *proceedings of the 10th international conference on advances in information technology*, 2018, pp. 1–8.

[262] A. S. Cowen and D. Keltner, "What the face displays: Mapping 28 emotions conveyed by naturalistic expression.," *American Psychologist*, vol. 75, no. 3, p. 349, 2020.

[263] A. Mollahosseini, B. Hasani and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 18–31, 2017.

[264] A. S. Ayoub Al-Hamadi, R. Niese, S. Handrich and H. Neumann, "Emotional trace: Mapping of facial expression to valence-arousal space,"

[265] Y. Zhang, S. Li, J. Wang *et al.*, "Pupil size estimation based on spatially weighted corneal flux density," *IEEE Photonics Journal*, vol. 11, no. 6, pp. 1–9, 2019.

[266] B. Laeng and D. Alnaes, "Pupillometry," *Eye movement research: An introduction to its scientific foundations and applications*, pp. 449–502, 2019.

[267] A. S. Ansari, J. Vehof, C. J. Hammond, F. D. Bremner and K. M. Williams, "Evidence that pupil size and reactivity are determined more by your parents than by your environment," *Frontiers in Neurology*, vol. 12, p. 651 755, 2021.

[268] R. H. Spector, "The pupils," *Clinical Methods: The History, Physical, and Laboratory Examinations. 3rd edition*, 1990.

[269] A. Mathur, J. Gehrmann and D. A. Atchison, "Influences of luminance and accommodation stimuli on pupil size and pupil center location," *Investigative ophthalmology & visual science*, vol. 55, no. 4, pp. 2166–2172, 2014.

[270] A. M. Miranda, E. J. Nunes-Pereira, K. Baskaran and A. F. Macedo, "Eye movements, convergence distance and pupil-size when reading from smartphone, computer, print and tablet," *Scandinavian Journal of Optometry and Visual Science*, vol. 11, no. 1, pp. 1–5, 2018.

[271] P. Binda, M. Pereverzeva and S. O. Murray, "Pupil size reflects the focus of feature-based attention," *Journal of neurophysiology*, vol. 112, no. 12, pp. 3046–3052, 2014.

[272] A. S. DiCriscio and V. Troiani, "Pupil adaptation corresponds to quantitative measures of autism traits in children," *Scientific reports*, vol. 7, no. 1, p. 6476, 2017.

[273] L. Kay, R. Keogh, T. Andrillon and J. Pearson, "The pupillary light response as a physiological index of aphantasia, sensory and phenomenological imagery strength," *Elife*, vol. 11, e72484, 2022.

[274] M. Nakayama, I. Yasuike and Y. Shimizu, "Controlling the effects of brightness on the measurement of pupil size as a means of evaluating mental activity," *Pupil Reactions in Response to Human Mental Activity*, pp. 1–14, 2021.

[275] P. Tarnowski, M. Kołodziej, A. Majkowski, R. J. Rak *et al.*, "Eye-tracking analysis for emotion recognition," *Computational intelligence and neuroscience*, vol. 2020, 2020.

[276] P. Raiturkar, A. Kleinsmith, A. Keil, A. Banerjee and E. Jain, "Decoupling light reflex from pupillary dilation to measure emotional arousal in videos," in *Proceedings of the ACM Symposium on Applied Perception*, 2016, pp. 89–96.

[277] S. Asano, M. Nakayama and Y. Shimizu, "A neural-network-based eye pupil reaction model for use with television programs," *Pupil Reactions in Response to Human Mental Activity*, pp. 31–47, 2021.

[278] S. Asano, I. Yasuike, M. Nakayama and Y. Shimizu, "Pupil reaction model using a neural network for brightness change," *Pupil Reactions in Response to Human Mental Activity*, pp. 15–30, 2021.

[279] W. Boucsein, *Electrodermal activity*. Springer Science & Business Media, 2012.

[280] Y. Pei, Y. Ma, Q. Ran, T. Jamoulle, S. Lv and J. Liu, "The effect of band-pass filter range and global signal regression on the amplitude of low-frequency fluctuations: A pilot study," *Journal of Medical Imaging and Health Informatics*, vol. 9, no. 4, pp. 813–818, 2019.

[281] D. Das, T. Bhattacharjee, S. Datta, A. D. Choudhury, P. Das and A. Pal, "Classification and quantitative estimation of cognitive stress from in-game keystroke analysis using eeg and gsr," in *2017 IEEE Life Sciences Conference (LSC)*, IEEE, 2017, pp. 286–291.

[282] G. Chanel, J. Kronegg, D. Grandjean and T. Pun, "Emotion assessment: Arousal evaluation using eeg's and peripheral physiological signals," in *International workshop on multimedia content representation, classification and security*, Springer, 2006, pp. 530–537.

[283] T. N. Nguyen, "An algorithm for extracting the ppg baseline drift in real-time," 2016.

[284] D. Han, S. K. Bashar, J. Lázaro *et al.*, "A real-time ppg peak detection method for accurate determination of heart rate during sinus rhythm and cardiac arrhythmia," *Biosensors*, vol. 12, no. 2, p. 82, 2022.

[285] M. N. Dar, M. U. Akram, S. G. Khawaja and A. N. Pujari, "Cnn and lstm-based emotion charting using physiological signals," *Sensors*, vol. 20, no. 16, p. 4551, 2020.

[286] M. Shantal, Z. Othman and A. A. Bakar, "A novel approach for data feature weighting using correlation coefficients and min–max normalization," *Symmetry*, vol. 15, no. 12, p. 2185, 2023.

[287] A. Greco, G. Valenza and E. P. Scilingo, *Advances in electrodermal activity processing with applications for mental health*. Springer, 2016.

[288] M. R. Ram, K. V. Madhav, E. H. Krishna, N. R. Komalla, K. Sivani and K. A. Reddy, "Ica-based improved dtcwt technique for ma reduction in ppg signals with restored respiratory information," *IEEE Transactions on Instrumentation and Measurement*, vol. 62, no. 10, pp. 2639–2651, 2013.

[289] M. A. Motin, C. K. Karmakar and M. Palaniswami, "Selection of empirical mode decomposition techniques for extracting breathing rate from ppg," *IEEE Signal Processing Letters*, vol. 26, no. 4, pp. 592–596, 2019.

[290] G. Nie, J. Zhu, G. Tang *et al.*, "A review of deep learning methods for photoplethysmography data," *arXiv preprint arXiv:2401.12783*, 2024.

[291] J. Z. Lim, J. Mountstephens and J. Teo, "Emotion recognition using eye-tracking: Taxonomy, review and current challenges," *Sensors*, vol. 20, no. 8, p. 2384, 2020.

[292] C.-L. Lee, W. Pei, Y.-C. Lin, A. Granmo and K.-H. Liu, "Emotion detection based on pupil variation," in *Healthcare*, MDPI, vol. 11, 2023, p. 322.

[293] W. Celniak and P. Augustyniak, "Eye-tracking as a component of multimodal emotion recognition systems," in *International Conference on Information Technologies in Biomedicine*, Springer, 2022, pp. 66–75.

[294] T. Božak, M. Luštrek and G. Slapničar, "Feature-based emotion classification using eye-tracking data," in *Information Society 2024*, Jožef Stefan Institute, 2024, pp. 7–11.

[295] S. Madhavan, "Emotions and pupil dilation: Understanding the psychological connection,"

[296] P. Arias Sarah, L. Hall, A. Saitovitch, J.-J. Aucouturier, M. Zilbovicius and P. Johansson, "Pupil dilation reflects the dynamic integration of audiovisual emotional speech," *Scientific reports*, vol. 13, no. 1, p. 5507, 2023.

[297] M. Oliva and A. Anikin, "Pupil dilation reflects the time course of emotion recognition in human vocalizations," *Scientific reports*, vol. 8, no. 1, p. 4871, 2018.

[298] A. Maza, B. Moliner, J. Ferri and R. Llorens, "Visual behavior, pupil dilation, and ability to identify emotions from facial expressions after stroke," *Frontiers in neurology*, vol. 10, p. 1415, 2020.

[299] L. Moharana and N. Das, "Analysis of pupil dilation on different emotional states by using computer vision algorithms," in *2021 1st Odisha International Conference on Electrical Power Engineering, Communication and Computing Technology (ODICON)*, IEEE, 2021, pp. 1–6.

[300] A. Gautam, N. Simões-Capela, G. Schiavone, A. Acharyya, W. De Raedt and C. Van Hoof, "A data driven empirical iterative algorithm for gsr signal preprocessing," in *2018 26th European Signal Processing Conference (EUSIPCO)*, IEEE, 2018, pp. 1162–1166.

[301] D. Ayata, Y. Yaslan and M. Kamaşak, "Emotion recognition via galvanic skin response: Comparison of machine learning algorithms and feature extraction methods," *IU-Journal of Electrical & Electronics Engineering*, vol. 17, no. 1, pp. 3147–3156, 2017.

[302] S. Chen, K. Jiang, H. Hu *et al.*, "Emotion recognition based on skin potential signals with a portable wireless device," *Sensors*, vol. 21, no. 3, p. 1018, 2021.

[303] E. Lutin, R. Hashimoto, W. De Raedt and C. Van Hoof, "Feature extraction for stress detection in electrodermal activity.," in *BIOSIGNALS*, Vienna, Austria, 2021, pp. 177–185.

[304] K. NISA'MINHAD, S. H. M. Ali and M. B. I. Reaz, "A design framework for human emotion recognition using electrocardiogram and skin conductance response signals," *J. Eng. Sci. Technol*, vol. 12, no. 11, pp. 3102–3119, 2017.

[305] J. Shukla, M. Barreda-Angeles, J. Oliver, G. C. Nandi and D. Puig, "Feature extraction and selection for emotion recognition from electrodermal activity," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 857–869, 2019.

[306] K. Nisa'Minhad, S. H. M. Ali, J. O. S. Khai and S. A. Ahmad, "Human emotion classifications for automotive driver using skin conductance response signal," in *2016 International Conference on Advances in Electrical, Electronic and Systems Engineering (ICAEES)*, IEEE, 2016, pp. 371–375.

[307] J. Schumm, M. Bächlin, C. Setz, B. Arnrich, D. Roggen and G. Tröster, "Effect of movements on the electrodermal response after a startle event," *Methods of Information in Medicine*, vol. 47, no. 03, pp. 186–191, 2008.

[308] J. Choi, B. Ahmed and R. Gutierrez-Osuna, "Development and evaluation of an ambulatory stress monitor based on wearable sensors," *IEEE transactions on information technology in biomedicine*, vol. 16, no. 2, pp. 279–286, 2011.

[309] S. A. H. Aqajari, E. K. Naeini, M. A. Mehrabadi, S. Labbaf, A. M. Rahmani and N. Dutt, "Gsr analysis for stress: Development and validation of an open source tool for noisy naturalistic gsr data," *arXiv preprint arXiv:2005.01834*, 2020.

[310] P. S. Kumar, P. K. Govarthan, A. A. S. Gadda, N. Ganapathy and J. F. A. Ronickom, "Deep learning-based automated emotion recognition using multi modal physiological signals and time-frequency methods," *IEEE Transactions on Instrumentation and Measurement*, 2024.

[311] F. Panahi, S. Rashidi and A. Sheikhani, "Application of fractional fourier transform in feature extraction from electrocardiogram and galvanic skin response for emotion recognition," *Biomedical Signal Processing and Control*, vol. 69, p. 102 863, 2021.

[312] Unknown, "New robotic technology detects human emotions through skin conductance measurements," *IEEE Access*, Dec. 2024, Accessed: 2024-12-25. [Online]. Available: `https://idtechwire.com/new-robotic-technology-detects-human-emotions-through-skin-conductance-measurements/`.

[313] A. Goshvarpour and A. Goshvarpour, "Poincaré's section analysis for ppg-based automatic emotion recognition," *Chaos, Solitons & Fractals*, vol. 114, pp. 400–407, 2018.

[314] L. Shu, Y. Yu, W. Chen *et al.*, "Wearable emotion recognition using heart rate data from a smart bracelet," *Sensors*, vol. 20, no. 3, p. 718, 2020.

[315] M. Mohammadpoor Faskhodi, M. Fernández-Chimeno and M. A. García-González, "Arousal detection by using ultra-short-term heart rate variability (hrv) analysis," *Frontiers in Medical Engineering*, vol. 1, p. 1 209 252, 2023.

[316] D. Dimitriev, O. Indeykina and A. Dimitriev, "The effect of auditory stimulation on the nonlinear dynamics of heart rate: The impact of emotional valence and arousal," *Noise and Health*, vol. 25, no. 118, pp. 165–175, 2023.

[317] J. Miles and M. Shevlin, "Applying regression and correlation: A guide for students and researchers," 2000.

[318] K. Rawal and A. Ahmad, "Feature selection for electrical demand forecasting and analysis of pearson coefficient," in *2021 IEEE 4th International Electrical and Energy Conference (CIEEC)*, IEEE, 2021, pp. 1–6.

[319] M. Awad and S. Fraihat, "Recursive feature elimination with cross-validation with decision tree: Feature selection method for machine learning-based intrusion detection systems," *Journal of Sensor and Actuator Networks*, vol. 12, no. 5, p. 67, 2023.

[320] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, no. 1, pp. 267–288, 1996.

[321] A. M. E. Saleh, M. Arashi, R. A. Saleh and M. Norouzirad, *Rank-based methods for shrinkage and selection: With application to machine learning*. John Wiley & Sons, 2022.

[322] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Frontiers in Bioinformatics*, vol. 2, p. 927 312, 2022.

[323] M. A. M. Hasan, M. Nasser, S. Ahmad and K. I. Molla, "Feature selection for intrusion detection using random forest," *Journal of information security*, vol. 7, no. 3, pp. 129–140, 2016.

[324] W. E. Marcílio and D. M. Eler, "From explanations to feature selection: Assessing shap values as feature selection mechanism," in *2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI)*, Ieee, 2020, pp. 340–347.

[325] Z. Atashgahi, G. Sokar, T. Van Der Lee *et al.*, "Quick and robust feature selection: The strength of energy-efficient sparse training for autoencoders," *Machine Learning*, pp. 1–38, 2022.

[326] L. S. K. Alnaber, "Comparative analysis of prediction models for diabetic patient readmission using explainable ai for feature selection and two-stage optimization techniques," M.S. thesis, State University of New York at Binghamton, 2024.

[327] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[328] M. Kuhn, K. Johnson *et al.*, *Applied predictive modeling*. Springer, 2013, vol. 26.

[329] L. Mazza, *Coarse-graining the cross-section: How regression-via-classification improves robustness in high-noise, small-sample-size domains such as cross-sectional asset pricing*, 2024.

[330] Ö. Çoban, M. Esit, S. Yalçın and F. Bozkurt, "Better with fewer features: Climate dynamics estimation for van lake basin using feature selection," *Environmental Science and Pollution Research*, pp. 1–25, 2025.

[331] S. Katsigiannis and N. Ramzan, "Dreamer: A database for emotion recognition through eeg and ecg signals from wireless low-cost off-the-shelf devices," *IEEE journal of biomedical and health informatics*, vol. 22, no. 1, pp. 98–107, 2017.

[332] 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, 2018.

[333] B. Nojavanasghari, T. Baltrušaitis, C. E. Hughes and L.-P. Morency, "Emoreact: A multimodal approach and dataset for recognizing emotional responses in children," in *Proceedings of the 18th acm international conference on multimodal interaction*, 2016, pp. 137–144.

[334] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (savee) database," *University of Surrey: Guildford, UK*, 2014.

[335] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.

[336] P. Schmidt, A. Reiss, R. Dürichen and K. Van Laerhoven, "Wearable-based affect recognition—a review," *Sensors*, vol. 19, no. 19, p. 4079, 2019.

[337] F. Abri, L. F. Gutiérrez, A. S. Namin, D. R. Sears and K. S. Jones, "Predicting emotions perceived from sounds," in *2020 IEEE International Conference on Big Data (Big Data)*, IEEE, 2020, pp. 2057–2064.

[338] Z. Lian, L. Sun, M. Xu *et al.*, "Explainable multimodal emotion reasoning," *CoRR*, 2023.

[339] I. O. Joudeh, A.-M. Cretu, S. Bouchard and S. Guimond, "Prediction of continuous emotional measures through physiological and visual data," *Sensors*, vol. 23, no. 12, p. 5613, 2023.

[340] R. Elliott, K. Lythe, R. Lee *et al.*, "Reduced medial prefrontal responses to social interaction images in remitted depression," *Archives of General Psychiatry*, vol. 69, no. 1, pp. 37–45, 2012.

[341] R. Elliott, R. Zahn, J. Deakin and I. M. Anderson, "Affective cognition and its disruption in mood disorders," *Neuropsychopharmacology*, vol. 36, no. 1, pp. 153–182, 2011.

[342] K. J. Ressler and H. S. Mayberg, "Targeting abnormal neural circuits in mood and anxiety disorders: From the laboratory to the clinic," *Nature neuroscience*, vol. 10, no. 9, pp. 1116–1124, 2007.

[343] S. Baron-Cohen, S. Wheelwright, R. Skinner, J. Martin and E. Clubley, "The autism-spectrum quotient (aq): Evidence from asperger syndrome/high-functioning autism, malesand females, scientists and mathematicians," *Journal of autism and developmental disorders*, vol. 31, pp. 5–17, 2001.

[344] A. Mühlberger, M. J. Wieser, M. J. Herrmann, P. Weyers, C. Tröger and P. Pauli, "Early cortical processing of natural and artificial emotional faces differs between lower and higher socially anxious persons," *Journal of neural transmission*, vol. 116, pp. 735–746, 2009.

[345] J. Hogeveen, G. Bird, A. Chau, F. Krueger and J. Grafman, "Acquired alexithymia following damage to the anterior insula," *Neuropsychologia*, vol. 82, pp. 142–148, 2016.

[346] E. Finch, A. Copley, P. Cornwell and C. Kelly, "Systematic review of behavioral interventions targeting social communication difficulties after traumatic brain injury," *Archives of Physical Medicine and Rehabilitation*, vol. 97, no. 8, pp. 1352–1365, 2016.

[347] R. M. Bagby, J. D. Parker and G. J. Taylor, "The twenty-item toronto alexithymia scale—i. item selection and cross-validation of the factor structure," *Journal of psychosomatic research*, vol. 38, no. 1, pp. 23–32, 1994.

[348] K. Kroenke, R. L. Spitzer and J. B. Williams, "The phq-9: Validity of a brief depression severity measure," *Journal of general internal medicine*, vol. 16, no. 9, pp. 606–613, 2001.

[349] R. L. Spitzer, K. Kroenke, J. B. Williams and B. Löwe, "A brief measure for assessing generalized anxiety disorder: The gad-7," *Archives of internal medicine*, vol. 166, no. 10, pp. 1092–1097, 2006.

[350] A. W. Loranger, N. Sartorius, A. Andreoli *et al.*, "The international personality disorder examination: The world health organization/alcohol, drug abuse, and mental health administration international pilot study of personality disorders," *Archives of general psychiatry*, vol. 51, no. 3, pp. 215–224, 1994.

[351] C. Bourke, K. Douglas and R. Porter, "Processing of facial emotion expression in major depression: A review," *Australian & New Zealand Journal of Psychiatry*, vol. 44, no. 8, pp. 681–696, 2010.

[352] N. Rickard, H.-A. Arjmand, D. Bakker, E. Seabrook *et al.*, "Development of a mobile phone app to support self-monitoring of emotional well-being: A mental health digital innovation," *JMIR mental health*, vol. 3, no. 4, e6202, 2016.

[353] S. Harrer, P. Shah, B. Antony and J. Hu, "Artificial intelligence for clinical trial design," *Trends in pharmacological sciences*, vol. 40, no. 8, pp. 577–591, 2019.

[354] Y. J. Wang and M. S. Minor, "Identifying psychiatric manifestations in schizophrenia and depression using machine learning," *Nature Mental Health*, 2022, Accessed: February 18, 2025. [Online]. Available: https://www.nature.com/articles/s41537-022-00287-z.

[355] M. d'Aquin *et al.*, "Introducing calmed: Multimodal annotated dataset for emotion detection in children with autism," *arXiv preprint arXiv:2307.13706*, 2023. [Online]. Available: https://arxiv.org/abs/2307.13706.

[356] N. Farhoumandi, S. Mollaey, S. Heysieattalab, M. Zarean and R. Eyvazpour, "Facial emotion recognition predicts alexithymia using machine learning," *Psychology & Marketing*, 2021, Accessed: February 18, 2025. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1155/2021/2053795.

[357] Tenovi, "Remote patient monitoring for mental health care," 2024, Accessed: February 18, 2025. [Online]. Available: https://www.tenovi.com/remote-patient-monitoring-for-mental-health/.

[358] 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, ISSN: 1939-1315. DOI: 10.1037/h0077714.

[359] J. O'Dwyer, R. Flynn and N. Murray, "Continuous affect prediction using eye gaze and speech," in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2017, pp. 2001–2007.

[360] J. Raju, Y. F. A. Gaus and T. P. Breckon, "Continuous multi-modal emotion prediction in video based on recurrent neural network variants with attention," in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, 2021, pp. 688–693.

[361] K. Brady, Y. Gwon, P. Khorrami *et al.*, "Multi-modal audio, video and physiological sensor learning for continuous emotion prediction," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 97–104.

[362] S. Patania, A. D'Amelio and R. Lanzarotti, "Exploring fusion strategies in deep multimodal affect prediction," in *International Conference on Image Analysis and Processing*, Springer, 2022, pp. 730–741.

[363] A. Dubbaka and A. Gopalan, "Detecting learner engagement in moocs using automatic facial expression recognition," in *2020 IEEE Global Engineering Education Conference (EDUCON)*, IEEE, 2020, pp. 447–456.

[364] W. Zhang, F. Qiu, C. Liu *et al.*, "An effective ensemble learning framework for affective behaviour analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4761–4772.

[365] A. Goshvarpour, A. Abbasi and A. Goshvarpour, "An accurate emotion recognition system using ecg and gsr signals and matching pursuit method," *Biomedical journal*, vol. 40, no. 6, pp. 355–368, 2017.

[366] W. Li, Z. Zhang and A. Song, "Physiological-signal-based emotion recognition: An odyssey from methodology to philosophy," *Measurement*, vol. 172, p. 108 747, 2021.

[367] C. Fawcett, E. Nordenswan, S. Yrttiaho *et al.*, "Individual differences in pupil dilation to others' emotional and neutral eyes with varying pupil sizes," *Cognition and Emotion*, vol. 36, no. 5, pp. 928–942, 2022.

[368] G. Cosme, P. J. Rosa, C. F. Lima *et al.*, "Pupil dilation reflects the authenticity of received nonverbal vocalizations," *Scientific Reports*, vol. 11, no. 1, p. 3733, 2021.

[369] A. Greco, "A new processing approach and modeling for the analysis of the electrodermal activity during multi-sensory affective stimulation," 2016.

[370] A. Al-Nafjan and M. Aldayel, "Anxiety detection system based on galvanic skin response signals," *Applied Sciences*, vol. 14, no. 23, p. 10 788, 2024.

[371] M. A. Nicolaou, H. Gunes and M. Pantic, "Output-associative rvm regression for dimensional and continuous emotion prediction," *Image and Vision Computing*, vol. 30, no. 3, pp. 186–196, 2012.

[372] M. Wegge, "Investigating topic bias in emotion classification," M.S. thesis, 2023.

[373] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions (IJSE)*, vol. 1, no. 1, pp. 68–99, 2010.

[374] S. Zamani, R. Sinha, M. Nguyen and S. Madanian, "Enhancing emotional well-being with iot data solutions for depression: A systematic review," *IEEE Journal of Biomedical and Health Informatics*, 2025.

[375] Y.-H. Yang, Y.-C. Lin, Y.-F. Su and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 2, pp. 448–457, 2008.

[376] B. Zupan and M. Eskritt, "Eliciting emotion ratings for a set of film clips: A preliminary archive for research in emotion," *The Journal of Social Psychology*, vol. 160, no. 6, pp. 768–789, 2020.

[377] K. Diconne, G. K. Kountouriotis, A. E. Paltoglou, A. Parker and T. J. Hostler, "Presenting kapodi–the searchable database of emotional stimuli sets," *Emotion Review*, vol. 14, no. 1, pp. 84–95, 2022.

[378] R. Christopherson, *Next-gen eye tracking is a game changer for disabled people | abilitynet*. [Online]. Available: https://abilitynet.org.uk/news-blogs/next-gen-eye-tracking-game-changer-disabled-people.

[379] Affectiva, *Imotions behavioral research with affectiva integration*, Accessed: 2025-02-04, 2025. [Online]. Available: https://www.affectiva.com/product/imotions-behavioral-research/.

[380] 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*, 2018, pp. 400–408.

[381] M. P. A. Ramaswamy and S. Palaniswamy, "Multimodal emotion recognition: A comprehensive review, trends, and challenges," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 14, no. 6, e1563, 2024.

[382] M. Magdin, L. Benko and Š. Koprda, "A case study of facial emotion classification using affdex," *Sensors*, vol. 19, no. 9, p. 2140, 2019.

[383] R. Salim, *What is eda peak detection and how does it work?* [Online]. Available: https://imotions.com/blog/learning/research-fundamentals/eda-peak-detection/.

[384] A. Greco, A. Lanata, G. Valenza, E. P. Scilingo and L. Citi, "Electrodermal activity processing: A convex optimization approach," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2014, pp. 2290–2293.

[385] B. Farnsworth, *Skin conductance response – what it is and how to measure it*. [Online]. Available: https://imotions.com/blog/learning/best-practice/skin-conductance-response/.

[386] A. BROCANELLI, "Galvanic skin response measurement data processing for user-related information extraction," 2018.

[387] D. Lulé, U. M. Schulze, K. Bauer *et al.*, "Anorexia nervosa and its relation to depression, anxiety, alexithymia and emotional processing deficits," *Eating and Weight Disorders-Studies on Anorexia, Bulimia and Obesity*, vol. 19, pp. 209–216, 2014.

[388] M. L. Phillips, W. C. Drevets, S. L. Rauch and R. Lane, "Neurobiology of emotion perception ii: Implications for major psychiatric disorders," *Biological psychiatry*, vol. 54, no. 5, pp. 515–528, 2003.

[389] M. D. Munir, "Prediction of heteroscedastic data using linear regression and various machine learning models," *Int. J. Sci. Res. in Mathematical and Statistical Sciences Vol*, vol. 10, no. 1, p. 20, 2023.

[390] T. Leinonen, D. Wong, A. Wahab, R. Nadarajah, M. Kaisti and A. Airola, "Empirical investigation of multi-source cross-validation in clinical machine learning," *arXiv preprint arXiv:2403.15012*, 2024.

[391] iMotions, *Galvanic skin response (gsr): The complete guide*, Accessed: March 20, 2025, 2023. [Online]. Available: https://imotions.com/blog/learning/research-fundamentals/galvanic-skin-response/.

[392] P. F. Grim and S. H. White, "Effects of stimulus change upon the gsr and reaction time.," *Journal of Experimental Psychology*, vol. 69, no. 3, p. 276, 1965.

[393] N. I. Technology, *What is galvanic skin response? a beginner's guide*, Accessed: March 20, 2025, Feb. 2025. [Online]. Available: https://noldus.com/blog/what-is-galvanic-skin-response.

[394] T. Dang, V. Sethu and E. Ambikairajah, "Factor analysis based speaker normalisation for continuous emotion prediction.," in *INTERSPEECH*, 2016, pp. 913–917.

[395] iMotions, *Pupillometry 101: How pupil responses reveal brain activity*, Accessed: February 12, 2025. [Online]. Available: https://imotions.com/blog/learning/best-practice/pupillometry-101/#eye-tracking.

[396] A. Jonnalagadda, M. Rajvir, S. Singh, S. Chandramouliswaran, J. George and F. Kamalov, "An ensemble-based machine learning model for emotion and mental health detection," *Journal of Information & Knowledge Management*, vol. 22, no. 02, p. 2 250 075, 2023.

[397] Y. J. Wang and M. S. Minor, "Validity, reliability, and applicability of psychophysiological techniques in marketing research," *Psychology & Marketing*, vol. 25, no. 2, pp. 197–232, 2008. DOI: 10.1002/mar.20206. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/mar.20206.

[398] W. Wen, G. Liu, N. Cheng, J. Wei, P. Shangguan and W. Huang, "Emotion recognition based on multi-variant correlation of physiological signals," *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 126–140, 2014.

[399] J. Khamthung, N. Lohia and S. Srivastava, "Multi-class emotion classification with xgboost model using wearable eeg headband data," *SMU Data Science Review*, vol. 8, no. 1, p. 7,

[400] K. M. Sakib, D. M. A. A. Salsabil Nishat Alam Md Golam Rabiul and U. M. Zia, "Cnn-xgboost fusion-based affective state recognition using eeg spectrogram image analysis," *Scientific Reports*, vol. 12, 2022. DOI: https://doi.org/10.1038/s41598-022-18257-x.

[401] Rishu, J. Singh and R. Gill, "Multimodal emotion recognition system using machine learning and psychological signals: A review," *Soft Computing: Theories and Applications: Proceedings of SoCTA 2020, Volume 1*, pp. 657–666, 2022.

[402] A. P. Muniyandi, K. Padmanandam, K. Subbaraj *et al.*, "An intelligent emotion prediction system using improved sand cat optimization technique based on eeg signals," *Scientific Reports*, vol. 15, no. 1, p. 8782, 2025.

[403] J. Kelly, P. Gooding, D. Pratt, J. Ainsworth, M. Welford and N. Tarrier, "Intelligent real-time therapy: Harnessing the power of machine learning to optimise the delivery of momentary cognitive–behavioural interventions," *Journal of Mental Health*, vol. 21, no. 4, pp. 404–414, 2012.

[404] I. Amaro, A. A. Citarella, F. De Marco *et al.*, "Ai-driven technologies in digital health & well being: Early detection and intervention strategies," in *CEUR WORKSHOP PROCEEDINGS*, CEUR-WS, vol. 3762, 2024, pp. 330–335.

[405] M. Benedek and C. Kaernbach, "A continuous measure of phasic electrodermal activity," *Journal of neuroscience methods*, vol. 190, no. 1, pp. 80–91, 2010.

[406] A. Greco, G. Valenza, E. P. Scilingo, A. Greco, G. Valenza and E. P. Scilingo, "Evaluation of cda and cvxeda models," *Advances in Electrodermal Activity Processing with Applications for Mental Health: From Heuristic Methods to Convex Optimization*, pp. 35–43, 2016.

[407] D. Chatterjee, R. Gavas and S. K. Saha, "Exploring skin conductance features for cross-subject emotion recognition," in *2022 IEEE Region 10 Symposium (TENSYMP)*, ISSN: 2642-6102, Jul. 2022, pp. 1–6. DOI: 10.1109/TENSYMP54529.2022.9864492. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9864492 (visited on 17/04/2024).

[408] O. Dehzangi, V. Sahu, V. Rajendra and M. Taherisadr, "Gsr-based distracted driving identification using discrete & continuous decomposition and wavelet packet transform," *Smart Health*, vol. 14, p. 100 085, 2019.

[409] P. Siirtola, S. Tamminen, G. Chandra, A. Ihalapathirana and J. Röning, "Predicting emotion with biosignals: A comparison of classification and regression models for estimating valence and arousal level using wearable sensors," *Sensors*, vol. 23, no. 3, p. 1598, 2023.