INTERNATIONAL HELLENIC UNIVERSITY

DEPARTMENT OF INFORMATION AND ELECTRONIC ENGINEERING

DIPLOMA THESIS

«Multimodal Fusion for Emotion Recognition: Leveraging Physiological Signals from Wearable Sensors»

Student
Ilias Chatzis
Student ID: 185308

Supervisor
Konstantinos Goulianas
Professor

Date 01-08-2025

Thesis Title. Multimodal Fusion for Emotion Recognition: Leveraging Physiological Signals from Wearable Sensors

Thesis ID. 25161

Student's full name. Ilias Chatzis

Supervisor's full name. Konstantinos Goulianas

Starting date. 10-03-2025

Completion date. 01-08-2025

I certify that I am the author of this thesis and that any assistance I had in its preparation is fully acknowledged and referenced in the thesis. I have also recorded any sources from which I have made use of data, ideas, images and text, whether quoted verbatim or paraphrased. Furthermore, I certify that this thesis was prepared by me personally, specifically as a thesis, at the Department of Information and Electronic Engineering of IHU.

The approval of the thesis by the Department of Information and Electronic Engineering of the International Hellenic University does not necessarily imply acceptance of the views of the author, on behalf of the Department.

«To those who Love and Help»

Prologue

Throughout the course of their studies, the author encountered a wealth of knowledge across diverse yet interconnected domains. However, it was the discovery of Machine Learning that truly ignited their intellectual curiosity and motivated a deeper commitment to learning. From that moment, it became clear that this was the field the author wished to pursue. Among the many captivating aspects of Machine Learning, the concept of multimodal fusion stood out, both in its technical complexity and its potential for meaningful real-world application. The feasibility of emotion recognition through wearable sensors, coupled with the broader vision of developing systems that could assist individuals in managing anxiety or support healthcare environments, became the central inspiration for this thesis. If this work contributes even in a small way to that vision, by advancing understanding or sparking further research, the author considers it a success. This thesis will remain a personal and academic milestone, continually reminding the author that the ultimate purpose of scientific advancement is to serve society and improve human well-being.

Abstract (EN)

The current thesis explores machine learning methods for emotion recognition through the multimodal fusion of wearable biometric sensor data. It leverages a dataset comprising physiological signals coupled with emotional tags, collected using an EmotiBit wearable device equipped with three-wavelength photoplethysmograph, galvanic skin response, and body temperature sensors. The research investigates how integrating heterogeneous sensor modalities can enhance the detection and monitoring of emotional states.

In the initial phase of the thesis, a comprehensive literature review will be conducted to understand the unique characteristics of each distinct physiological signal type, as well as to evaluate techniques for their processing and extracting pertinent features. The insights gathered also inform the fusion strategies to be used.

The study systematically tackles challenges in data preprocessing, feature extraction and selection, model development, and both feature-level and decision-level fusion. Through rigorous evaluation, it assesses the robustness and accuracy of various multimodal fusion strategies and classification algorithms applied to the emotion recognition task.

Ultimately, this work aims to demonstrate that the strategic integration of wearable biometric sensor data can significantly improve the performance of emotion recognition systems. In these frameworks, fusion is used to exploit the distinct characteristics of each sensor, showcasing their unique advantages while mitigating individual weaknesses. It is anticipated that the insights derived from this research will contribute to the development of reliable affective technologies -with applications in healthcare, stress management, and diagnostic support for emotional well-being- all achieved using only readily available wearable devices.

Abstract (EL)

Η παρούσα Διπλωματική Εργασία διερευνά μεθόδους μηχανικής μάθησης για την αναγνώριση συναισθημάτων μέσω της σύντηξης πολυτροπικών δεδομένων από φορετούς βιομετρικούς αισθητήρες. Αξιοποιεί μία βάση δεδομένων αποτελούμενη από σήματα φυσιολογίας σε συνδυασμό με ετικέτες συναισθημάτων, τα οποία συλλέχθηκαν από μία φορετή συσκευή EmotiBit εξοπλισμένη με φωτοπληθυσμογράφο τριών μήκων κύματος, αισθητήρα γαλβανικής απόκρισης δέρματος και αισθητήρα θερμοκρασίας σώματος. Αυτή η έρευνα εξετάζει πώς η ενσωμάτωση ετερογενών τύπων αισθητήρων μπορεί να βελτιώσει την ανίχνευση και την παρακολούθηση συναισθηματικών καταστάσεων.

Στην αρχική φάση της εργασίας, θα πραγματοποιηθεί μια εκτενής βιβλιογραφική έρευνα με σκοπό την κατανόηση των μοναδικών χαρακτηριστικών κάθε ξεχωριστού τύπου φυσιολογικού σήματος, καθώς και για την αξιολόγηση τεχνικών για την επεξεργασία τους και την εξαγωγή των σχετικών χαρακτηριστικών τους. Οι γνώσεις που συλλέγονται παρέχουν επίσης πληροφορίες για τις στρατηγικές σύντηξης που θα χρησιμοποιηθούν.

Η έρευνα αντιμετωπίζει συστηματικά τις προκλήσεις στην προεπεξεργασία δεδομένων, την εξαγωγή και επιλογή χαρακτηριστικών, την ανάπτυξη μοντέλων και τη σύντηξη τόσο σε επίπεδο χαρακτηριστικών όσο και σε επίπεδο αποφάσεων. Μέσω ενδελεχούς αξιολόγησης, εκτιμάται η ευρωστία και η ακρίβεια διαφόρων στρατηγικών πολυτροπικής σύντηξης και αλγορίθμων ταξινόμησης που εφαρμόζονται στην εργασία της αναγνώρισης συναισθημάτων.

Τελικά, η εργασία αυτή έχει ως στόχο να καταδείξει ότι η στρατηγική σύντηξη των δεδομένων των φορετών βιομετρικών αισθητήρων μπορεί να βελτιώσει σημαντικά την απόδοση των συστημάτων αναγνώρισης συναισθημάτων. Σε αυτά τα πλαίσια, η σύντηξη χρησιμοποιείται για την εκμετάλλευση των ξεχωριστών χαρακτηριστικών του κάθε αισθητήρα, αναδεικνύοντας τα μοναδικά τους πλεονεκτήματα και ταυτόχρονα περιορίζοντας τις επιμέρους αδυναμίες τους. Αναμένεται ότι οι πληροφορίες που θα προκύψουν από την παρούσα έρευνα θα συμβάλουν στην ανάπτυξη αξιόπιστων τεχνολογιών -με εφαρμογές στην Υγεία, τη διαχείριση του άγχους και τη διαγνωστική υποστήριξη της συναισθηματικής ευεξίας- οι οποίες θα υλοποιούνται μόνο με τη χρήση εύκολα προσβάσιμων φορετών συσκευών.

Acknowledgements

This thesis has been a profound opportunity for the author to delve deeper into the field of Machine Learning, a domain that ignites their passion and drives their curiosity. It has provided valuable insights, expanded the author's knowledge, and showed the transformative potential of artificial intelligence, a field that remains inherently human-centered and aims at enhancing humans' lives. The author expresses heartfelt gratitude to all the professors who have guided their academic journey, equipping them with the skills and knowledge essential for the successful completion of this work. In particular, the author wishes to express deep appreciation to Professor Konstantinos Goulianas, whose unwavering passion for the discipline and the Department has been a source of inspiration from the very beginning. His example encouraged the author to pursue Machine Learning, an academic path for which the author will remain forever grateful. Special thanks are due to CERTH for providing the dataset used in this thesis and for entrusting the author with a topic of both academic and real-world significance. The author also wishes to express heartfelt thanks to Vasilis Xefteris, whose guidance, insights, and generosity in offering this opportunity played a crucial role in shaping the outcome of this thesis. The opportunity to work alongside such dedicated and passionate researchers was a privilege. The author also thanks their friends for their patience, encouragement, and practical support throughout this attempt, helping to maintain resilience during challenging times. Lastly, the deepest gratitude goes to the author's parents, whose unwavering support, understanding, and belief provided the foundation upon which this academic journey was built. To all of you, thank you, from the bottom of the author's heart.

Contents

List of Figures

List of Tables

List of Mathematical Relations

List of Abbreviations

AC	Alternating Current
AI	Artificial Intelligence
ANS	Autonomic Nervous System
AUC	Area Under Curve
CERTH	Centre for Research and Technology Hellas
CNN	Convolutional Neural Network
DC	Direct Current
DL	Deep Learning
ECG	Electrocardiograph / Electrocardiography
EDA	Electrodermal Activity
ER	Emotion Recognition
ERF	Effective Receptive Field
FACS	Facial Action Coding System
GSR	Galvanic Skin Response
HCI	Human-Computer Interaction
HF	High Frequency
HR	Heart Rate
HRV	Heart Rate Variability
HVHA	High Valence, High Arousal
HVLA	High Valence, Low Arousal
IBI	Inter-Beat Intervals
IHU	International Hellenic University
IR	Infrared
KNN	k-Nearest Neighbors
LF	Low Frequency
LSTM	Long Short-Term Memory
LVHA	Low Valence, High Arousal
LVLA	Low Valence, Low Arousal
ML	Machine Learning
NLP	Natural Language Processing

PAD	Pleasure-Arousal-Dominance
PPG	Photoplethysmograph / Photoplethysmography
PRV	Pulse Rate Variability
PSD	Power Spectral Density
RF	Random Forest
RNN	Recurrent Neural Network
SCL	Skin Conductance Level
SCR	Skin Conductance Response
SVC	Support Vector Classifier
SVM	Support Vector Machine
THERM	Temperature
VR	Virtual Reality

Introduction

We are moving towards an era where computers become an irreplaceable part of our society. Our field develops rapidly and new technologies emerge at an unprecedented pace. The advancement of Artificial Intelligence in particular already shows its vast and diverse capabilities, and its potential to enrich our lives and evolve the human potential.

One long-standing puzzle for scientists and AI has been the task of Emotion Recognition. It's the process of predicting the emotional state of a person, by using Machine Learning pipelines, and data derived from a diverse group of sensors, either on-body, or out-of-body. Its importance mostly lies in human-computer interaction and healthcare, such as stress management.

Sensors are becoming more robust, better documented and well built. This allows for better representation of the data derived from them, thus increasing our ability to extract meaningful features from them for usage in ML pipelines. The emergence of commercially available and well-integrated multi-sensors in particular, opens the door to the processing of different types (modes) of data simultaneously, aligned on the time axis. On top of that, most sensors are nowadays built compact, and could be integrated into uniform wearable devices, thus enabling the deployment of ML-driven systems on the go.

One of the technologies that makes such processing of different types of data possible, is called Multimodal Fusion. Multimodal Fusion pipelines aim to align and fuse the data derived from these multi-sensors, and use the combined inputs or features to train a ML model. Compared to unimodally trained models, it is used to exploit the distinct characteristics of each sensor, showcasing their unique advantages while mitigating their individual weaknesses.

This thesis aims to show that Multimodal Fusion systems outperform conventional models when evaluated on the task of ER. This insight makes the use of wearable biometric sensors more feasible than before for the task, which in turn opens the door to the development of consumer-grade, readily available and reliable affective technologies.

The dataset used for the thesis was created and kindly provided to me by CERTH, and contained different modalities of real physiological data derived from a wearable multi-sensor, such as PPG, GSR, and skin temperature readings, accompanied with the ground truth responses of the emotional state of 25 subjects.

Besides a comprehensive literature review, the presentation of the methodology, results, figures and tables, the deliverables of this thesis include the full Python scripts used for the processing of the data, feature extraction, visualization of graphs, and training of the models.

In Chapter 1, a literature review is conducted, where foundational concepts such as Emotion Recognition, Multimodal Fusion, Physiological Signals and common ML pipelines are explained and analysed.

In Chapter 2, the data processing pipeline is showcased, the different methods of fusion that were used, ranging from classical feature-based approaches to a CNN-based multimodal fusion architecture, the manual feature extraction pipeline, and the different models that were trained and their hyperparameters, as well as the decisions that were made during each part.

In Chapter 3, the results are presented, including performance comparisons and visualizations that highlight the details and the contributions of each approach, as well as the interpretation and justification of the aforementioned results.

In Chapter 4, a discussion regarding the findings is provided, along with the implications and limitations of both this thesis and the task in general, followed by proposed directions for future research.

The Appendices contain supporting materials, such as seed and DL model implementation code, as well as code usage instructions.

Chapter 1: Literature Review

1.1 Introduction to Literature Review

In this chapter, the foundational concepts of this thesis are reviewed and analysed. Emotion Recognition is introduced along with its definition, its importance in human-computer interaction, healthcare, affective computing, and stress management, as well as the common emotion models used in research.

The chapter then examines the groups of modalities often used for ER tasks, such as visual-based, audio-based, text-based, and physiological, and compares their respective strengths and limitations, with a focus on physiological modalities due to their relevance in this work. Physiological signals such as PPG, GSR, respiration, and skin temperature are described, as well as their connection to the autonomic nervous system and emotional states.

Multimodal Fusion is defined, along with its motivations, and fusion strategies such as feature-level and decision-level fusion are discussed. Finally, the chapter presents machine learning approaches used for ER, including both classical and deep learning models, and compares feature engineering with feature learning. Challenges associated with data acquisition for ER, such as inter-subject variability, sensor noise and artifacts, data scarcity and real-time processing constraints, are also highlighted.

The goal of this review is to provide the necessary theoretical background to motivate and contextualize the methodological choices presented in Chapter 2.

1.2 Emotion Recognition

1.2.1 Introduction to Emotion Recognition

Emotion Recognition is the field in which the representation and the interpretation of human emotions by computers is studied. It is the task where we use and manipulate affective data and apply computations to it, in order to predict the emotional state of an individual.

It’s a complex classification problem that often requires advanced pattern recognition, since the boundaries between emotion labels are often obscured, and the relationships between raw sensor data and emotions are non-linear [1], [2]. This makes non-deterministic solutions, such as Machine Learning models, promising candidates for the task.

ER is considered a stepping stone toward advanced human-computer interaction, as humans interact naturally and socially with computers, making affective recognition necessary for understanding user intent [3]. Furthermore, understanding intent and connecting it with a measurable emotion could open the door for greater understanding of human emotions in general, and provide insights in neuroscience [4].

A variety of modalities can be used for this pattern recognition task, including visual-based, text-based, audio-based, and physiological signals [1]. What makes ER a particularly complex problem is the difficulty of reliably expressing, differentiating, and labeling emotions and their nuances [5], further complicated by inter-subject variability, both in how individuals express emotions and in how consistent such expressions are over time. These factors also introduce challenges in data acquisition.

1.2.2 Definition of Emotion Recognition

Emotion Recognition is the computational task of detecting and interpreting human emotional states from various types of data, often derived from sensors. It involves mapping inputs, such as physiological signals, facial expressions, speech, or text, to either predefined emotion labels for classification tasks or continuous affective values, such as Valence and Arousal, for regression tasks [1].

As a pattern recognition problem, ER typically includes signal acquisition, preprocessing, feature extraction, and the training and inference of machine learning models [2]. Originating from the broader field of affective computing [3], ER aims to enable the development of emotionally intelligent systems that can perceive and respond to user affect.

While implementations may vary depending on the application, goal or modalities used, the core objective remains to algorithmically bridge the gap between measurable input signals and the affective states they reflect.

1.2.3 Importance in HCI, healthcare, affective computing

As stated, Emotion Recognitions stems from the broader field of affective computing, which seeks to bridge the gap between the user's affective state and the computer's response to it [3]. There are three theoretical levels: computers that detect emotion, computers that express emotion in a human-like manner, and computers that have emotions, comparable to humans [2], [3]. ER is only one part of the process, but even standing alone it has many uses, such as in human-computer interaction and healthcare.

Emotions have computational value because they are a form of expression, based on neuroscience. Findings suggest that when emotions occur, specific regions of the conscious brain show increased activity, and directly influence cognition, perception and decision-making [4]. Therefore, measurable emotions are strong indicators to the user's intent [3], [4].

This lays the foundation for advanced HCIs, where systems use ER to detect precise intent, interpret the user's goals effectively, and respond with a complex, human-like nature. Example applications include intelligent robotics, uses in entertainment industry, such as adaptive gaming and VR implementations [6], uses in education systems, where affect-aware systems identify confusion or frustration in learners and detect complications in current educational methods, uses in work environments, where affective feedback surfaces hidden problems in productivity and well-being, and uses in websites and apps, where targeted solutions or suggestions are given based on the intention of the user [7], [8]. In summary, implementation of ER in HCIs has the potential to enhance the quality of the user's experience and streamline their work flow.

More importantly, ER could be used in healthcare, since emotions alter bodily functions, such as metabolism, body temperature, heart rate, and blood pressure. Moreover, information on the emotional state of the patient could give crucial information to their caregivers, as to the direction of the treatment or rehabilitation. However, it should be stated that since ML models are probabilistic, more research is required for them to be reliably deployed in healthcare systems, and the development of Explainable AI for such systems is important, for trust and interpretability [5], [9]. ER could nowadays be used for non-diagnostic, but nonetheless information-rich tasks, such as early stress recognition, for people who suffer from chronic or acute stress syndromes.

In summary, ER is a versatile tool with applications across domains. Its ability to transform the emotional state into a digital input has the potential to enhance the quality of the interaction and possibilities of computers, focusing on intelligent and human-centric technologies.

1.2.4 Challenges in Emotion Recognition

Despite its growing importance, emotion recognition faces numerous technical, conceptual, and ethical challenges. A core issue lies in the inherent ambiguity and subjectivity of emotions. Individuals often struggle to accurately express or even self-identify their emotional states, making ground truth labeling difficult and noisy [2], [3], [5]. Furthermore, many existing datasets are skewed toward discrete and intense emotional categories, like anger, fear, joy, leaving subtler or mixed affective states, like mild anxiety and discomfort, or more complex ones, like love and guilt, underrepresented, thus limiting the validity of trained models [3]. In data acquisition, ground truth emotion annotations are prone to variability and subjectivity, especially near the regions of ambiguous emotional states, where classification is inherently harder.

The computational burden also grows with the number of emotional categories subject to classification, requiring more data and larger models to capture fine-grained distinctions between emotional nuances [2], [3]. This is further complicated by inter-subject variability, as individuals express the same emotion differently, and physiological responses are highly personalized, influenced by factors such as age, gender, baseline physiology, and even cultural background [10]. This variability poses a major challenge for building well generalizable models.

Additionally, sensor inconsistencies, such as electrode placement, environmental noise, and signal corruption, introduce further unpredictability in data acquisition. Lastly, ethical concerns must be addressed. Emotions are deeply private and context-dependent, and the use of physiological or behavioral data for ER raises issues of consent, surveillance, and emotional manipulation. As Rosalind Picard notably argued, affective computing systems must be designed with a strong emphasis on transparency, user control, and ethical safeguards [5].

1.3 Emotion models

1.3.1 Definition of emotion models

In order to predict an emotional state using supervised ML, it is first necessary to define the target labels for classification models or the axes of measurement for regression models. This requires a formalized model of emotion, which is a framework used to interpret, and quantify emotional states. Over the years, psychologists have proposed several theories for the definition of emotion, for the number of emotions that exist, and whether we can interpret emotions in discrete and separate states, or a more complex, continuous way is required. They fall into two main categories: discrete, or categorical models, and dimensional models [11 - 14].

1.3.2 Discrete models

In discrete or categorical emotion models, emotional states are represented using class tags. One of the most influential models in this category is Paul Ekman’s six basics emotions model, namely happiness, anger, sadness, fear, surprise and disgust [12]. These emotions are considered distinct, and each is linked to a specific set of behaviors, facial expressions, and unique physiological signatures. Other researchers expanded on this idea, and added more nuanced emotional states on the existing list, such as love,

anticipation, or guilt. Robert Plutchik proposed a wheel emotion model, with categorical labels, but accounting for polarity and similarity in their placement on the wheel [15].

Discrete models have the advantage that they are easy to understand, and fit particularly well in classification tasks in ML, as they map directly to fixed output labels. However, they don't account for the nuances of emotions, they can't quantify the strength of the emotion, they overlook the influence of concept, culture and individual differences [13].

1.3.3 Dimensional models

Dimensional models, in contrast, aim to interpret emotions as points in a continuous space, rather than as distinct categories. The most widely adopted dimensional framework is the Valence-Arousal model, attributed to James Russell's Circumplex Model of Affect [11]. In this model, Valence describes the positivity or negativity of the emotional state, ranging from pure happiness to total displeasure, while Arousal describes the intensity of the emotion, ranging from pure stimulation to complete dullness. Albert Mehrabian proposed a model that introduces a third axis on this schema, called Dominance axis, thus creating the PAD model [16]. The Dominance axis describes the control of the individual over the emotion and vice-versa, ranging from total to complete lack of control.

Dimensional models are particularly useful for regression-based tasks, as they better capture the gradations of emotional states. They suggest that any emotion can be described as a combination of Valence and Arousal points. A particular advantage also lies in data collection, as it is easier for the participants to provide ground truth responses in a schema-like manner, rather than having to choose one label out of a set of distinct, widely variable ones. Their downside is the interpretability of the results, especially in applications requiring classification [13].

This thesis uses the Valence-Arousal model, with responses given between a fixed space of negative and positive integer values for each axis. These values are then encoded to labels, based on their position on the schema.

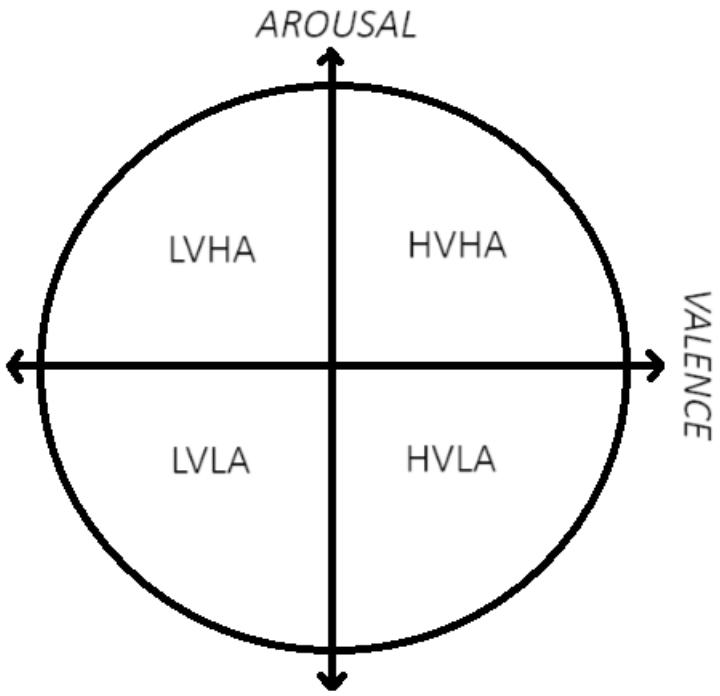

Figure 1.1: Valence-Arousal schema. Regions shown counter-clockwise: High Valence - High Arousal, Low Valence - High Arousal, Low Valence - Low Arousal, High Valence - Low Arousal.

1.4 Modalities for Emotion Recognition

1.4.1 Overview of Modalities used for ER

With the term modality, we refer to a distinct type of data than can be used as input for a ML task. In Emotion Recognition, several modalities can be used depending on the goal and restrictions, as each one captures a different facet of emotional expression, or a combination of them [1]. Commonly used modalities in ER include visual-based, which capture visual behavioral cues from facial movements and body language, text-based, which extract affective value from written text, audio-based, which detect affective states from changes in vocal tones in speech, and physiological-based, which consist of physiological data derived from sensors, such as ECG, and use signal variations over a period of time to detect an emotional state [14].

Visual emotion recognition primarily relies on facial expressions, body gestures, and eye movement. The most common visual method is facial expression analysis, since certain emotional expressions are universal, as suggested by Ekman and which led to the creation of the Facial Action Coding System [17]. Facial expression analysis methods often fall into either the geometry-based, or the appearance-based category. Limitations include noise, artifacts, insufficient lighting and obstructions [14]. Furthermore, these methods raise privacy-related concerns.

Audio-based emotion recognition analyses speech prosody, including pitch, tempo, and rhythm. Emotional states modulate features like vocal fundamental frequency and speech rate, thus making them detectable through signal processing and ML methods [18], [19]. Audio data is non-invasive, and occurs naturally in a conversation, however it is prone to noise, it is language-dependent, and raise privacy issues.

Text-based emotion recognition uses natural language processing, to infer emotion from word usage, syntax choices, and semantics of language. One approach relies on the usage of lexicons, which contains pre-defined affective word dictionaries, which map a certain word to an affective state [14], [20]. However, this method does not account for word placement in a sentence and overall context. More sophisticated approaches include the utilization of DL models and advanced architectures, such as the transformer, to extract contextual emotion cues [20]. Text-based methods are effective in digital communication, but they lack the physiological or acoustic depth present in biosignals and speech.

Physiological signals reflect autonomic nervous system activity, making them suitable for objective affect detection, particularly in contexts where verbal or facial cues are unreliable. They are commonly referred as biosignals, since they are directly related with vital functions of the body [10]. Common signals include:

- Electrodermal Activity (EDA): Measures changes in skin conductance due to sympathetic nervous system activation. Associated with sweat gland activity, which correlates with emotional Arousal and is frequently used in stress and affective state detection [10].
- Electrocardiography (ECG): Records the electrical activity of the heart via electrodes placed on the skin. It allows measurement of heart rate and heart rate variability, both of which provide information to emotional stress, cognitive load, and affective Valence [10], [21].
- Photoplethysmography (PPG): Uses varying light wavelengths to detect blood volume changes in peripheral tissues through the skin, typically on the finger or the wrist [22]. It is less invasive than ECG, easily integrated into wearables because of compact sensor size, and while slightly less accurate, it effectively estimates heart rate and HRV [23]. PPG works by emitting light into the skin and measuring the amount reflected back to the sensor or absorbed, which varies with blood flow.
- Skin Temperature: Emotionally driven autonomic responses, such as vasoconstriction under stress and vasodilation in calm settings, lead to measurable changes in peripheral skin temperature, typically decreasing during negative Arousal states and increasing in positive ones [7].
- Electroencephalography (EEG): Captures electrical brain activity through electrodes placed on the scalp. EEG is used to link cognitive and emotional processes with specific patterns in frequency bands of the brain, like alpha, beta, and theta [10], [21]. Though effective, it is considered an invasive method, mainly due to the large number of electrodes required for data acquisition and the difficulty associated with applying it and debugging it in non-laboratory settings.

These modalities often require specialized equipment and careful calibration but offer insights that are not possible with visual or audio data alone, mainly because they are objective, and less susceptible to masking or deception [3].

This thesis uses Physiological Signals, namely PPG in three wavelengths, EDA, and a THERM signal. Below, a more comprehensive review of these signals is conducted.

1.5 Physiological Signals in Emotion Recognition

1.5.1 Description of Physiological Signals

Physiological signals focus on the internal states of the human body and offer insights into the functioning of the autonomic nervous system, which plays a critical role in emotional processing. Unlike behavioral modalities like facial expressions and speech, physiological responses are involuntary and less susceptible to conscious control, making them particularly valuable for emotion recognition tasks where authenticity and consistency are essential.

These signals provide deep computational value, because they are generated by complex physiological processes, such as cardiovascular, electrodermal, and thermoregulatory activity, that are modulated by affective stimuli. By capturing and analysing these biosignals, we aim to infer emotional states with higher objectivity and reliability. Additionally, the use of physiological data, when combined with machine learning models, enables real-time, continuous, and non-invasive assessment of affective states.

In this thesis, five such modalities were used for the Emotion Recognition task: photoplethysmography in three wavelengths (PPG), electrodermal activity (EDA), and skin temperature (THERM). Each of these signals offers a unique physiological perspective, reflecting sympathetic and/or parasympathetic activity, and will be discussed in detail in the following sections.

1.5.2 Photoplethysmography (PPG)

Photoplethysmography (PPG) is an optical measurement technique used to capture blood volume changes in the microvascular bed of tissue. It operates by emitting light, typically in the green, red and infrared spectrum, into the skin and measuring the variations in light absorption or reflection caused by pulsatile changes in blood flow [24].

These fluctuations correspond to cardiac cycles, making PPG a non-invasive proxy for heart activity. Unlike electrocardiography, PPG does not require direct electrical contact and can be implemented with compact wearable sensors, making it more practical for deployment in non-laboratory settings [23]. PPG reflects both sympathetic and parasympathetic modulation of heart rate, which are known to vary with emotional Arousal and stress.

The signal typically consists of its pulsatile (AC) and baseline (DC) components, and each pulse has a form of a sharp peak, followed by the dicrotic notch [7], [23], [25]. From the PPG waveform, several time-domain features can be extracted, such as heart rate, inter-beat intervals, and pulse rate variability, which approximate heart rate variability, a key biomarker of autonomic activity [10], [26]. Frequency-domain features can also be derived through the usage of filters or transforms to assess low-frequency and high-frequency bands, associated with sympathetic and parasympathetic balance.

Preprocessing of PPG usually involves band-pass filtering (such as 0.5 - 5 Hz) to isolate the cardiac component, followed by normalization techniques, and peak detection algorithms to identify systolic peaks [26]. However, PPG is susceptible to motion artifacts, poor skin contact, and ambient light interference, which necessitates robust filtering and signal quality assessment. Despite these limitations, PPG remains a core signal in emotion recognition systems due to its practicality, physiological relevance, and rich feature space.

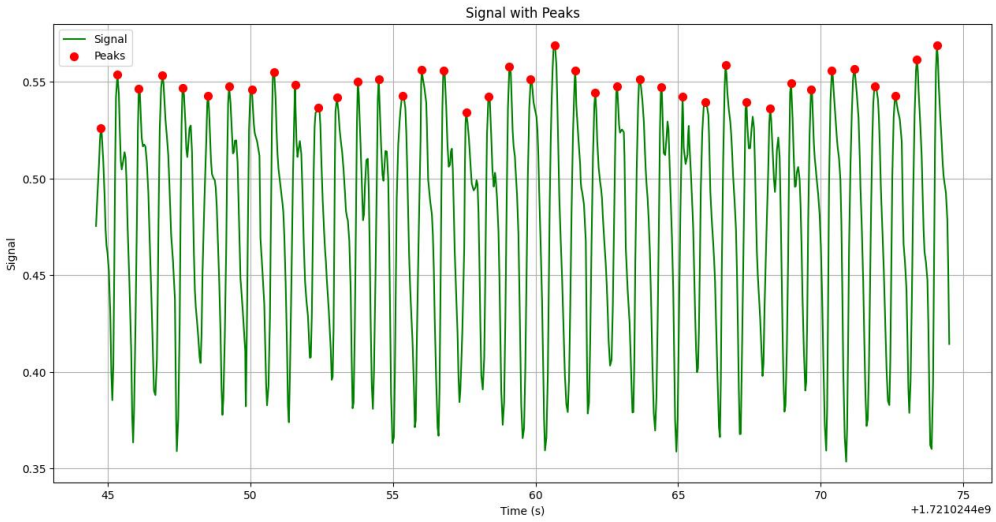

Figure 1.2: A PPG signal with detected peaks.

1.5.3 Electrodermal Activity (EDA)

Electrodermal Activity (EDA), also referred to as Galvanic Skin Response (GSR), reflects changes in the electrical conductance of the skin, which is modulated by sweat gland activity under the control of the sympathetic branch of the autonomic nervous system. As such, it serves as a direct and robust physiological correlate of Arousal [7].

EDA is typically acquired using surface electrodes placed on areas with high sweat gland density, such as the fingers or palms. The signal is composed of two components: the tonic component, or Skin Conductance Level (SCL), which reflects slow, baseline-related changes, and the phasic component, or Skin Conductance Response (SCR), which captures rapid, transient changes in conductance related to spontaneous Arousal fluctuations. To effectively analyse EDA, the signal is often decomposed into its tonic and phasic components, as these carry different types of information [27]. Decomposition enhances the interpretability of the signal and allows for the extraction of more meaningful features such as SCR amplitude, latency, and frequency, which are closely tied to emotional and cognitive states.

The EDA signal is commonly low-pass filtered, normalized to reduce inter-subject variability, decomposed, and then subjected to feature extraction. Common features include statistical features, such as mean and standard deviation, peak analysis for the phasic component, such as number and amplitude of SCR, and signal derivatives [10], [27].

Despite its high sensitivity to Arousal-related emotional states, EDA is prone to motion artifacts and environmental influences, and inter-subject variability remains a significant challenge [7]. Nonetheless, its high temporal resolution and direct link to sympathetic activation make it a valuable signal in emotion recognition tasks, especially when fused with complementary modalities.

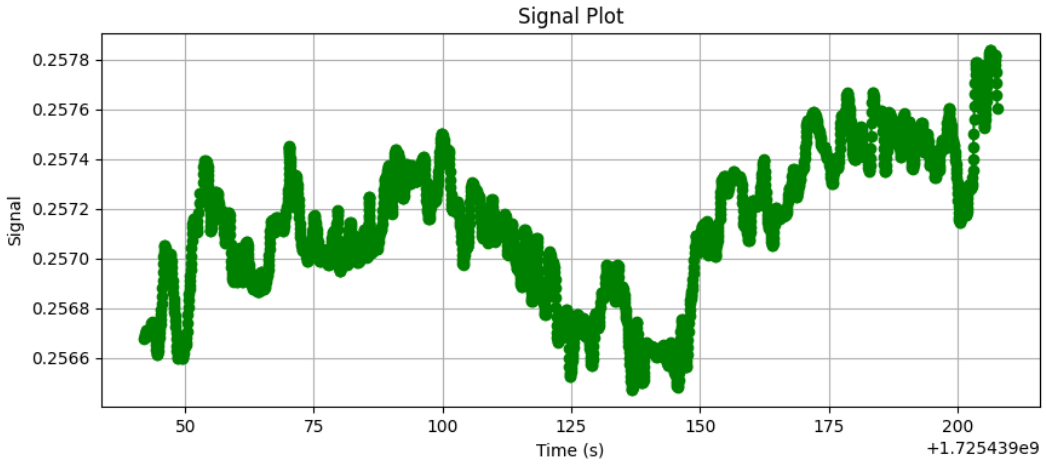

Figure 1.3: An EDA signal.

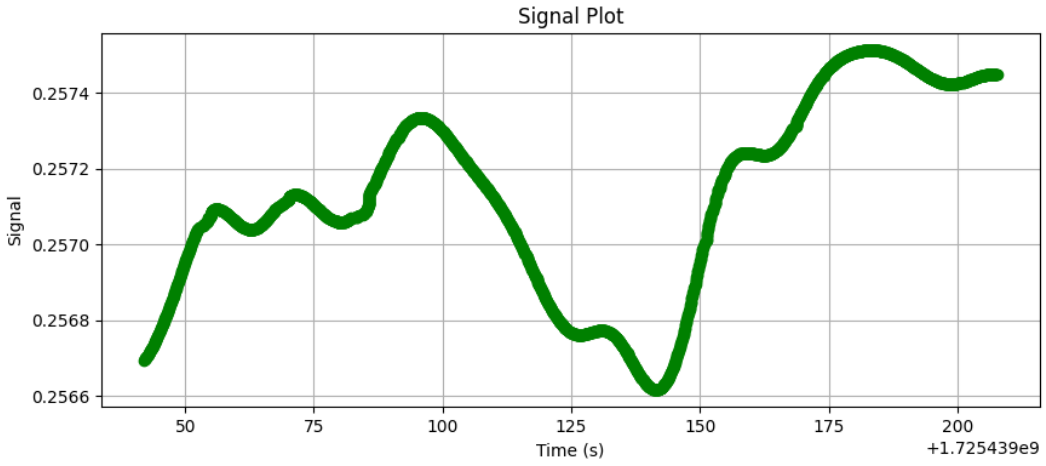

Figure 1.4: The tonic component (SCL) of the signal.

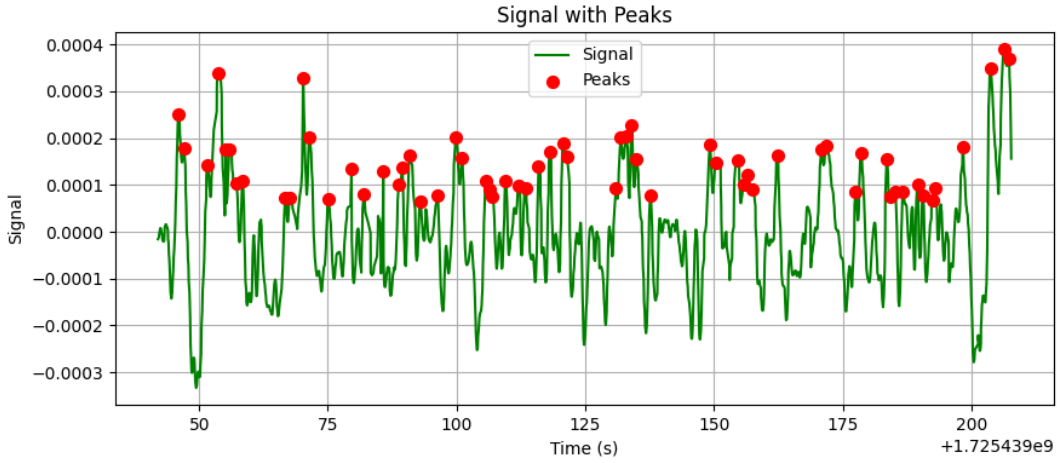

Figure 1.5: The phasic component (SCR) of the signal, with the detected peaks.

1.5.4 Skin Temperature (THERM)

Skin Temperature is a slow-varying physiological signal that reflects thermoregulatory processes governed by the autonomic nervous system, particularly via sympathetic vasoconstriction and

vasodilation of peripheral blood vessels [10]. Emotional states, especially those linked to stress or Arousal, can subtly influence skin temperature due to changes in blood flow. For instance, anxiety or fear may lead to peripheral vasoconstriction, reducing skin temperature at extremities. In this work, derived skin temperature features include simple statistical features, like mean and standard deviation, and first derivatives to capture low-frequency trends. Regarding preprocessing, no filtering steps were required, as the produced THERM signals were stable and did not suffer from motion artifacts or disruption like PPG, making normalization the only processing method that was applied.

Although skin temperature alone does not offer strong temporal resolution and is affected by ambient conditions, like room temperature or sensor placement, it can serve as a complementary modality in multimodal emotion recognition when combined with higher-resolution signals like PPG or EDA. Recent literature supports its inclusion in affective recognition [7], [10], and lists skin temperature among the viable biosignals for emotion-related state inference, particularly in wearable systems. However, due to its low reactivity, skin temperature is generally not used as a standalone predictor but contributes valuable context during fusion in multimodal setups.

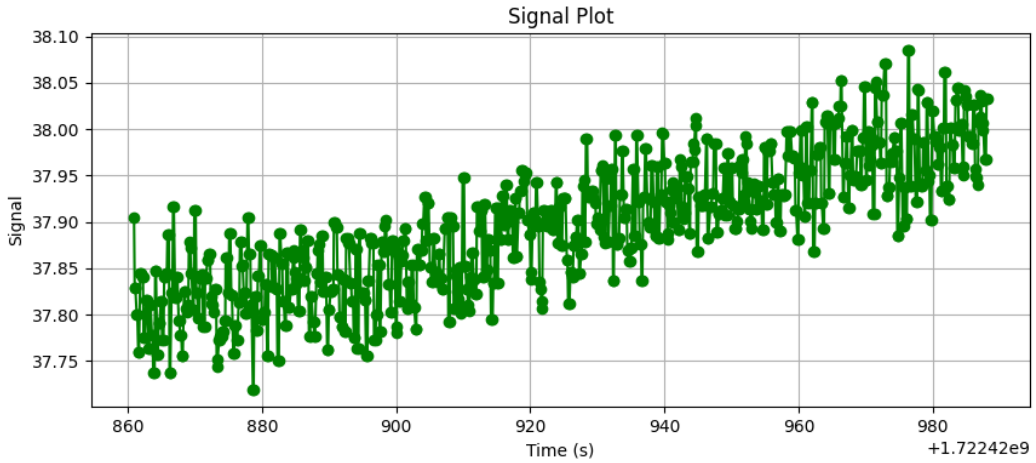

Figure 1.6: A THERM signal.

1.5.5 Summary of Physiological Signals, Importance and Challenges

Physiological signals offer a robust and objective approach to emotion recognition, as they reflect autonomic nervous system activity that often precedes or bypasses conscious control. Unlike facial expressions, speech, or behavioral cues, which can be voluntarily masked, culturally biased, or dependent on environmental context, physiological signals such as PPG, EDA, and skin temperature provide continuous, real-time insights into internal affective states [2], [7], [10]. Modern physiological sensors are typically small, wearable, and mostly non-invasive, enabling unobtrusive data acquisition even in real-world settings. This makes them particularly valuable in applications requiring passive monitoring, such as mental health assessment, stress detection, or adaptive human-computer interaction.

However, these signals also introduce significant challenges. These include low signal-to-noise ratios, susceptibility to motion artifacts, inter-subject variability, skin properties, ambient temperature, and electrode placement. Moreover, there are modality-specific limitations, since no single physiological signal provides a complete representation of an individual's emotional state. For example, PPG primarily reflects cardiovascular reactivity, EDA is closely tied to sympathetic Arousal but lacks

Valence specificity, and skin temperature reflects thermoregulatory processes that evolve over longer time scales.

To mitigate these limitations and increase generalizability, researchers have increasingly turned to multimodal fusion [1], [2], [7], [14], which leverages complementary information across signals to enhance robustness, accuracy, and resilience against sensor-specific constraints, noise or signal dropout. The following section addresses the rationale, techniques, and implementation strategies for physiological signal fusion in the context of emotion recognition.

1.6 Multimodal Fusion

1.6.1 Introduction to Multimodal Fusion

Emotion recognition systems increasingly leverage multimodal data sources to improve their robustness, generalizability, and accuracy. Traditional unimodal approaches, relying on a single source of input such as facial expression, speech, or physiological signals, are often limited by noise, modality-specific shortcomings, or situational constraints. For instance, facial recognition can be compromised by occlusion or lighting, while speech-based systems may fail in noisy environments. These limitations have led to the growing adoption of multimodal fusion, which integrates complementary information from multiple channels, such as physiological signals, audio, video, or text, to create a more holistic and resilient method for predicting affective states [2], [7], [14].

The central idea is that by combining different modalities, a system can compensate for the weaknesses of individual signals while amplifying their strengths. This approach is especially relevant in affective computing, where emotions manifest through diverse and complex physiological and behavioral pathways. Importantly, this principle aligns with how humans naturally perceive and interpret emotions, through an inherently multimodal process that integrates visual, auditory, and physiological cues [14]. Since human affect perception relies on the fusion of multiple sensory channels, multimodal systems are not only more robust but also more biologically and cognitively suitable. Fusion techniques are thus increasingly seen as essential for reliable emotion recognition, particularly in real-world applications involving variability in users, contexts, and environments [28].

In recent years, advances in machine learning and deep learning have further propelled interest in multimodal systems, enabling sophisticated architectures capable of learning cross-modal relationships via feature engineering or feature learning and fusion methods. These developments have made multimodal fusion not only feasible but also practical for domains such as healthcare, education, human-computer interaction, and adaptive systems [14].

1.6.2 Definition and Motivations of Multimodal Fusion

Multimodal fusion refers to the integration of information from multiple heterogeneous data sources, often referred as modalities, to improve the performance and robustness of ML systems, such as for emotion recognition tasks. In affective computing, these modalities typically include physiological signals, like PPG, EDA, skin temperature, visual inputs, like facial expressions, body language and gesture recognition, audio, like vocal prosody in speech, and textual data. The aim is to combine these complementary modes in a way that maximizes their informative value and minimizes individual weaknesses, leading to more reliable and generalizable predictions of emotional states [7].

The motivation for multimodal fusion is both empirical and theoretical. From a theoretical standpoint, emotional experiences are inherently multimodal, involving simultaneous changes in physiology,

expression, and behavior. Relying on a single modality often results in incomplete or ambiguous representations of affective states due to noise, occlusion, voluntary control, or modality-specific limitations. For example, facial expressions may be masked or culturally modulated, while physiological signals may be susceptible to noise or individual variability. Multimodal systems can cross-validate affective information across modalities, increasing the reliability of emotion detection [14].

From a machine learning perspective, multimodal fusion enhances feature diversity and enables models to learn richer, more abstract representations. It allows learning algorithms to detect associations between modalities, which can lead to better generalization, especially in complex tasks or noisy environments. Deep learning models, such as convolutional or recurrent neural networks, are particularly well-suited for multimodal learning, as they can learn nonlinear relationships across diverse feature sets. Moreover, fusion allows flexible adaptation to missing or degraded input channels, a common scenario in real-world deployments [28]. This suggests that in the case of a sensor failure during deployment, the system can continue functioning.

Multimodal fusion is also crucial for user-centric and context-aware affective systems, where robustness across individuals and environmental conditions is necessary. Systems that integrate distinct physiological signal modalities, for instance, are more resilient to voluntary masking of affect, while also being more sensitive to subtle or internalized emotional states that may manifest in one physiological channel but not others, thus accounting for inter-subject variability [5], [10].

Ultimately, multimodal fusion represents a biologically inspired and practically necessary strategy to approach the complexity of human emotions in computational systems. Below, an overview of the methods used for multimodal fusion is presented.

1.6.3 Fusion Levels

1.6.3.1 Introduction to Fusion Levels

Multimodal fusion strategies are commonly categorized based on the stage at which data integration occurs within the processing pipeline. Commonly referred levels are early (feature-level) fusion, late (decision-level) fusion, and hybrid fusion. They represent distinct approaches to combining information from multiple modalities, each with its own advantages and limitations. Early fusion involves combining engineered or learned features extracted from each modality into a unified representation before classification, allowing the model to exploit complementary information while maintaining modality-specific representations. Late fusion integrates the outputs of modality-specific classifiers trained separately, offering robustness to missing or noisy modalities. Hybrid fusion strategies combine both feature-level and decision-level methods, to leverage the advantages of both approaches. Understanding these levels is essential for selecting a fusion strategy that aligns with the characteristics of the modalities involved, the goals of the emotion recognition task, and the constraints of real-world deployment scenarios.

1.6.3.2 Early Fusion

Feature-level fusion, commonly referred to as early fusion, involves the integration of multiple modalities by concatenating or combining their extracted features into a single, unified feature set before the learning phase. This approach enables the model to learn cross-modal relations directly, allowing the network to capture subtle interactions between modalities such as photoplethysmography (PPG), electrodermal activity (EDA), and skin temperature (THERM). These interactions can be especially

informative in emotion recognition, where different physiological pathways may respond complimentarily to affective stimuli.

Early fusion offers several advantages. By allowing the model to consider the full context of multimodal information simultaneously, it becomes fitting for capturing subtle patterns that may only emerge through the interaction of different physiological signals. This strategy has been shown to improve the robustness and accuracy of emotion recognition systems, particularly when modalities provide complementary information [2], [14]. This fusion method preserves intra-modality specificity while enabling inter-modality integration. Such an approach is particularly suitable for deep learning models, which benefit from having access to the full multimodal context during feature learning and classification [29], [30], [31].

Nevertheless, early fusion also presents specific challenges. These include issues related to the different nature of features for each modality, time synchronization, dimensionality mismatches, differences in sampling rates, missing or noisy data, and the risk of overfitting due to the high dimensionality of concatenated feature sets. These issues must be carefully addressed during preprocessing and model design, such as architecture choices and hyperparameter values. Despite these drawbacks, early fusion remains one of the most widely adopted strategies in affective computing due to its superior ability to model shared representations across physiological signals [7], [14], [30].

In some works, data (signal-level) fusion is proposed as an additional category. Data fusion involves the direct integration of raw or processed signals, typically through operations such as averaging time-series data. While conceptually straightforward, this approach introduces significant practical challenges, including mismatched sampling rates across modalities and distortion of the physiological characteristics inherent to each signal type. Moreover, data fusion is inherently limited, as it cannot be feasibly applied to modalities such as video and text [2].

In this thesis, early fusion was chosen for the implementation of multimodal fusion in the emotion recognition task, both for the classical ML models, as well as the CNN-based approach.

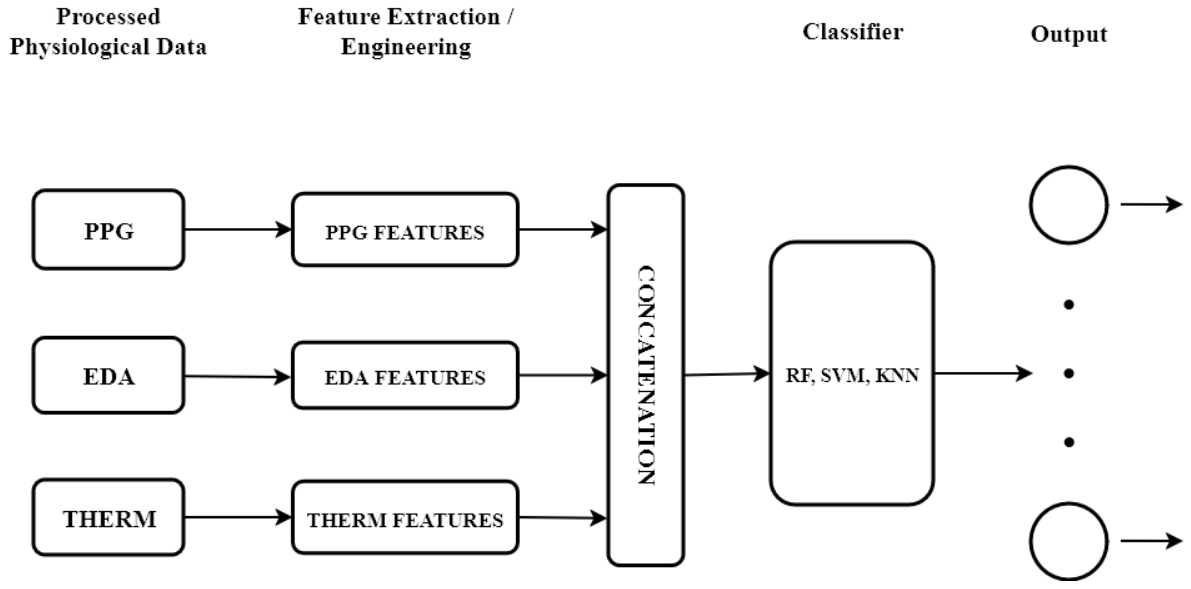

Figure 1.7: Early Fusion pipeline.

1.6.3.3 Late Fusion

Late fusion, also known as decision-level fusion, refers to the integration of predictions from multiple unimodal models, each independently trained on a distinct modality. In this approach, the outputs, typically class probabilities or confidence scores from each modality-specific classifier are combined using rule-based strategies, like majority voting and weighted averaging, or by training a meta-classifier model that uses the modality-specific predictions as inputs to make a final judgment. Unlike early fusion, where features are fused before learning, late fusion enables each modality to be modeled independently, allowing for modality-specific architectures and preprocessing techniques to be optimized without cross-modality constraints.

One of the main advantages of late fusion lies in its modularity and robustness to missing data. If one modality is unavailable or corrupted at test time, the system can still rely on other available modalities. Additionally, decision-level fusion reduces the risk of cross-modal interference that can occur in early fusion, such as noise from a particular sensor, or when signal characteristics differ greatly in temporal or spatial resolution. However, it also comes with limitations. Since the integration occurs after the individual learning stages, the model cannot learn shared representations or exploit complementary low-level features across modalities. This often results in suboptimal synergy, particularly when the emotional expression is subtle and distributed across multiple physiological channels.

In emotion recognition systems, late fusion may be preferable when modalities are highly heterogeneous, such as when combining physiological, visual, and audio data, when dataset completeness varies, or when flexibility, modular testing and interpretability are more important than maximal performance [2], [14], [28], [29], [30], [31].

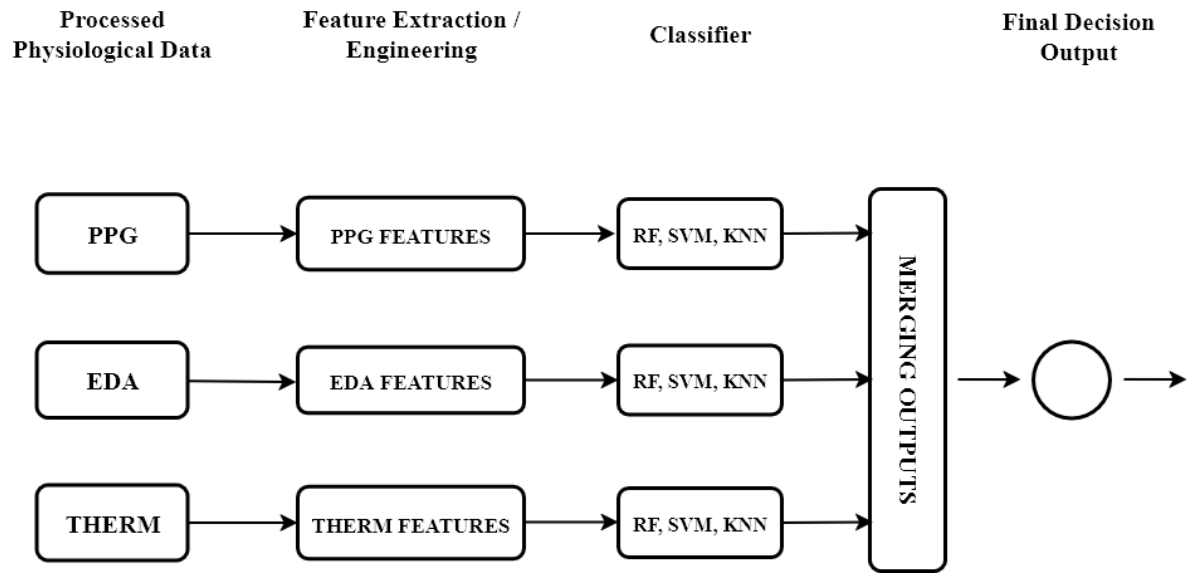

Figure 1.8: Late Fusion pipeline.

1.6.3.4 Hybrid Fusion

Hybrid fusion strategies combine elements of both early and late fusion, aiming to exploit their respective advantages while mitigating their limitations. For example, certain modalities may be fused at the feature level to capture joint representations, while others are integrated later at the decision level to preserve modality-specific insights or maintain robustness to missing data. This layered approach allows for greater flexibility in handling multimodal heterogeneity, and can be tailored to the nature of

each modality and its contribution to the affective state. While potentially more powerful, hybrid fusion typically involves increased system complexity and requires careful architectural design to ensure effective integration [14], [29], [30].

1.6.4 Summary of Multimodal Fusion

In summary, multimodal fusion is a cornerstone of modern affective computing, enabling systems to capture the multifaceted nature of emotional expression across diverse input channels. By integrating information at various levels, from raw signals to final decisions, fusion strategies enhance robustness, sensitivity, and adaptability to real-world variability. These benefits, however, are only fully acquired when paired with machine learning models, capable of capturing complex, nonlinear relationships across modalities. The following section outlines how such models are employed in the context of emotion recognition.

1.7 Machine Learning for Emotion Recognition

1.7.1 Introduction to Machine Learning for ER

Machine learning has become an essential component of emotion recognition systems, enabling the transformation of raw signals into meaningful predictions of emotional states. This process is particularly challenging due to the complex, nonlinear, and often subtle relationships between physiological or behavioral data and the emotional states they depict. In classical approaches, models such as Support Vector Machines (SVM), Random Forests (RF), and k-Nearest Neighbors (KNN) rely on handcrafted features, such as statistical and frequency-based features extracted from physiological signals, to perform classification or regression tasks [32], [33]. These models tend to work effectively with smaller datasets, but their performance is limited by the quality of feature engineering and their inability to capture complex, nonlinear dependencies.

With the rise of deep learning, architectures like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer models, have become prominent. Such models perform feature learning by automatically extracting temporal patterns from raw or minimally preprocessed data, that may not be captured through manual feature engineering. CNNs are particularly adept at capturing local spatial or temporal structures, RNNs excel at modeling sequential dependencies, and Transformers leverage attention mechanisms to capture both local and global contextual relationships [32 - 35]. Their ability to achieve end-to-end learning makes them especially powerful for large, multimodal datasets, and complex classification tasks.

However, deep learning models often require significant amounts of labeled data and have higher computational demands. A growing trend in emotion recognition tasks is the application of deep learning techniques in order to overcome the limitations posed by scarce labeled data, particularly in physiological modalities, such as transfer learning [36]. Deep architectures can automatically learn discriminative and abstract representations from raw sensor inputs, making them well-suited for wearable physiological data that are complex, noisy, and temporally structured. As a result, deep learning has emerged as a powerful tool in building robust, scalable ER systems capable of capturing subtle affective cues across individuals and contexts.

1.7.2 Classical Machine Learning Models

1.7.2.1 Overview of Classical ML for ER

Classical machine learning models have been extensively employed in emotion recognition tasks, particularly when working with physiological signals. These approaches rely heavily on the use of handcrafted features derived from domain knowledge in physiology and medicine. Features such as heart rate variability, skin conductance level, temperature gradients, and frequency-domain characteristics are often selected for their established clinical relevance and interpretability. However, the extraction of such features requires careful signal preprocessing, including filtering, detrending, normalization, and artifact removal, to ensure that they accurately reflect underlying physiological processes.

Preprocessing is therefore a critical step, as improperly treated signals can result in misleading feature values, ultimately degrading model performance. Given the need for clearly defined and well-understood input data, classical models such as Support Vector Machines, Random Forests, and k-Nearest Neighbors remain popular choices [33]. These algorithms are capable of handling relatively small datasets and offer strong performance when the input features are appropriately engineered and noise is minimized. Nonetheless, their reliance on manual feature design can limit both scalability and performance across users and contexts, as handcrafted features often fail to capture the deeper, more nuanced patterns that can only be learned directly by the model.

In addition to exploring deep learning-based feature learning approaches, this work also incorporates manually engineered features rooted in physiological domain knowledge. Initially, physiological signals were preprocessed, through filtering and normalization. Subsequently, physiological features were extracted from the processed signals and were used to train classical machine learning models such as SVM, Random Forest, and KNN. This dual approach allows for a comparative evaluation of traditional handcrafted pipelines against model-derived features, providing insights into their respective strengths in physiological emotion recognition tasks.

1.7.2.2 Support Vector Machines

Support Vector Machines are widely used in emotion recognition due to their robustness in handling small datasets with moderately high-dimensional feature spaces, conditions commonly encountered in physiological data. SVMs aim to find an optimal hyperplane that maximizes the margin between data classes, and their flexibility stems from the use of kernel functions, such as polynomial, which enable non-linear separation in transformed feature spaces [37].

In ER, SVMs have shown reliable performance for both binary and multiclass classification tasks, particularly for Valence-Arousal recognition, where emotional states are mapped onto a two-dimensional space. They are especially effective when paired with well-engineered features derived from various physiological signals [32].

1.7.2.3 Random Forest

Random Forest, a classical ML model based on multiple decision trees, is known for its interpretability, robustness to noise and outliers, and relatively low risk of overfitting. It works by training multiple decision trees on bootstrapped subsets of data and features, and aggregating their outputs, typically via majority voting for classification [33].

In ER, RF have been particularly effective in stress detection, as it can handle complex feature interactions without requiring heavy preprocessing or strict assumptions about data distributions. Their interpretability also makes them attractive in domains where feature importance is a critical concern, such as in medical-grade affective assessments [10], [32], [33].

1.7.2.4 K-Nearest Neighbors

KNN is a lazy learning algorithm that assigns class labels based on the majority vote among the k closest data points in feature space, using a chosen distance metric, commonly Euclidean distance. Its simplicity makes it appealing for baseline comparisons or low-resource systems, and it doesn't require an explicit training phase.

However, it can be computationally expensive at inference time, since it the entire dataset must be retained, and is sensitive to the choice of distance metric and feature scaling, which is especially crucial in physiological signals with different units and dynamics.

In ER, KNN is often used as a benchmark to compare the performance of more sophisticated classifiers [14], [37].

1.7.3 Deep Learning Models

1.7.3.1 Overview of Deep Learning for ER

Deep learning has become a powerful framework in emotion recognition, particularly for handling complex, high-dimensional, and heterogeneous data like physiological signals. Unlike classical models that depend on handcrafted features, deep neural networks can perform feature learning during the training phase, uncovering intricate patterns and non-linear relationships directly from raw data, that may not be prominent with manual feature extraction.

This is especially advantageous in ER, where emotional states are often encoded in subtle and distributed temporal patterns across various modalities. Architectures such as Convolutional Neural Networks, Recurrent Neural Networks, and Transformers, have demonstrated superior performance in various ER tasks across modalities, from physiological signals, to speech, video, and text [32 - 35].

These models are particularly well-suited to wearable and applications taking advantage of physiological signals, where the data are sequential and often noisy, but rich in affective cues. As such, deep learning continues to enable increasingly robust and scalable emotion recognition systems by bypassing the limitations of manual feature engineering and adapting flexibly to diverse user contexts and signal types.

1.7.3.2 CNN

Convolutional Neural Networks are a class of deep learning models originally developed for visual data processing, but they have since found widespread use across a variety of domains, including speech, text, and physiological signal analysis. Their fundamental building block, the convolutional layer, applies a set of learnable filters (or mask) across the input data to detect local patterns. This local connectivity allows CNNs to automatically extract hierarchical features, capturing low-level patterns like frequency fluctuations or local trends in earlier layers, and more abstract representations, like combinations of signal events and trends, in deeper layers [38], [39], [40].

CNNs are particularly well-suited for emotion recognition tasks due to their ability to model spatial or temporal correlations in structured input data. In the context of physiological signals, local patterns such

as sudden changes in heart rate, increases in skin conductance, or variations in temperature can carry affective significance. By leveraging weight sharing, which means that the same filter is applied across the entire input, CNNs generalize these local patterns regardless of their position in the signal, which is crucial when signal alignment varies between individuals or sessions. Furthermore, CNNs are computationally efficient due to this weight sharing and localized computation compared to fully-connected layers, making them attractive for real-time or wearable implementations [38].

1.7.3.3 1D-CNN

While 2D-CNNs are conventionally used for image data, 1D-CNNs are particularly effective for temporal signals like those acquired from wearable physiological sensors. In a 1D-CNN, the convolution operation is performed across a single temporal dimension, enabling the model to learn sequential dependencies and localized temporal features directly from time-series data such as electrodermal activity, photoplethysmography, or skin temperature [39 - 41].

A key strength of 1D-CNNs lies in their ability to capture temporal structures, such as peaks, notches, trends, and rhythms, that are often indicative of emotional responses. Unlike classical feature engineering methods that require domain expertise to define relevant features, 1D-CNNs learn discriminative patterns from the data itself during the training process. This removes the need for manual feature selection and allows the model to discover subtle, complex characteristics that may otherwise go unnoticed.

Moreover, the weight-sharing property of convolution ensures that a learned feature, such as a stress-related spike in EDA, can be recognized regardless of its position within the signal segment. This makes 1D-CNNs both powerful and robust for processing physiological data, which often exhibit inter-subject and intra-subject variability in temporal dynamics and signal morphology.

These advantages make 1D-CNNs a compelling choice for emotion recognition pipelines using raw or lightly preprocessed physiological signals, especially when real-time performance and adaptability are required, such as in wearable applications.

For these reasons, this thesis adopts a multimodal fusion approach based on 1D-CNNs.

1.7.3.4 RNN

Recurrent Neural Networks are another deep learning architecture commonly applied to emotion recognition, particularly in tasks involving sequential data. Unlike feedforward models, RNNs are designed to capture temporal dependencies by maintaining a form of memory across time steps, making them suitable for analysing time-series signals such as physiological data, speech, or video frames [38]. Their ability to model the dynamic evolution of emotional responses has made them popular in affective computing, especially for detecting changes in emotional state over time [32], [33], [39], [41]. Variants such as Long Short-Term Memory have been especially effective due to their capacity to handle longer temporal dependencies and mitigate issues like vanishing gradients [38]. While not the focus of this work, RNNs represent an important class of models for understanding how emotions unfold across time and have contributed significantly to advancing emotion recognition tasks.

1.7.3.5 Transformer

Transformers have emerged as powerful architectures for emotion recognition, particularly in multimodal contexts. Unlike sequential models like RNNs, Transformers rely on self-attention mechanisms that allow the model to weigh the importance of different parts of the input sequence,

regardless of their position. This makes them especially effective at capturing complex dependencies across time and modalities. In multimodal emotion recognition, attention mechanisms enable the model to dynamically focus on the most informative aspects of each modality, whether physiological, visual, or auditory, at every time step [30], [42]. This selective integration enhances interpretability and robustness, particularly in scenarios where the contribution of each modality may fluctuate due to contextual or sensor-related variability. While computationally intensive, Transformer-based models represent a growing direction in affective computing research.

1.7.4 Feature Engineering vs. Feature Learning

A fundamental distinction in machine learning pipelines, as well as in emotion recognition, lies in how input features are obtained from raw data. The two common approaches are feature engineering and feature learning.

Feature engineering refers to the manual extraction of features based on prior knowledge and domain expertise [43]. In the context of physiological signals, this typically involves deriving statistical, temporal, and frequency-domain features such as heart rate variability, skin conductance level or mean signal amplitude. These features are designed to reflect known physiological responses to emotional states, often grounded in biomedical literature and clinical practice. The effectiveness of classical machine learning algorithms such as SVM, RF, or KNN is heavily dependent on the quality of these engineered features. However, this approach requires extensive preprocessing, such as filtering, segmentation, artifact removal, and often fails to generalize across subjects and contexts due to inter-subject variability and hidden dependencies.

Feature learning, on the other hand, is a core strength of deep learning models. Rather than relying on manually handcrafted inputs, models such as CNNs or RNNs learn hierarchical feature representations directly from raw or minimally processed data during training [43]. This allows them to capture complex, nonlinear, and high-level abstractions that may be imperceptible or challenging to define through manual engineering. For physiological signals, this could mean learning intricate waveform patterns or subtle changes in dynamics over time that correlate with emotional states. Deep learning models can therefore scale more effectively to large datasets and adapt more robustly to varying input conditions.

Nonetheless, feature learning also comes with challenges. It requires significant amounts of labeled data, careful architectural choices, and greater computational resources. Moreover, interpretability is often lower compared to manually engineered features, which can be a drawback in applications like healthcare or psychology, where explainability and understanding the physiological basis of emotion are crucial.

In practice, many systems combine both strategies, using domain-informed preprocessing and handcrafted features alongside learned representations. This approach can balance interpretability with modeling power, offering better performance and flexibility.

While a direct comparison between feature engineering and feature learning would require their evaluation within the same model architecture, this thesis presents both approaches through separate pipelines, classical machine learning with handcrafted features and deep learning with feature learning, thereby offering an indirect yet informative perspective on their respective advantages in emotion recognition tasks.

1.7.5 Summary of ML models in ER

Classical and deep learning approaches each offer distinct advantages for emotion recognition, with classical models excelling in low-data scenarios and deep learning models better suited for complex, high-dimensional data. The choice of model should align with the nature and volume of available data, as well as the modalities involved. Ultimately, machine learning forms the computational backbone of modern emotion recognition systems, enabling multimodal fusion, generalization across users, and real-time inference.

1.8 Challenges

1.8.1 Inter-Subject and Intra-Subject Variability

One of the most common challenges in emotion recognition is the variability in how emotional states are physiologically expressed, both between individuals (inter-subject) and within the same individual over time (intra-subject). These differences complicate model generalization and reduce the ability to transfer learned features across users or sessions.

Inter-subject variability arises from differences in baseline physiology, emotional expression tendencies, age, sex, health conditions, and even cultural factors. For instance, a specific heart rate pattern may correspond to heightened Arousal in one person, but not in another. Intra-subject variability, on the other hand, includes temporal fluctuations in physiological responses due to fatigue, hydration, stress, circadian rhythms, or even sensor placement. This leads to inconsistency in how the same emotional state manifests across time for the same individual [5], [10].

Moreover, the subjective nature of emotion ground-truth labeling further worsens these issues. The same physiological signature might be labelled differently by different individuals, or even by the same individual under different contexts. This makes the training of generalized emotion recognition models especially difficult, pushing the field toward personalization techniques or larger, more diverse training datasets [5], [10].

In multimodal setups, variability also affects how each modality contributes to the overall prediction. Some individuals may express affect more strongly through speech, others through physiology, further highlighting the importance of flexible fusion strategies that can adapt to user-specific expression patterns.

1.8.2 Signal Quality and Sensor Limitations

Physiological signals, such as PPG, EDA, and skin temperature, offer unique insight into autonomic emotional responses, but their effectiveness heavily depends on signal quality. These signals are highly susceptible to motion artifacts, environmental noise, and sensor placement inconsistencies [10], [21], [26]. For example, PPG signals can become unreadable during hand movements, while EDA measurements are influenced by skin hydration, ambient temperature, or electrode pressure. Even small deviations in sensor alignment or contact pressure can produce significant variations in signal amplitude and morphology.

Wearable devices, while enabling real-world deployment and passive emotion monitoring, often sacrifice precision compared to laboratory-grade equipment. This trade-off introduces variability in signal integrity across devices, users, and contexts. Moreover, the limited sampling rates and resolution

of many consumer-grade sensors can constrain the extraction of subtle temporal or frequency-domain features essential for emotion recognition.

Preprocessing steps, such as filtering, detrending, or artifact correction, are thus essential to recover meaningful information from noisy inputs [23]. However, these steps can be difficult to integrate, particularly in real-time applications, and may unintentionally distort genuine physiological patterns if poorly tuned.

Collectively, these limitations challenge both classical and deep learning models, as the quality of input signals directly affects feature extraction and, as a result, model performance. These issues also reinforce the need for robust sensor calibration and fusion strategies that account for signal quality in multimodal systems.

1.8.3 Data Scarcity and Label Ambiguity

One of the most persistent challenges in emotion recognition is the limited availability of high-quality, labeled datasets, particularly for physiological signals. Emotion data acquisition is inherently resource-intensive, often requiring controlled setups, ethical approvals, and specialized hardware. This limits both the scale and diversity of datasets available for training machine learning models.

Moreover, emotions are highly subjective experiences, which makes labeling inherently ambiguous. Ground-truth labels are typically derived from self-reports, a method that often introduces noise. Self-reports may suffer from recall bias, emotional suppression, or social desirability effects [3].

Additionally, the emotional labels are often discrete, covering only extreme affective states, such as anger, fear, happiness, which limits the model's ability to learn complex, and nuanced or blended emotions like love, mild frustration or nostalgia. Thus, label sparsity and ambiguity can significantly degrade model performance, especially when trying to map complex physiological patterns to single emotion tags [2], [3].

The imbalanced distribution of resulting emotional classes further complicates model training. This can bias classifiers toward dominant classes and suppress sensitivity to less frequent but equally important emotions.

Together, these limitations make it difficult to train deep learning models effectively, which require lots of data by nature. Addressing these issues also require deeper understanding of human emotion from a neuroscientific and psychological perspective [3], [4]. In parallel, advances in wearable sensors may help capture more diverse, realistic, and high-resolution emotional responses, ultimately reducing ambiguity and improving generalizability.

1.8.4 Real-Time and Resource Constraints

Implementing emotion recognition in real-time systems, such as wearable devices or in HCI, imposes strict constraints on latency, computational load, memory usage, and energy efficiency. Deep learning models, while powerful, often involve significant inference expense, making them less suitable for deployment on edge devices without optimization [32]. Moreover, real-time applications require streaming data processing, where signals must be correctly segmented to match model input sizes and analysed incrementally with minimal delay, often under conditions of incomplete or noisy data. Achieving these demands lightweight architectures, efficient preprocessing pipelines, and potentially model compression techniques. The trade-off between model complexity and responsiveness remains a central design consideration for real-world affective computing systems.

1.8.5 Ethical and Privacy Considerations

Emotion recognition systems in general raise important ethical and privacy concerns. Emotions are inherently personal and sensitive, and the act of monitoring them, especially through involuntary or unconscious responses, can feel intrusive to users [3], [5]. Issues include informed consent, data ownership, potential misuse of affective data in surveillance or manipulation, and the risk of emotional profiling. These concerns are amplified in real-time applications, such as wearables or smart environments, where continuous monitoring may occur without explicit user awareness. As such, ethical design, transparency, and strict data governance are essential for the responsible deployment of ER technologies.

1.8.6 Multimodal Integration Challenges

While multimodal fusion enhances robustness and accuracy in emotion recognition, it introduces unique challenges. Integrating diverse data streams, each with different sampling rates, noise artifacts, and temporal alignments, requires precise synchronization and preprocessing strategies. Time alignment is particularly difficult when modalities exhibit asynchronous responses to emotional stimuli, such as delayed physiological reactions relative to facial expressions or speech. Moreover, real-world systems must account for missing or corrupted modalities at runtime, requiring dynamic fusion methods that can adaptively reweight or compensate for unavailable data without degrading performance [28]. Additionally, designing systems that generalize across users and contexts becomes more complex as modality combinations grow. Finally, computational cost increases with the number of inputs, posing practical constraints for real-time and embedded applications. These factors make multimodal integration both a powerful asset and a significant engineering puzzle.

1.9 Summary of Literature Review

This literature review has examined the theoretical foundations and current practices in emotion recognition, emphasizing the value of physiological signals and multimodal fusion. Emotions are complex, multifaceted phenomena that manifest across behavioral, physiological, and expressive domains. While many approaches rely on observable cues such as facial expressions or speech, these can be consciously masked or biased by context. In contrast, physiological signals, such as photoplethysmography (PPG), electrodermal activity (EDA), and skin temperature, are closely tied to the autonomic nervous system and thus offer a more objective, unconscious window into affective states. Their continuous, low-latency nature makes them particularly well-suited for wearable and real-time applications.

Moreover, given that humans perceive and interpret emotions multimodally, fusing information across modalities is a natural and necessary step toward reliable and generalizable ER systems. The integration of physiological signals, in particular, holds great promise due to their complementarity and minimal invasiveness. Despite the challenges posed by variability, signal quality, and synchronization, multimodal fusion has the potential to compensate for the limitations of unimodal approaches and enhance robustness in dynamic environments. The central hypothesis of this thesis is that a well-designed fusion architecture built on physiological modalities will outperform unimodal baselines, offering both accuracy and practical viability for emotion-aware systems.

Chapter 2: Methodology

2.1 Introduction to Methodology

This chapter puts emphasis on the methodological framework that was applied to evaluate and compare different machine learning approaches for emotion recognition using physiological data. The primary goal is to design and evaluate a multimodal emotion recognition system, and investigate whether multimodal integration of signals, specifically through feature-level fusion, enhances classification performance over unimodal approaches, and to assess the benefits of classical machine learning versus deep learning models in this context. Furthermore, this approach utilized physiological signals, which are objective, non-invasive, and suitable for real-time systems, particularly when coupled with multimodal fusion that may improve performance.

The methods that were used fall into one of these broad categories: signal preprocessing, feature extraction and engineering, label encoding, temporal and dimensionality coherence checks, model architectural design and hyperparameter tuning, training, evaluation and comparison. Each stage is carefully designed to respect the physiological characteristics of each modality, while maximizing the capacity of the resulting models.

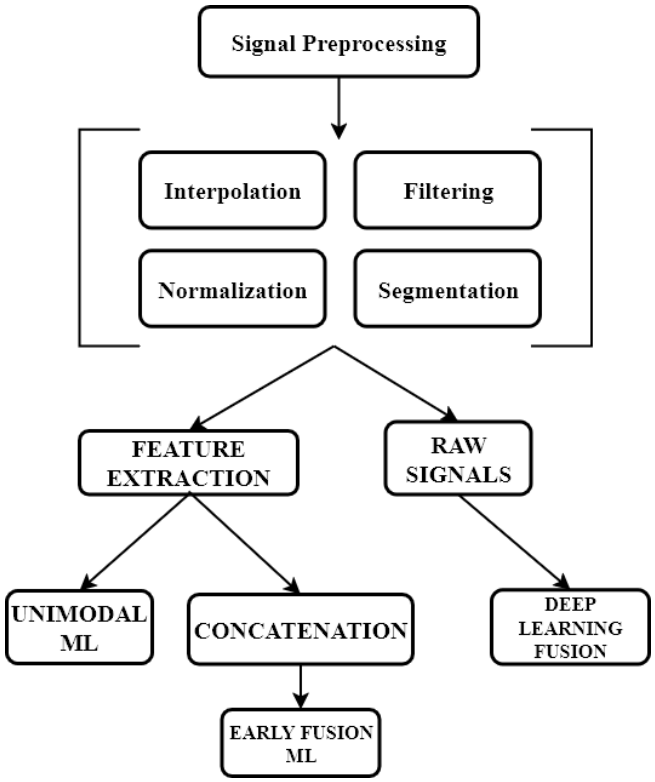

Figure 2.1: A schematic diagram of the full pipeline.

The dataset used in this work was kindly provided to the author by CERTH and included synchronized recordings of five physiological modalities: photoplethysmography in three wavelengths (PPG Green, PPG Infrared, PPG Red), electrodermal activity (EDA), and skin temperature. Along with the data, the ground truth responses of Valence and Arousal from the conducted data acquisition experiment were provided. Each modality underwent preprocessing tailored to its signal characteristics, including EDA decomposition, and noise filtering, in order to produce clean inputs for feature engineering and learning.

Signal normalization in particular was conducted in a per-modality, per-subject manner, an approach that had the goal to mitigate inter-subject variability.

The raw response values of Valence and Arousal were encoded into six discrete classes, derived from their distribution within the Valence-Arousal space. In addition to the four primary quadrants typically used in dimensional emotion models, two intermediate classes were included to capture samples that exhibited pronounced activation in only one dimension. The data was split into training and test sets using a stratified strategy, ensuring balanced class distributions across training and test sets. The same data splits were preserved across all modeling approaches to allow for direct and fair comparison, using seeds.

Besides the unimodal ML baseline for reference, two fusion modeling strategies were implemented. The first involved classical machine learning algorithms trained on manually extracted and engineered features from each modality. Feature extraction was modality-specific and based on established physiological markers such as heart rate variability, skin conductance level, and temperature derivatives. An early fusion (concatenation) strategy was used to combine the features from all five modalities into a single input set for classification. This early integration was preferred over decision-level fusion, as it enables the model to learn cross-modality interactions that may contribute to more nuanced emotion representations.

The second strategy employed a 1D-CNN-based model trained end-to-end on the preprocessed signals. This deep learning approach allowed the model to learn temporal features directly from raw data, reducing dependence on manual feature engineering while leveraging the generalization power of convolutional operations.

Three model types were evaluated:

- Unimodal models trained separately for each physiological signal,
- Feature-level multimodal models using classical algorithms and handcrafted features,
- A deep learning end-to-end multimodal model.

Model performance was assessed using accuracy as the primary metric, chosen for its balanced reflection of classification success in an emotion classification task. Accuracy is defined as the ratio of correctly predicted samples to the total number of predictions. To ensure robustness and reproducibility, all models were evaluated and averaged across five randomly selected seeds (7, 11, 21, 35, 42). These seeds controlled all sources of randomness, including data splits and model initialization, and were generated independently from the implementation code. Due to the computational demands of training deep learning models, and because different seeds ensured different splits, cross-validation was substituted with seed-based repeated runs, and this methodology was applied uniformly to classical models for consistency and fair comparison. Finally, among the five seeded splits, the best-performing seed was selected based on performance, and the model was subsequently trained twice using the original 7-class labels, once for Valence and once for Arousal classification.

$$Accuracy = \frac{Correct\ predictions}{Total\ predictions} \tag{2.1}$$

2.2 Dataset

The dataset used in this thesis was provided by the Centre for Research and Technology Hellas (CERTH) and originates from an emotion recognition experiment designed to record physiological responses

corresponding to distinct emotional states. It contains multimodal biosignal data acquired using a single wearable device, the EmotiBit [44], which integrates multiple sensors into a compact, non-invasive form factor. The use of a wearable system rather than laboratory-grade equipment aligns with the broader aim of developing real-time, real-world emotion recognition systems.

The recorded physiological modalities include:

- Photoplethysmography (PPG) in three wavelengths (green, infrared, and red), sampled at 25 Hz,
- Electrodermal Activity (EDA) at 15 Hz,
- Skin Temperature at 7.5 Hz.

In total, 25 subjects participated in the experiment. Each participant was exposed to the same set of 16 videos, presented in a fixed sequence. These audiovisual stimuli were curated to evoke a broad range of emotional states, each lasting approximately 2 - 3 minutes. Following each video, participants reported their perceived emotional state using two self-assessment scores: Valence and Arousal. Both were rated as integers on a discrete [-3, 3] scale, resulting in seven possible values per axis.

These values were later discretized into six classes, based on their position in the Valence-Arousal dimensional model. The classification scheme preserved the four quadrants typically defined in affective computing research, while also accommodating additional classes for responses lying along the axes (namely, “pure Valence” or “pure Arousal”), to account for the occurrence of responses with neutral intensity on one axis but expression on the other. This expanded label set better reflects the subjectivity and continuous nature of emotional experiences, while maintaining interpretability for classification tasks.

Each participant’s data was stored in an individual JSON file. The structure followed a nested format: each file contained a list of 16 entries, corresponding to the video stimuli. For each video, a set of sensor readings was provided per modality (PPG Green, PPG IR, PPG Red, EDA, and temperature), each accompanied by its own timestamp list. This hierarchical structure required parsing and reorganization to align modalities temporally and associate them with the corresponding ground-truth labels during preprocessing.

The dataset is fully anonymized. While the data acquisition was not performed by the author of this thesis, it is understood that ethical considerations, including participant consent and privacy protection, were addressed by the original investigators.

2.3 Signal Processing

2.3.1 General Pipeline

The preprocessing pipeline was designed with a strict quality control mindset, prioritizing signal reliability over data quantity. Given the relatively small dataset size, no attempts were made to impute or repair corrupted signals. Instead, entire segments or recordings were discarded when quality criteria were not met. This ensured that the data retained for further analysis reflected high-confidence physiological responses, ultimately strengthening the validity of the classification results. This principle guided both preprocessing and feature extraction decisions.

Data was loaded in a nested loop structure, first iterating over subjects, then over each of the 16 video stimuli per subject. The pipeline applied the following steps for each recording:

2.3.1.1 Empty Segment Check

Videos with no recorded data were immediately excluded. This was necessary, as some files were partially corrupted due to device malfunction or dropout during recording.

2.3.1.2 Flattening and Timestamp Handling

For each signal within a video, measurements were flattened from nested JSON structures. It was frequently observed that the number of signal samples exceeded the number of timestamps, commonly at a ratio of about 3:1. This mismatch suggested two possible explanations:

- The sensor may have buffered multiple readings under a single timestamp.
- Timestamps were missing or dropped, likely due to hardware limitations.

Two strategies were tested: Averaging repeated values per timestamp, and interpolating missing timestamps using linear interpolation.

The former yielded an effective sampling rate of ~8 Hz, significantly below the documented rate for the device. In contrast, linear interpolation restored the sampling rate to approximately 25 Hz for PPG signals, aligning with EmotiBit's specifications [45]. Based on this evidence, interpolation of timestamps was adopted as the correct reconstruction method.

2.3.1.3 Sampling Rate Verification

After interpolation, the actual sampling rate was calculated and compared against expected values, which were 25 Hz for PPG, 15 Hz for EDA, and 7.5 Hz for temperature. Segments with deviations exceeding 10% of the expected rate were flagged as irregular and discarded.

2.3.1.4 Edge Trimming

To avoid potential artifacts at signal boundaries, the first and last 5 seconds of each recording were removed. This accounted for both physiological latencies, such as delayed sympathetic response, and practical transitions, such as video onset and user attention shifts.

2.3.2 PPG Processing

PPG data were collected across three optical channels, in green, infrared, and red wavelengths, each of which underwent identical preprocessing. The preprocessing pipeline for PPG signals was designed to restore signal orientation, remove noise, and eliminate baseline wander, ensuring high-quality inputs for further analysis.

2.3.2.1 Signal Orientation Correction

Raw PPG signals, as exported from the recording device, were inverted along the y-axis, showing troughs instead of peaks. This inversion was corrected by multiplying the signal by -1.

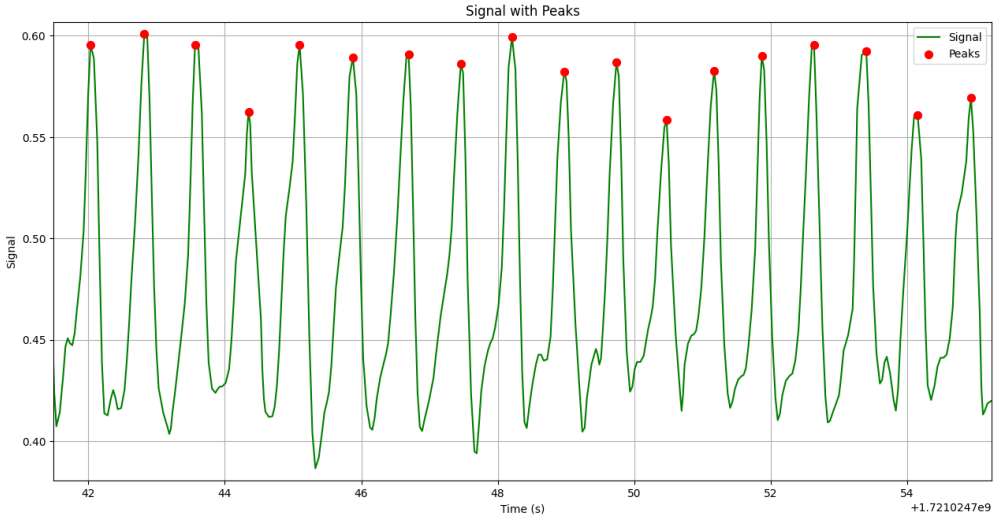

Figure 2.2: A raw PPG signal that is flipped on the y-axis.

2.3.2.2 Bandpass Filtering

A third-order Butterworth bandpass filter was applied to isolate the frequency range containing physiologically relevant cardiovascular features. The selected cutoff frequencies were 0.7 Hz for the low cut, and 4.6 Hz for the high cut.

These values were determined based on two criteria:

- Empirical analysis using Welch power spectral density (PSD) plots [46] across multiple signals, which consistently showed cardiac activity and harmonic content within this range.
- Consistency with literature, which typically reports useful PPG frequency content between 0.5 Hz and 3.5-5 Hz for pulse wave analysis and related metrics[25], [26], [47 - 52].

The slightly conservative upper limit of 4.6 Hz was retained to avoid eliminating potentially informative harmonics.

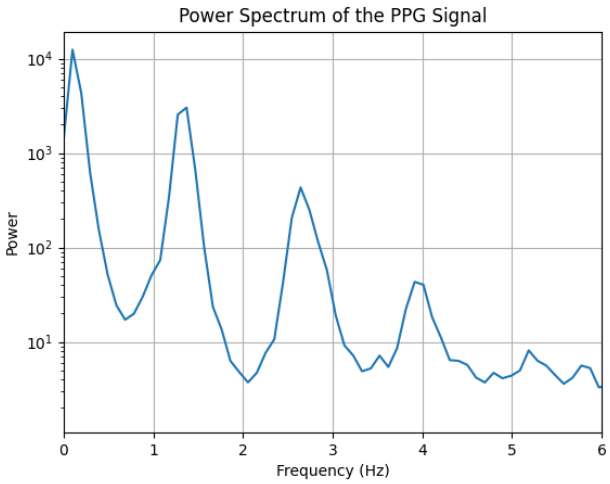

Figure 2.3: A PSD plot using Welch's method.

2.3.2.3 Baseline Wander and Trend Detection

To confirm the effectiveness of the filter in removing baseline drift, two mathematical checks were implemented:

- Linear Trend Detection: The slope of a linear regression (using SciPy [53]) was computed across the signal. Slopes smaller than 0.001 were considered effectively trend-free.
- Baseline Wander Detection: A 5-second moving average was applied to capture slow oscillatory components. The standard deviation of this baseline was then compared to the standard deviation of the original signal. A ratio greater than 0.1 was taken as an indication of significant low-frequency wander.

Both thresholds were empirically selected and consistently indicated that the applied bandpass filter was sufficient for removing unwanted drift. Visual inspection supported these findings across multiple subjects and signals.

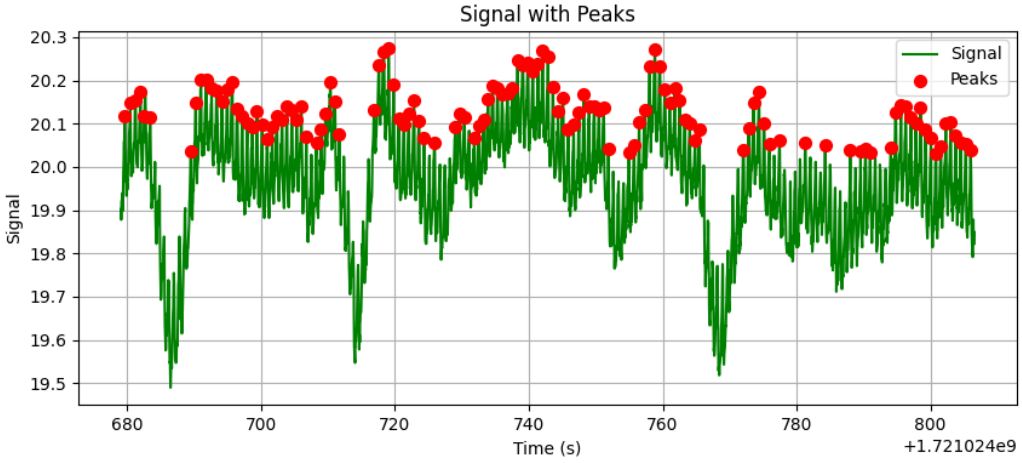

Figure 2.4: Unfiltered PPG signal.

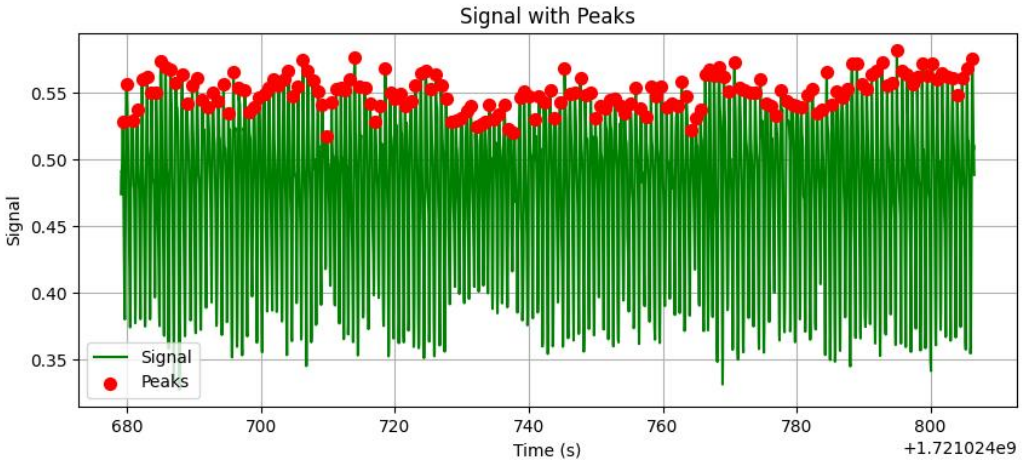

Figure 2.5: Filtered PPG signal.

2.3.2.4 Modality-Specific Note

These baseline correction tests were conducted only on PPG, as this modality is particularly susceptible to respiration-induced drift and motion artifacts [26], [47], [48], [50], [51], [54], [55].

EDA signals, by contrast, have baseline shifts, commonly called Skin Conductance Level, that are physiologically meaningful and must not be removed as artifacts [7], [10], [27], [56].

Skin temperature signals showed no measurable baseline trends or noise artifacts, and were not subjected to detrending procedures.

2.3.3 EDA Processing

2.3.3.1 Low-Pass Filtering

Electrodermal activity signals were processed with the aim of removing high-frequency noise and decomposing the signal into its meaningful components. A third-order low-pass Butterworth filter with a cutoff frequency of 1 Hz was applied to denoise the raw EDA signal. This value was chosen based on both physiological literature [57] and empirical verification, as it preserves the slow dynamics of skin conductance while attenuating sensor noise and movement artifacts.

2.3.3.2 Signal Dropout Handling

EDA was found to be particularly susceptible to signal dropouts, likely due to its dependency on continuous and stable skin contact. Visual and statistical inspection revealed that some EDA segments were nearly flat, often with extreme constant values. These manifested as either abnormally high conductance levels, suggestive of excessive skin moisture, such as wet hands, or near-zero conductance, indicative of poor or lost contact between sensor and skin. Such flat or saturated segments were deemed unreliable and were systematically discarded, along with the corresponding segments from all other modalities for that video trial, to maintain multimodal consistency.

2.3.3.3 Decomposition

After filtering, normalization, and quality control, the EDA signal was subjected to physiological decomposition into tonic (SCL) and phasic (SCR) components. Tonic activity was extracted using another low-pass Butterworth filter, again of third order, with a cutoff frequency of 0.05 Hz. This value was selected empirically to isolate the slow-varying baseline, and is slightly more conservative than values found in literature [27] to better suppress slow phasic contamination. The phasic component was computed by subtracting the tonic signal from the filtered EDA, thus isolating the faster, event-related conductance fluctuations.

The effectiveness of this decomposition approach was validated through visual inspection and comparison across trials. Representative examples of raw and decomposed EDA signals are illustrated in Figures 1.3, 1.4, 1.5 (presented previously), demonstrating the separation of baseline trends from short-term activity.

2.3.4 THERM Processing

The skin temperature (THERM) signal, exhibited high stability and minimal susceptibility to noise or dropout artifacts throughout the dataset. This robustness is likely attributed to the simplicity and reliability of temperature sensors in wearable settings, which typically operate independently of skin contact impedance and are less sensitive to motion-induced artifacts. Consequently, no additional preprocessing or filtering was applied to the THERM signal. The raw data was used as-is for normalization and feature extraction, ensuring that the signal retained its full temporal and thermal resolution without unnecessary transformation.

2.3.5 Normalization

2.3.5.1 Normalization Rationale

Normalization was a necessary step in the preprocessing pipeline. Many features of interest, particularly amplitude-based features, depend on the absolute values of the signal, rather than purely on temporal or frequency-domain characteristics. Thus, normalization was required to ensure comparability across different modalities and subjects. Without normalization, signals with inherently different units, ranges, or scaling could dominate or distort learning, particularly in multimodal fusion settings where balanced feature contribution is critical [58].

2.3.5.2 Min-Max Normalization

To address these needs, a Min-Max normalization strategy was implemented manually for all signals. While Min-Max normalization is known to be sensitive to outliers, it was intentionally preferred over alternatives like z-score normalization. The reason lies in the fact that z-score scaling transforms data based on mean and standard deviation, which may distort physiologically meaningful amplitude relationships, such as stress-induced intensity changes, or rise and decay times in PPG. In contrast, Min-Max scaling preserves the true relative amplitude distribution, making it more faithful to the original physiological signal characteristics, all while scaling to the specific [0, 1] range [58].

Importantly, Min-Max normalization was not applied individually per signal segment. Had that been the case, all signals would be stretched to fit the [0, 1] range, leading to the loss of important morphological and amplitude distinctions, and potentially misleading the model. Instead, the normalization was conducted per subject and per modality, across all available videos for that subject. This decision addressed two critical sources of variation, inter-modality and inter-subject variability.

2.3.5.3 Inter-Modality and Inter-Subject Variability

First, physiological modalities differ significantly in their units, dynamic ranges, and signal morphology. For example, EDA is measured in microsiemens, PPG in voltage, and skin temperature in degrees Celsius. As a result, it is necessary to normalize each modality independently to avoid unequal scaling and feature distortion during fusion.

Second, inter-subject variability can significantly affect general physiological baselines. For instance, a lower-amplitude PPG waveform in one subject may reflect a naturally subdued cardiovascular response, not an absence of emotional Arousal. Furthermore, individual differences in skin type affect optical signal absorption and quality in PPG, thus contributing to natural amplitude variability between subjects [48]. Similar sources of variability exist in other modalities as well. For example, some individuals may have more reactive or sensitive sweat glands, resulting in higher baseline or exaggerated electrodermal activity, even under moderate stimuli. Furthermore, skin temperature baselines can differ due to individual metabolic rates, peripheral circulation, or hormonal influences. These physiological traits are not always indicative of emotional intensity, but they do impact the raw signal ranges observed [5], [10].

2.3.5.4 Normalization Method

To account for these, the minimum and maximum values used for normalization were derived independently for each subject and modality, but across all trials (videos), under the assumption that each video stimuli constituted a stable reference, inducing a certain emotional state, across subjects.

This approach ensured that physiological responses were normalized in a way that respected individual baselines while retaining inter-trial dynamics, leading to more "natural" and meaningful signal profiles.

It also supports fairer cross-subject comparison during model training, especially under a multimodal fusion framework.

2.3.6 Segmentation

After normalization, each signal was subjected to a segmentation procedure designed to extract temporally localized information suitable for classification. A fixed window length of 30 seconds was used, with 50% overlap between sequential segments. This configuration balances temporal resolution with statistical robustness, ensuring enough signal data per window while increasing the number of available samples.

Each segment was then forwarded to modality-specific feature extraction pipelines. These pipelines returned structured feature sets, along with their labels, corresponding to each window.

2.3.7 Quality Control and Dimensionality Checks

To maximize data quality and ensure that only physiologically valid and computationally usable segments were retained, a strict quality control protocol was employed. Segments were discarded entirely if they exhibited any of the following issues:

- Contained NaN (Not a Number) or Inf (Infinity) values after feature extraction.
- Returned empty feature sets, likely due to excessively short or flat signals.
- Caused runtime errors during feature extraction, most often due to corrupted morphology. A common example involved PPG signals with too few valid peaks to compute inter-beat intervals (IBIs).

In any such case, not only was the faulty segment discarded, but the entire corresponding trial (video) was also removed from all modalities to preserve alignment, and ensure fusion compatibility and label matching. This quality-first approach prioritized the physiological validity of the dataset over raw sample size.

After segment-level quality control, it was verified that:

- All modalities had the same number of segments per trial.
- Each modality had the same number of resulting features across all trials.
- Each raw signal segment had identical temporal duration across modalities.
- If signal lengths differed, they were truncated from the end, maintaining synchronization.

This ensured temporal and dimensional alignment of signals and features across modalities, a critical requirement for multimodal fusion.

2.3.8 Data Storage and Output Structure

Two parallel storage structures were maintained:

- Processed raw signal segments, post-normalization and quality assessment, saved modality-wise.
- Corresponding extracted feature sets from each segment, also saved modality-wise.

This separation ensured compatibility between classical machine learning and deep learning pipelines, and enabled fair comparison across models.

Across all output file types, the number of samples was constant (2083), and sample order was preserved. Each sample was associated with two labels, Valence and Arousal, as self-reported by the subjects after each trial. All files were saved in the NumPy [59] format, chosen for its fast performance and compatibility with Python-based machine learning workflows. This consistent structuring and formatting guaranteed full alignment across modalities, segment length, and sample order, supporting fair and reproducible comparisons across a variety of strategies, including unimodal baseline approaches, early fusion, and Deep Learning.

Three categories of output files were produced:

2.3.8.1 Raw Signal Files (per modality)

Each file (total five files) contained vertically stacked samples (segments) from a single modality:

- PPG (all wavelengths): 693 time points per sample + 2 labels
- EDA: 415 time points per sample + 2 labels
- THERM: 202 time points per sample + 2 labels

2.3.8.2 Manual Feature Files (per modality)

Feature sets (total five files) extracted from the above raw signals:

- Each PPG channel: 31 features + 2 labels
- EDA: 29 features + 2 labels
- THERM: 10 features + 2 labels

2.3.8.3 Combined Feature-Level File

All manually extracted features concatenated (one file) across modalities, aligned per segment:

- Total: 132 features + 2 labels

2.4 Feature Extraction

This section details the manual feature extraction and engineering process applied to each modality, which formed the foundation for the early fusion method and unimodal classical machine learning baselines. It is important to note that the features described here are distinct from the learned representations used in the deep learning pipeline, which rely on end-to-end learning from raw signals. Instead, the present approach focuses on extracting interpretable, engineered features from each 30-second segment (as described in the previous section), using domain-specific signal processing methods tailored to the physiological characteristics of each modality. Separate feature extraction procedures were implemented for PPG, EDA, and THERM signals, using handcrafted mathematical methods, and are presented in the following subsections.

2.4.1 PPG Feature Extraction

This subsection outlines the manual feature extraction techniques applied to PPG signals across all wavelengths. The goal was to derive a robust set of time-domain, frequency-domain, and morphology-based features from each 30-second segment, enabling comparisons in classical machine learning pipelines.

2.4.1.1 Peak Detection

To extract cardiovascular dynamics, a custom peak detection algorithm was developed. This method defined a dynamic threshold set at:

$$Threshold = mean + 0.5 * std \quad (2.2)$$

Peaks were defined as points that exceeded this threshold and were the local maximum between the previous and next point. They were subsequently subject to a minimum inter-peak distance of 0.4 seconds (150 BPM upper bound) from the previous peak, to reject physiologically unrealistic detections. In cases of closely spaced peaks, the higher of the two was retained. This method proved effective across PPG morphology and was also later reused for detecting phasic peaks in EDA.

Following peak detection, inter-beat intervals were computed as the time difference between successive peaks. These intervals formed the basis for both heart rate and heart rate variability feature extraction.

2.4.1.2 Time-Domain Features

Time-domain features computed from the IBIs included [10], [26], [27], [47], [55], [56], [60 - 62]:

- Avg_BPM and SD_BPM: Mean and standard deviation of heart rate.
- Mean_IBI, Median_IBI, Skew_IBI, Kurtosis_IBI: Statistics of the IBIs.
- RMSSD, pNN50, and SDNN: Standard HRV metrics reflecting beat-to-beat variability.

2.4.1.3 Nonlinear HRV Features

Additional features were derived from Poincaré metrics [10], [63]:

- SD1 and SD2: Short-term and long-term variability indices.
- SD1_SD2 (ratio).

2.4.1.4 Frequency-Domain Features

Spectral analysis of the IBIs was performed using Welch's method to estimate power in standard HRV bands [10], [26], [27], [47], [55], [60], [61]:

- LF (Power band: 0.04 - 0.15 Hz) and HF (Power band: 0.15 - 0.4 Hz).
- LF_HF (ratio): An indicator of cardiovascular balance.
- LF_Norm, HF_Norm, and Total_Power were also calculated for additional detail.

2.4.1.5 Morphological and Pulse Metrics

Pulse morphology was assessed using amplitude-based and shape-related metrics [10], [25], [47], [48], [50], [51], [61]:

- Mean_Amp, SD_Amp, Median_Amp, MAD_Amp: Statistics of peak amplitudes.
- Skew_Amp, Kurtosis_Amp.
- AC_Component, DC_Component, AC_DC (ratio): Features characterizing signal dynamics relative to baseline.
- Rise_Mean and Decay_Mean: Average time from DC component to peak and peak to DC component of each pulse, respectively.
- Rise_Decay (ratio) and AUC (Area Under Curve): Capturing waveform symmetry and area under the curve.

2.4.1.6 Summary of PPG Features

The above features were computed for each segment independently, resulting in a total of 31 features per segmented PPG signal. These features aimed to capture meaningful cardiovascular dynamics relevant to emotion-driven physiological responses.

2.4.2 EDA Feature Extraction

For electrodermal activity, manual feature extraction was applied separately to the tonic and phasic components of the 30-second segments. These components were previously preprocessed and decomposed as described in earlier sections.

The extracted features were grouped into three primary domains, tonic (SCL), phasic (SCR), and frequency-domain features derived from the phasic signal.

2.4.2.1 Tonic Component Features

Derived from the low-frequency baseline of the EDA signal, these features include [10], [56], [57], [60]:

- SCL_Mean, SD_SCL, SCL_Min, SCL_Max, SCL_Range: Basic statistics.
- SCL_Slope: Linear trend.
- SCL_MeanDer, SCL_MaxDer: First derivatives.
- SCL_AUC: Area under the curve.

2.4.2.2 Phasic Component Features

These features describe the shape and variability of the high-frequency, event-related SCR waveform [10], [56], [57], [60]:

- SCR_SD, SCR_Min, SCR_Max, SCR_Range: Basic statistics.
- SCR_MeanDer, SCR_MaxDer: First derivatives.
- SCR_AUC: Area under the curve.
- SCR_Peaks, SCR_PeakRate, SCR_PeakMeanAmp: Peak count, rate, and amplitude.
- SCR_Zero, SCR_ZeroRate: Zero-crossing metrics.
- SCR_Skew, SCR_Kurtosis: Distribution of the shape of the phasic signal.

2.4.2.3 Frequency-Domain Features

Using Welch's method to estimate spectral power from the phasic signal [56], [64]:

- LF (Power band: 0.05 - 0.25 Hz), HF (Power band: 0.25 - 0.5 Hz), LF_HF (ratio): Low and high frequency power and their ratio.
- LF_Norm, HF_Norm: Normalized band power.
- Total_Power: Full-band power.

2.4.2.4 Summary of EDA Features

The above features were computed independently for each EDA segment, yielding a total of 29 features per window. These features were designed to characterize both baseline GSR and event-related skin conductance responses, capturing dynamics relevant to autonomic Arousal and emotion-related activity.

2.4.3 THERM Feature Extraction

Compared to PPG and EDA, the THERM signal encodes less direct physiological information related to emotional Arousal or autonomic reactivity. As such, only a limited set of features was extracted per 30-second segment, focusing on general signal characterization rather than complex feature engineering.

- Therm_Mean, Therm_Median, SD_Therm, Therm_Min, Therm_Max, and Therm_Range: Statistical metrics that capture the variability and overall dynamic range of the temperature signal.
- Therm_MeanDer and Therm_MaxDer: First derivatives.
- Therm_Skew and Therm_Kurtosis: Distribution of the shape of the signal.

These computations resulted in a total of 10 features per skin temperature segment. While thermoregulatory responses typically lack the physiological depth of cardiovascular or electrodermal dynamics, these features were retained as complementary information reflecting sympathetic activity and stress-related vasoconstriction patterns, especially when integrated in a multimodal fusion framework.

2.4.4 Summary of Feature Extraction

In total, 31 features were extracted from each distinct-wavelength PPG segment, 29 from EDA, and 10 from THERM, resulting in a concatenated feature set of 132 features per segment. These manually crafted features were designed to capture domain-specific physiological information across cardiovascular, electrodermal, and thermoregulatory systems. They formed the input to all classical machine learning models in the early fusion and unimodal pipelines.

2.5 Data Labeling and Emotion Annotation

2.5.1 Emotion Representation Model

Emotions in this thesis were modeled using the Valence-Arousal dimensional framework, a widely adopted approach in affective computing. This model positions emotional states within a two-dimensional continuous space, where:

- Valence reflects the degree of pleasure or displeasure, representing how positive or negative an emotion is,
- Arousal represents the level of physiological activation or alertness.

This approach offers several advantages over categorical emotion models, as it captures a richer and more continuous range of affective experiences, accommodates the ambiguity of real-world emotion labeling, and aligns more naturally with nuanced physiological responses that may not correspond cleanly to discrete categories.

The Valence-Arousal space used in this thesis is illustrated in Figure 1.1, where common emotional regions are mapped according to their relative Valence and Arousal scores. This framework serves as the basis for both the original annotation and the modeling of affect in the thesis.

By adopting this dimensional model, each label could be annotated by binning these two emotional values, Valence and Arousal, forming the target outputs for machine learning tasks.

2.5.2 Label Binning Strategy

Emotion annotations in the dataset were provided via self-assessment, with participants rating their affective experience following each video trial. For every trial, subjects reported two integer values, one for Valence and one for Arousal, each drawn from a 7-point scale ranging from [-3, 3]. These original annotations formed the basis for emotion labeling.

To convert this data into a format suitable for classification, a label binning strategy was implemented, inspired by the geometric structure of the Valence-Arousal model. Specifically, the two-dimensional values were discretized into six classes based on the emotional quadrant they occupied in the affective space. Four classes represent the classical quadrants of the circumplex model, while two additional classes capture axis-aligned affective states, specifically high or low Arousal with neutral Valence, and

high or low Valence with neutral Arousal. These axis states are affectively meaningful and were thus explicitly retained.

- Label 0 (HVHA): V > 0, A > 0,
- Label 1 (LVHA): V < 0, A > 0,
- Label 2 (LVLA): V < 0, A < 0,
- Label 3 (HVLA): V > 0, A < 0,
- Label 4 (Neutral Valence, High/Low Arousal): V = 0, A != 0,
- Label 5 (High/Low Valence, Neutral Arousal): V != 0, A = 0.

A theoretical seventh class (V = 0, A = 0), representing complete neutrality, was found to contain only a single sample and was therefore excluded from the dataset entirely. This filtering was performed programmatically prior to training and evaluation.

Finally, for comparative purposes, the best-performing seed of the CNN-based model was also trained on the original 7-point scales, using two separate classification tasks, one for Valence and one for Arousal. This allowed direct evaluation of the model's ability to learn affective distinctions beyond the categorical abstraction used in the 6-class labeling strategy.

2.5.3 Class Distribution

After binning, the final class distribution across the full dataset of 2083 samples was:

- Label 0 (HVHA): 33.1%
- Label 1 (LVHA): 38.0%
- Label 2 (LVLA): 3.8%
- Label 3 (HVLA): 12.5%
- Label 4 (Neutral Valence, High/Low Arousal): 7.1%
- Label 5 (High/Low Valence, Neutral Arousal): 5.5%

This distribution reflects the natural affective tendencies of the participant population and was intentionally left unbalanced. No class balancing techniques were applied, such as resampling or synthetic augmentation, in order to respect the population structure of the self-reported emotional data.

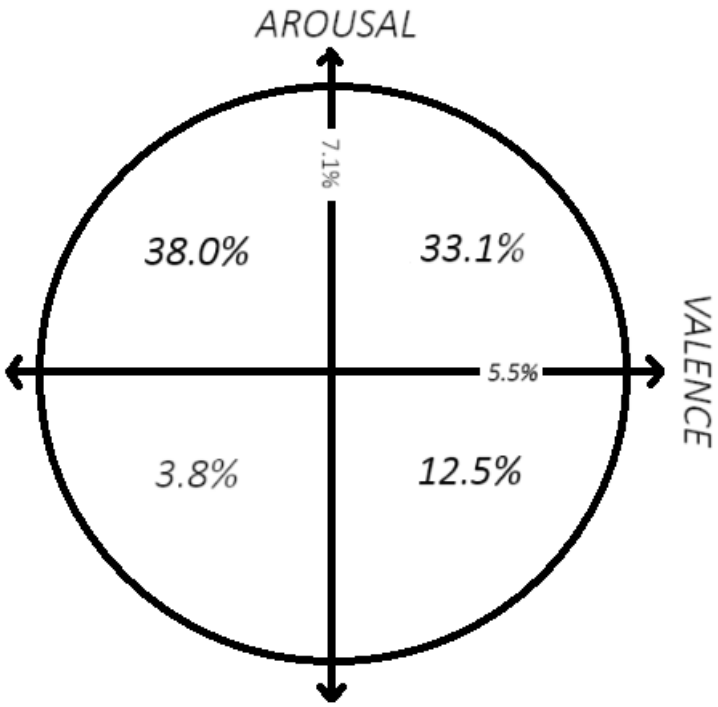

Figure 2.6: Class distribution percentages on the Valence-Arousal schema.

2.6 Modeling

2.6.1 Frameworks and Tools

All components of the pipeline were implemented in Python 3.11.9, using a combination of scientific computing and machine learning libraries tailored for signal processing and model development.

Internal signal manipulation and preprocessing operations, such as Butterworth filtering, were carried out using NumPy and SciPy libraries [53], [59]. Data type for the input files was also selected to be the NumPy file type. The handcrafted feature extraction routines were implemented using native functions, with the exception of HeartPy [65], which was used solely to estimate the sampling rate of signals. For data visualization, Matplotlib was employed to produce plots and schematic figures.

Classical machine learning models were developed using scikit-learn [66], which also provided functionality for stratified data splitting. Deep learning experiments were implemented in PyTorch [67], utilizing the CUDA framework for GPU acceleration [68]. These models were trained locally on a system equipped with an NVIDIA RTX 3060 GPU.

To ensure reproducibility and stability across different runs and modeling approaches, five distinct random seeds were selected and consistently applied across all configurations. Random seed initialization was handled appropriately for each framework. In the classical machine learning pipeline, seeds were set for Python's random module and for NumPy. For the deep learning models, seeds were additionally set for PyTorch and CUDA, and runtime determinism was enforced by disabling non-deterministic cuDNN operations. This seed-setting code is included in the Appendix for transparency.

This setup ensured consistency in data shuffling, train-test splits, model initialization, and training behavior across both classical and deep learning pipelines.

2.6.2 Data Splitting and Loading

To ensure fair and consistent evaluation across all setups, unimodal baselines, early fusion models, and the deep learning approach, the data was split in an identical manner across all runs and modeling strategies. This was considered essential for isolating the effects of the modeling approach and input representation, thereby enabling direct and valid performance comparisons.

Given the computational demands of training deep learning models, cross-validation was deemed infeasible. Instead, a fixed train-test split was used for all models. Despite the relatively limited dataset size (2083 samples), a 10% test set was consistently extracted and withheld from training. This proportion represented a practical compromise, maintaining sufficient data for model training, while still preserving a meaningful and representative hold-out set for evaluation.

In the deep learning pipeline, the remaining 90% training data was further split to assign a 5% validation set, equaling to 5% of the full dataset. This internal validation set was not used in training, but exclusively to inform mechanisms such as early stopping and learning rate scheduling.

For deep learning inputs, the NumPy arrays were converted to PyTorch tensors. A custom PyTorch Dataset class was implemented to handle multimodal sample alignment, and data loading was managed using PyTorch's DataLoader. The training set was shuffled, while validation and test sets were kept in order. All DataLoader instances used a batch size of 64 and had memory pinning enabled to improve GPU transfer efficiency.

Furthermore, a stratified method was used to split the dataset. This guaranteed that the class distribution was preserved across training, validation, and testing subsets, critical for avoiding performance artifacts.

Together with the strict enforcement of random seed initialization, these design choices ensured a fully deterministic and reproducible data handling pipeline. All models were trained and evaluated on identical sample subsets, eliminating variability in data exposure.

2.6.3 Machine Learning Approaches

2.6.3.1 Introduction to Machine Learning Approaches

In this section, the modeling strategies used for emotion classification are described, encompassing both traditional machine learning and deep learning approaches. These strategies were applied across different representations of the physiological data: unimodal features, early-fused features, and raw physiological signals. The classical ML approaches relied on handcrafted features derived from domain-specific knowledge, while the DL architecture leveraged feature learning from raw time-series inputs. All models were trained and evaluated on the same dataset partitions and under controlled random seed settings, ensuring consistent and fair comparisons.

The section first describes the unimodal ML baselines that rely on handcrafted features. This is followed by the early fusion classical ML models, which combine features across modalities. The final part details the deep learning architecture that learns representations directly from the aligned raw signals.

2.6.3.2 Unimodal Classical Models

To establish a comparative baseline, unimodal classification models were trained independently on each modality's handcrafted feature set. This setup served as a benchmark against which to evaluate the benefits of early fusion across modalities. For each unimodal case (PPG, EDA, and THERM), three classical machine learning algorithms were applied:

- Random Forest with 100 estimators,
- K-Nearest Neighbors with k = 5,
- Support Vector Classifier (SVC).

These models were selected for their widespread use in classification tasks. Basic hyperparameter configurations were used based on established practices, without extensive tuning, as the primary focus was on evaluating feature-level modality performance. Training and inference were performed separately for each model and each modality.

2.6.3.3 Early Fusion Classical Models

To evaluate the benefits of combining modalities at the feature level, early fusion models were developed using the same classical machine learning algorithms and hyperparameters as the unimodal baselines. This design ensured that the fusion strategy was the only varying factor between the two approaches, allowing for a direct and fair comparison. In this setup, the handcrafted features from all modalities (PPG, EDA, and THERM) were concatenated into a single input set per sample. Each of the three classifiers, namely Random Forest (100 estimators), K-Nearest Neighbors (k=5), and SVC, was trained on the combined feature set. This fusion approach enabled the models to exploit inter-modality relationships and complementary signal characteristics, which led to improved performance compared to unimodal models.

2.6.3.4 Deep Learning Model

2.6.3.4.1 Introduction and Model Architecture

2.6.3.4.1.1 Model Overview

This section details the deep learning model designed for multimodal emotion recognition. The architecture is a feature-level convolutional and dense neural network, trained in an end-to-end manner, aiming to learn informative signal representations beyond the constraints of handcrafted features. Unlike the previous approaches, this model is fed the preprocessed raw physiological signals directly, thus avoiding the potential biases and limitations introduced by manual feature extraction techniques. The model was implemented in PyTorch and trained using the CUDA framework on an NVIDIA 3060 GPU.

As in the previous approaches, the same data handling protocols were applied, including stratified data splitting and reproducible random seeds. However, the deep learning model introduced a 5% validation subset carved from the training data, forming an 85% - 5% - 10% split into training, validation, and test sets respectively. This was essential for in-training techniques such as early stopping and learning rate scheduling.

2.6.3.4.1.2 Model Architecture

The model consists of five parallel 1D-CNN branches, each corresponding to one physiological modality: PPG:GRN, PPG:IR, PPG:RED, EDA, THERM. Each branch processes its raw input signal independently, producing a learned feature set that is ultimately concatenated and passed to a fully connected classification dense network. Input lengths varied per modality, and were handled by the code by determining the input size of each branch. All five branches share an identical architecture. The code of the model's definition is added to the Appendix.

2.6.3.4.1.2.1 CNN-Branch Design

Each 1D-CNN branch consists of:

- Five convolutional layers (Conv1d in PyTorch)
- Each followed by Batch Normalization (BatchNorm1d in PyTorch) and Max Pooling (MaxPool1d in PyTorch, with kernel_size = 2)
- The final pooling layer is an Adaptive Average Pooling layer (AdaptiveAvgPool1d in PyTorch), which ensures fixed-length feature sets across branches regardless of input length.

All convolutional layers use:

- kernel_size = 3
- stride = 1 (default)
- dilation = 1 (default)
- padding = 'same', to ensure same input and output sizes.

Each convolutional layer uses the ReLU activation function:

$$ReLU(x) = \max(0, x) \tag{2.3}$$

The number of out_channels (feature maps) per layer is as follows:

- Conv1: 320
- Conv2: 480
- Conv3: 640
- Conv4: 960
- Conv5: 1280

Given the fixed kernel_size = 3, default stride and dilation, and 5 convolutional layers, the effective receptive field (ERF) [38] in terms of time-points is:

$$ERF = 1 + (k - 1) * L = 1 + 2 * 5 = 11 \tag{2.4}$$

To convert this into time duration for each modality, we divide by sampling rate respectively:

- For PPG (25 Hz): 11/25 = 0.44s
- For EDA (15 Hz): 11/15 = 0.73s
- For THERM (7.5 Hz): 11/7.5 = 1.47s

ERF represents the maximum temporal context each output unit at the final layer considers.

2.6.3.4.1.2.2 Dense Network

The output feature sets from each CNN branch (modality) are concatenated along the feature axis using PyTorch's concatenation function. This fused feature set is then passed through a fully connected classification network composed of:

- Dense Layer 1: 1024 neurons
- Dense Layer 2: 768 neurons
- Dense Layer 3: 512 neurons

- Dense Layer 4: 256 neurons
- Dense Layer 5: 128 neurons
- Output Layer: num_classes neurons (parametrically defined)

Each dense layer also uses the ReLU activation, with no dropout by default. The model flows fully end-to-end, as backpropagation flows from the final output back to each modality-specific CNN branch.

2.6.3.4.1.2.3 Loss Function

The model is trained using cross-entropy loss with label smoothing, set to 0.1. The smoothed cross-entropy loss function is:

$$Loss_{CE} = -\sum_{i=1}^{C}(1-\epsilon)y_i \log p_i + \frac{\epsilon}{C}\log p_i \quad (2.5)$$

Where:

- C is the number of classes,
- ϵ is the smoothing factor,
- y_i is the ground-truth label,
- p_i is the predicted probability.

This technique helps the model generalize better by preventing it from becoming overconfident.

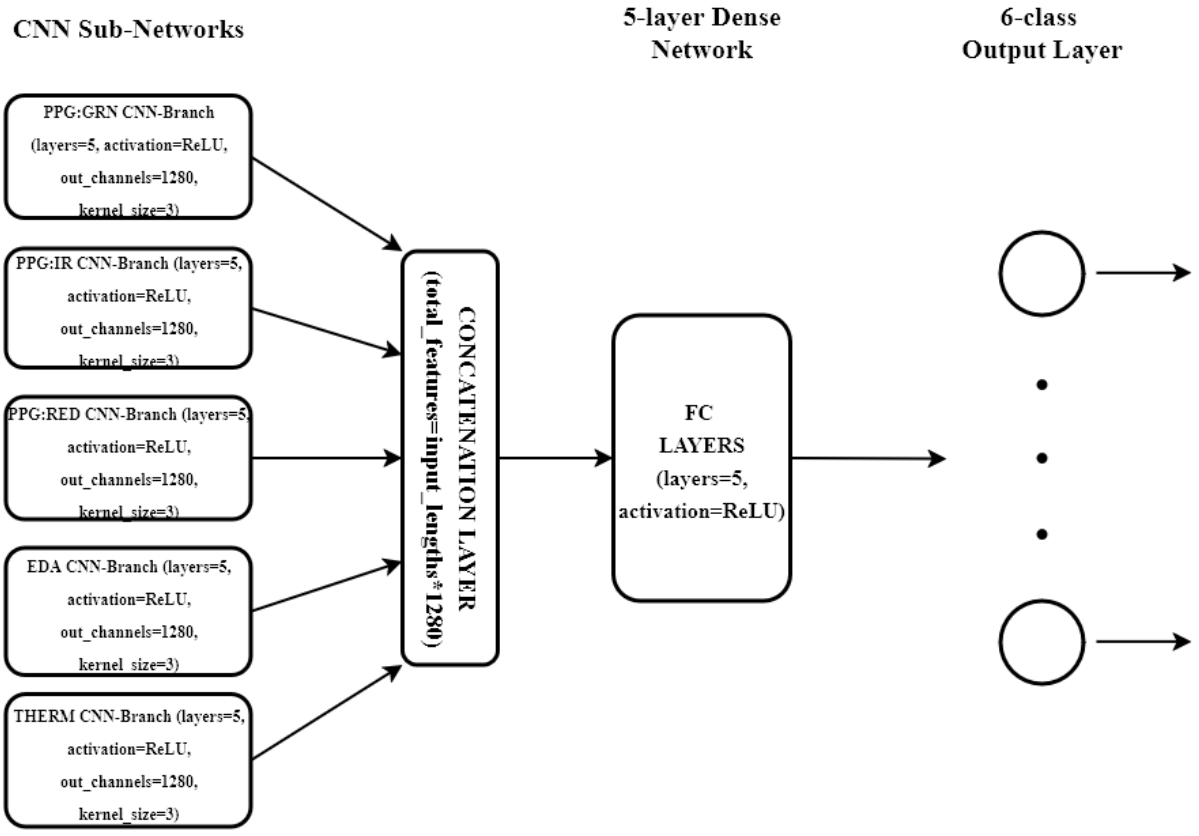

Figure 2.7: Illustrated model architecture.

2.6.3.4.2 Training Procedure and Hyperparameter Tuning

2.6.3.4.2.1 Training Procedure

The model was trained using the Adam optimizer, with the following configuration:

- Initial learning rate: 1e-4
- Weight decay: 1e-4

A ReduceLROnPlateau scheduler was applied to adapt the learning rate based on the validation loss:

- Mode: 'min' (monitors minimum validation loss)
- Factor: 0.5 (reduces learning rate by half)
- Patience: 5 epochs (waits 5 epochs of no improvement)

An early stopping mechanism monitored the same validation loss. The patience was set to 30 epochs, which is approximately one-quarter of the maximum number of epochs, which was set to 125. Both mechanisms were based on the 5% untouched validation set, ensuring robust in-training generalization tracking.

2.6.3.4.2.1.1 Training Details

- During each epoch, the model computed training loss and updated its parameters accordingly.
- After each epoch, the model switched to evaluation mode on the validation set.
- If the validation loss was lower than any previously recorded value, a deep copy of the model state was stored.
- Training was terminated if the validation loss failed to improve for 30 consecutive epochs.
- After training completed, the model state with the lowest validation loss was loaded, and the final performance was evaluated on the untouched test set.

The test set remained isolated during both training and validation, making its performance a legitimate measure of generalization.

2.6.3.4.2.2 Hyperparameter Selection and Tuning Strategy

Given computational constraints, grid search or automated hyperparameter tuning methods were not employed. Instead, an iterative manual tuning approach was adopted based on empirical validation feedback. The following strategy was used to balance underfitting and overfitting.

2.6.3.4.2.2.1 Underfitting

Underfitting was identified when:

- Both training and validation accuracies were low,
- Training and validation losses were high and close in value.

The corrective strategy was to increase model capacity, in small increments to avoid overfitting:

- First, by increasing depth (adding more layers to increase receptive field),
- Then, by increasing width (higher out_channels values per layer).

Regularization values, such as dropout, were sometimes reduced cautiously to allow the model to better capture patterns, and indirectly increase capacity. Throughout tuning, maximum training epochs were

always kept high enough to trigger early stopping, not finish training normally. Various kernel_sizes were tested because they directly influence the effective receptive field.

2.6.3.4.2.2.2 Overfitting

Overfitting was identified when:

- Training accuracy was high,
- Validation accuracy was low, with validation loss stagnating or increasing.

Corrective actions included increasing regularization, such as:

- Adding dropout layers (typically 0.1 - 0.2 in the first dense layers),
- Increasing weight decay,
- Applying label smoothing,

In order to "close the gap" between training and validation metrics, often with some train accuracy penalty. Simplifying the model was also a valid option, though this was less frequently needed.

Dropout was found to easily cause underfitting in this model, likely due to relatively small data volume. Batch normalization was already implemented after each convolutional layer and was kept consistent.

Batch size was fixed at 64, which offered a good compromise between stability and memory usage.

Label smoothing effectively increased the absolute loss values, but the relation between training and validation loss remained.

2.6.3.4.2.2.3 Summary of Tuned Hyperparameters

Capacity-related:

- Number of convolutional layers (result: 5 layers).
- out_channels: incrementally increased per layer.
- Kernel size: fixed at 3, while experimentation confirmed its adequacy.

Regularization-related:

- Weight decay: finally chosen as 1e-4.
- Label smoothing: finally set at 0.1.
- Dropout: tested but not used in final model due to observed underfitting.
- Batch normalization: used throughout convolutional layers.
- Batch size (indirectly).

Other hyperparameters:

- Learning rate: 1e-4.
- Epochs: 125 maximum.
- Early stopping patience: 30.
- Scheduler: ReduceLROnPlateau.
- Scheduler factor and patience: 0.5 and 5, respectively.
- Activation function: ReLU.
- Padding: ‘same’.
- Pooling: max pooling, final adaptive average pooling.
- Optimizer: Adam.
- Stride.
- Dilation.

2.6.3.4.3 Runtime and Disk Space

The full multimodal deep learning model required approximately 21 minutes and 37 seconds to train on a single GPU using the CUDA framework. The training time for the separate Valence and Arousal models was comparable. Upon completion, the saved model state occupied approximately 160 MB of disk space per model.

2.6.3.4.4 Final Evaluation and Model Selection

This model consistently outperformed the classical early fusion ML models. After training with five distinct random seeds, the best-performing model (based on test accuracy) was selected. This model, with seed = 21, was also re-trained using the original 7-class Valence and Arousal labels independently. These additional experiments provided comparative insight into performance when learning Valence and Arousal separately.

The test accuracies of these models are collectively reported in the Results section.

2.7 Summary of Methodology

This chapter detailed the complete methodological pipeline for the multimodal emotion classification task, from raw signal processing to model design and training.

First, the dataset was described, consisting of synchronized physiological recordings (PPG in three wavelengths, EDA, THERM) and associated self-assessed Valence-Arousal ratings per trial. Each modality underwent tailored preprocessing procedures to correct for noise, alignment, and sampling inconsistencies, ensuring clean and time-aligned inputs across modalities.

Feature extraction was implemented using handcrafted features for classical machine learning models, while the deep learning approach used the raw, preprocessed signal inputs. For the classical models, statistical, temporal, and frequency-domain features were extracted per modality. The labels were derived from the original 7-point self-assessed Valence and Arousal scales using a structured binning strategy based on the circumplex model of affect. Six classes were used in the final experiments.

All approaches followed a consistent data split: 85% training, 5% validation (used only in deep learning), and 10% testing. Stratified splitting preserved class distributions, and random seeds ensured reproducibility.

Three machine learning approaches were implemented:

- Unimodal classical models such as Random Forest, KNN, and SVM trained on single-modality features.
- Early fusion models used concatenated features across modalities and were trained with the same algorithms and hyperparameters, isolating fusion as the independent variable.
- A deep learning model was designed using five parallel CNN branches (one per modality), which learned representations directly from raw signals. The architecture was trained end-to-end using cross-entropy loss with label smoothing, Adam optimizer, early stopping, and a learning rate scheduler.

Hyperparameter tuning for the deep learning model was performed empirically through comparisons of training and validation behavior, balancing underfitting and overfitting. Regularization strategies such as weight decay, label smoothing, and batch normalization were crucial to improving generalization without over-constraining model capacity.

The next chapter presents the performance results and comparative evaluation across all approaches.

Chapter 3: Results

3.1 Introduction to Results

This chapter presents the performance results of the three modeling approaches developed for the task of emotion recognition using physiological signals: unimodal classical models, early fusion classical models, and a deep learning architecture. The primary objective was to investigate the impact of multimodal fusion and learned representations on model performance, with a particular focus on scenarios close to real-world wearable emotion recognition.

The metric used for evaluation was accuracy, due to its simplicity, widespread acceptance in classification problems, and its effectiveness as a balanced metric. Each approach was evaluated across five distinct random seeds to account for variability and randomness in training and dataset splitting. For each model, the average accuracy and standard deviation across seeds are reported. In the case of unimodal classical models, results from each modality were first averaged per seed to reflect performance in a practical unimodal deployment setting, before computing overall statistics.

3.2 Unimodal Classical Models

The unimodal models served as baselines, using handcrafted features from each modality separately. Three algorithms were applied: Random Forest (100 estimators), K-Nearest Neighbors (k=5), and Support Vector Machine (SVC). Each model was trained and evaluated using the same seeded data splits, and the final accuracy was averaged across modalities for each seed.

Table 3.1: average accuracy per seed, and overall mean and standard deviation across the five seeds.

Model	Seed 7	Seed 11	Seed 21	Seed 35	Seed 42	Mean Accuracy	Std Dev
Random Forest	0.6444	0.6612	0.6784	0.6622	0.6172	0.6527	±0.0207
KNN	0.4919	0.5033	0.4897	0.5196	0.4966	0.5002	±0.0108
SVM	0.4373	0.4373	0.4268	0.4555	0.4172	0.4348	±0.0128

Across all seeds, the Random Forest model consistently outperformed both KNN and SVM. The KNN model showed a moderate degree of consistency, while the SVM model was the least effective across all runs, demonstrating poor generalization when trained on unimodal handcrafted features.

3.3 Early Fusion Classical Models

To examine the effect of feature-level multimodal fusion, the same three classifiers were trained using concatenated features across all five modalities. All hyperparameters were held constant to isolate the effect of fusion. This strategy enabled the models to learn inter-modal feature relationships, thereby improving emotion classification.

The early fusion models consistently outperformed their unimodal counterparts in the case of Random Forest, and in some cases for KNN and SVM as well. The table below presents the performance of each classifier across the five seeds:

Table 3.2: Accuracy of early fusion classical models across seeds.

Model	Seed 7	Seed 11	Seed 21	Seed 35	Seed 42	Mean Accuracy	Std Dev
Random Forest	0.8230	0.8134	0.8756	0.8852	0.7895	0.8373	±0.0369
KNN	0.4115	0.5263	0.5215	0.5502	0.4545	0.4928	±0.0516
SVM	0.3923	0.4498	0.3971	0.4258	0.3732	0.4076	±0.0270

Among the three models, the Random Forest classifier saw the most pronounced benefit from early fusion, achieving a significant improvement over its unimodal variant (mean accuracy 83.73% vs. 65.27%). KNN exhibited a mild increase in performance compared to the unimodal version, whereas SVM remained similarly limited in effectiveness, with relatively low accuracy.

These findings reinforce that classical models can benefit from feature-level multimodal fusion, making it a viable approach in low-complexity or embedded systems, where deep learning may be infeasible.

3.4 Deep Learning Model Performance

The end-to-end 1D-CNN-based deep learning model trained directly on preprocessed raw signals delivered the highest performance across all approaches. As detailed in the previous chapter, the architecture included five identical CNN branches, one per modality, whose outputs were fused and passed through a dense network. Training utilized the same data partitions, but added a 5% validation set to guide learning via early stopping and adaptive scheduling.

The model was trained using the same 5 seeds. For each seed, the best model was saved based on validation loss and evaluated on the untouched test set. The table below shows the test accuracy for each seed, along with the mean and standard deviation:

Table 3.3: Accuracy of deep learning model across seeds.

Seed	Accuracy
7	0.9426
11	0.9282
21	0.9474
35	0.9474
42	0.9378
Mean	0.9407
Std Dev	±0.0072

The model exhibited high performance with remarkably low variance, consistently outperforming all other methods. The results indicate that learning representations directly from preprocessed raw signals, coupled with the complexity and capacity that characterizes DL models, offers a substantial improvement in classification accuracy and robustness. Compared to early fusion Random Forest (mean accuracy of 83.73%) and unimodal baselines (max ~65%), the deep learning model achieves the highest average accuracy of 94.07%, validating its capacity to generalize across seeds.

The deep learning model demonstrated a marked improvement over classical approaches, with an average accuracy boost of nearly 10 percentage points over early fusion RF models, suggesting not just statistical but also practical significance in real-world settings.

Additionally, it showed not just high mean accuracy, but low standard deviation, suggesting consistency.

In addition to the 6-class emotion model, the best seed (Seed 21) was used to re-train the model on the original 7-class Valence and Arousal label sets. The final test accuracies obtained for these models were:

- Valence Model (7-class): 0.9474
- Arousal Model (7-class): 0.9378

Finally, the training and validation learning curves for all five seeds are shown in Figures 3.1 - 3.10. The Figures 3.11 - 3.14 represent the models trained on the 7-class tasks. These curves provide a visual inspection of training stability, convergence behavior, and overfitting.

3.4.1 Training and Validation Accuracy and Loss Curves per Seed

The oscillatory behavior observed in the validation accuracy and loss curves can be attributed to the small size of the validation set, which comprised only 5% of the total dataset, approximately 104 samples. With such a limited number of validation examples, small fluctuations in prediction performance can result in relatively large changes in the computed validation metrics across epochs. Increasing the validation set size might have mitigated these oscillations by offering a more stable estimate of generalization. However, doing so would have reduced the amount of data available for training, which is critical given the overall small dataset. On the other hand, reducing the size of the test set was not a viable solution, as it would compromise the integrity of the final evaluation. Consequently, the chosen split reflects a necessary trade-off between reliable validation monitoring and maximizing training efficiency.

Additionally, the absolute loss values are high due to label smoothing, however the relation between the training and validation loss curves remained.

3.4.1.1 Seed = 7

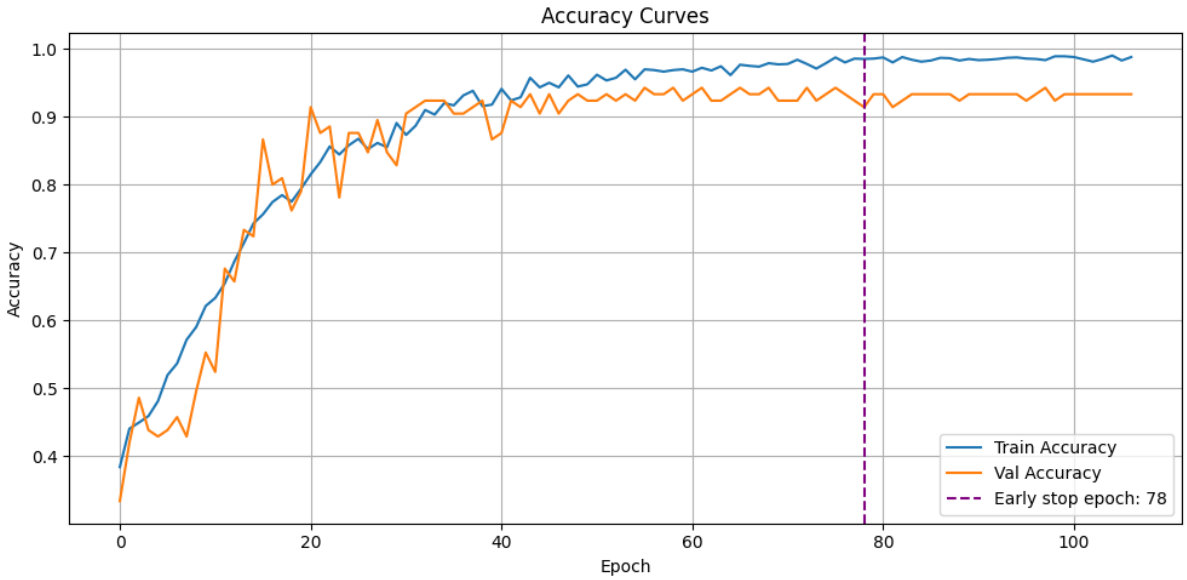

Figure 3.1: Accuracy curves for seed = 7.

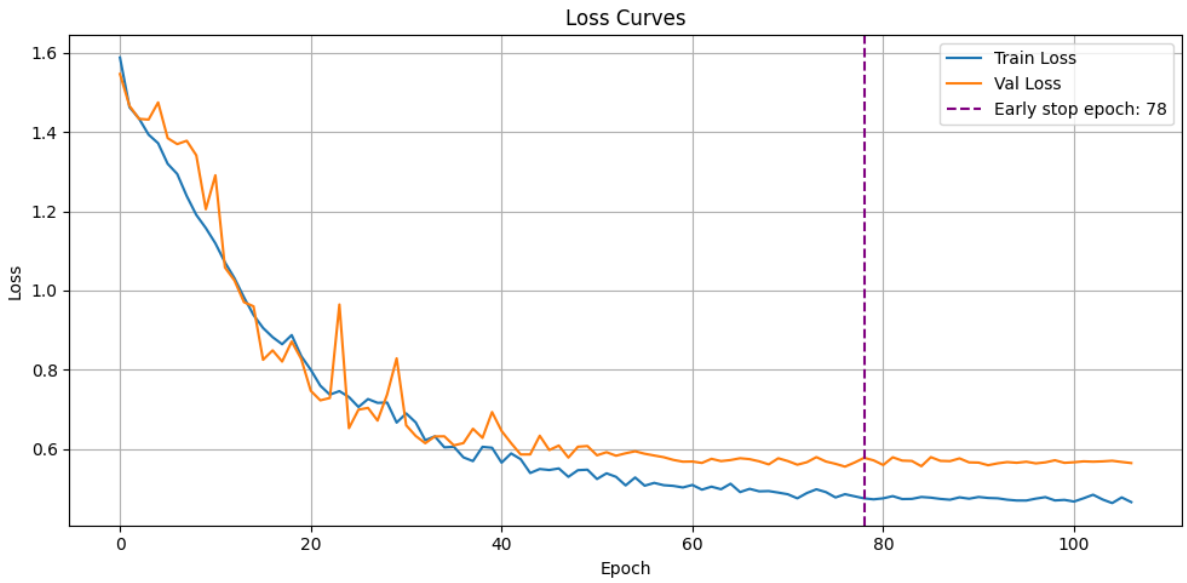

Figure 3.2: Loss curves for seed = 7.

3.4.1.2 Seed = 11

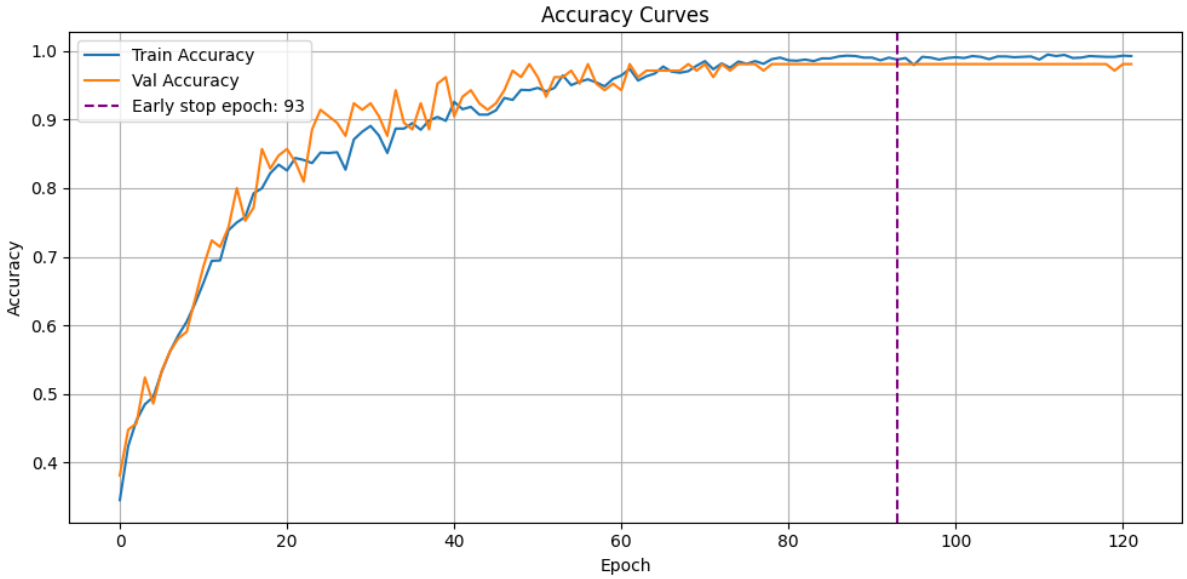

Figure 3.3: Accuracy curves for seed = 11.

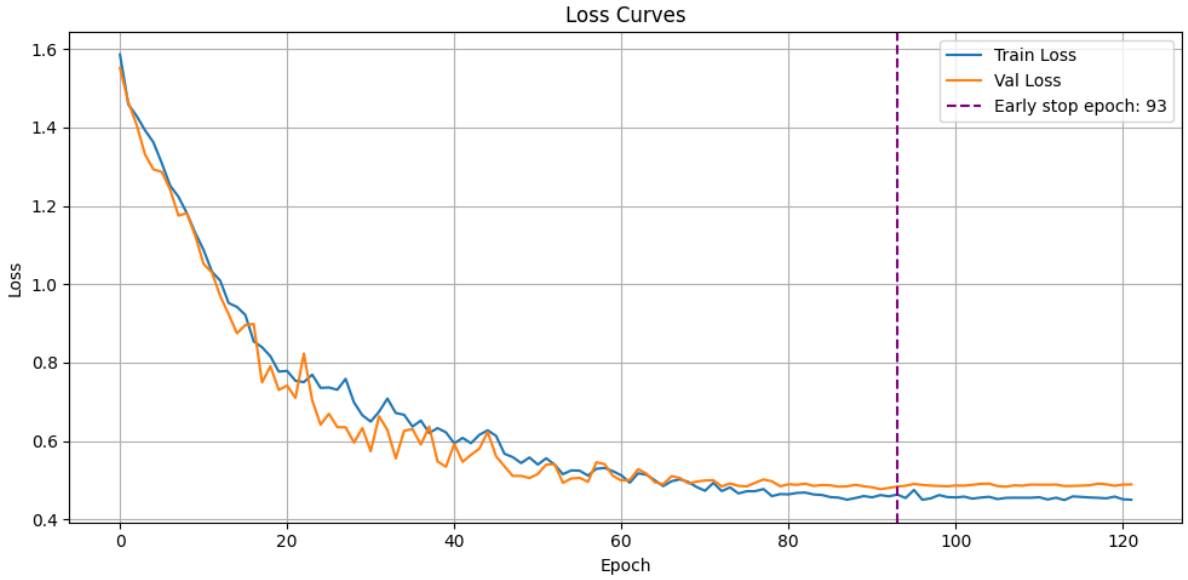

Figure 3.4: Loss curves for seed = 11.

3.4.1.3 Seed = 21

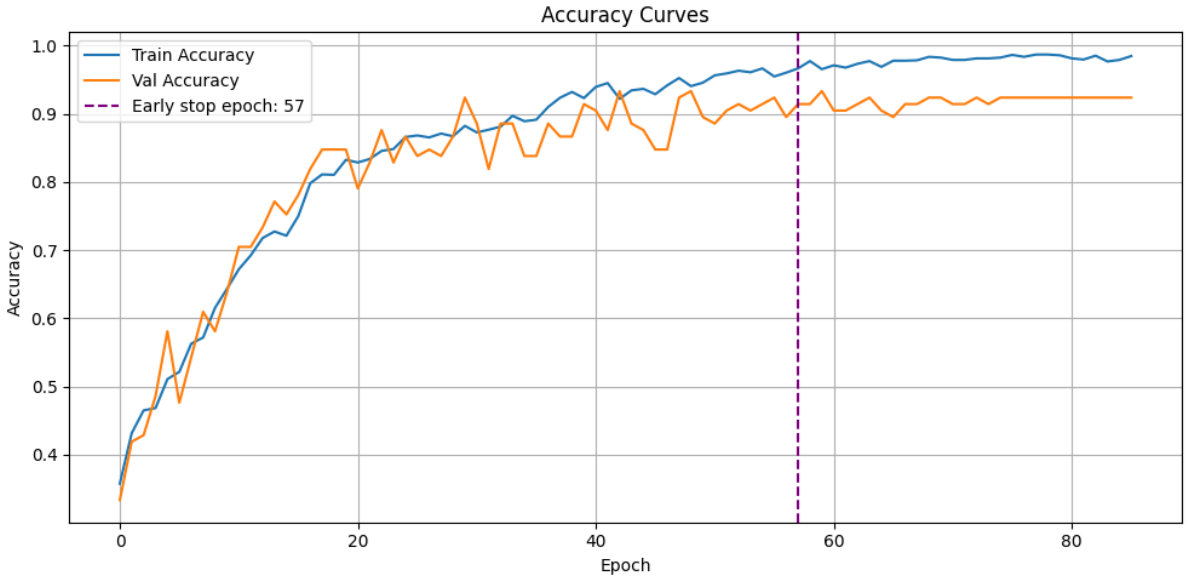

Figure 3.5: Accuracy curves for seed = 21.

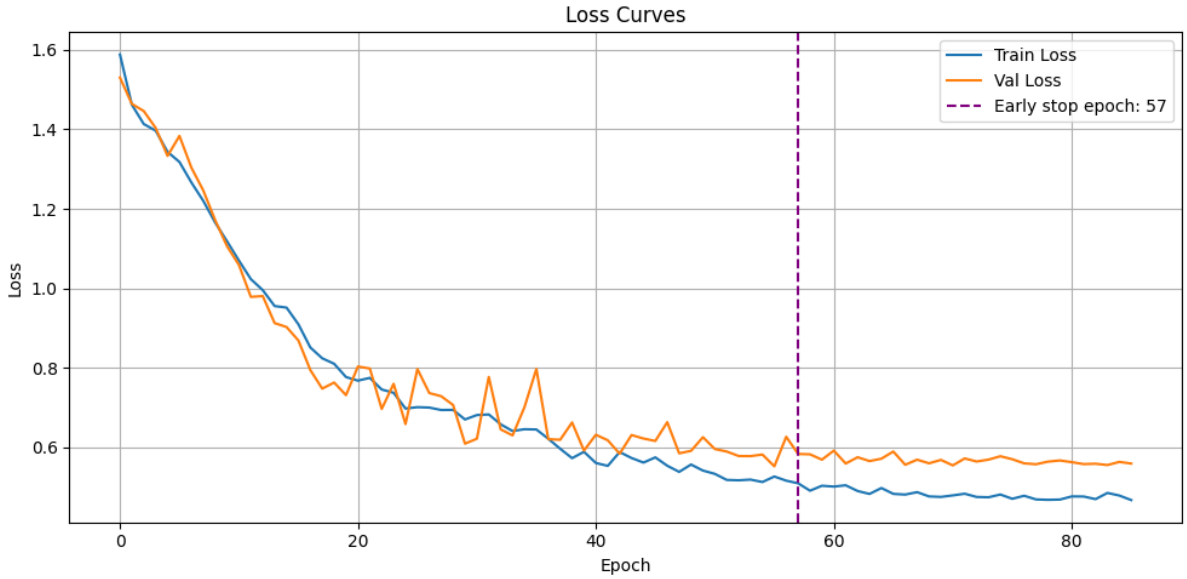

Figure 3.6: Loss curves for seed = 21.

3.4.1.4 Seed = 35

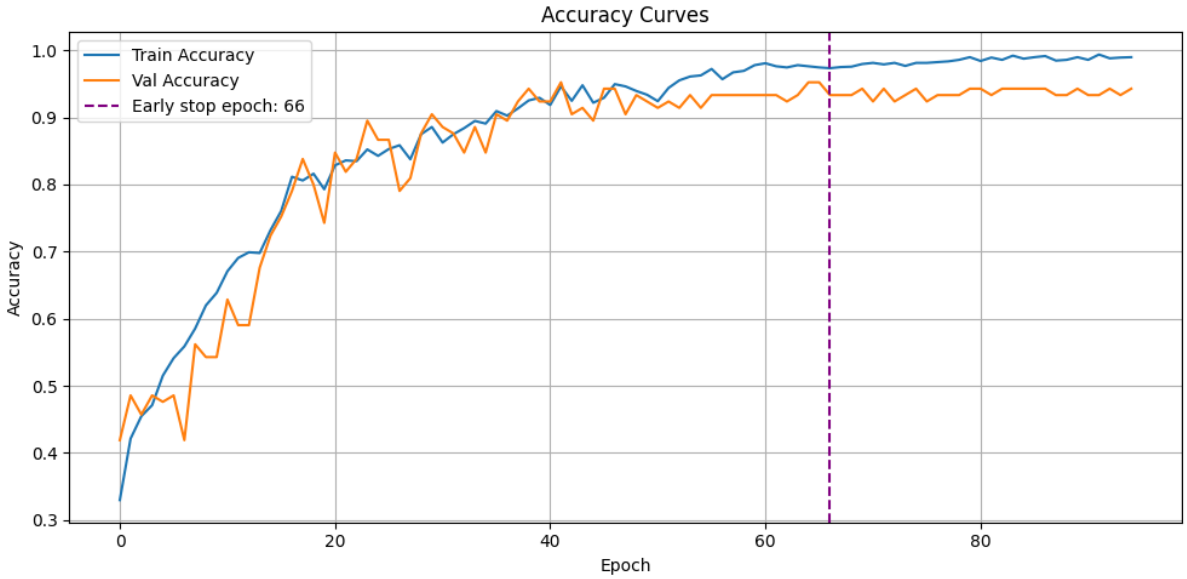

Figure 3.7: Accuracy curves for seed = 35.

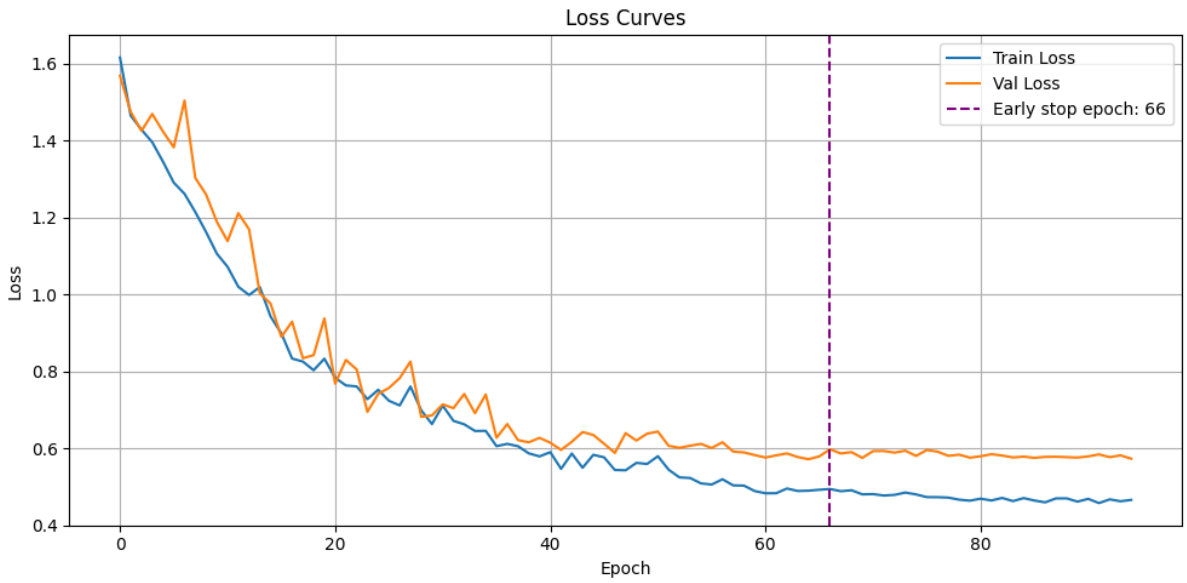

Figure 3.8: Loss curves for seed = 35.

3.4.1.5 Seed = 42

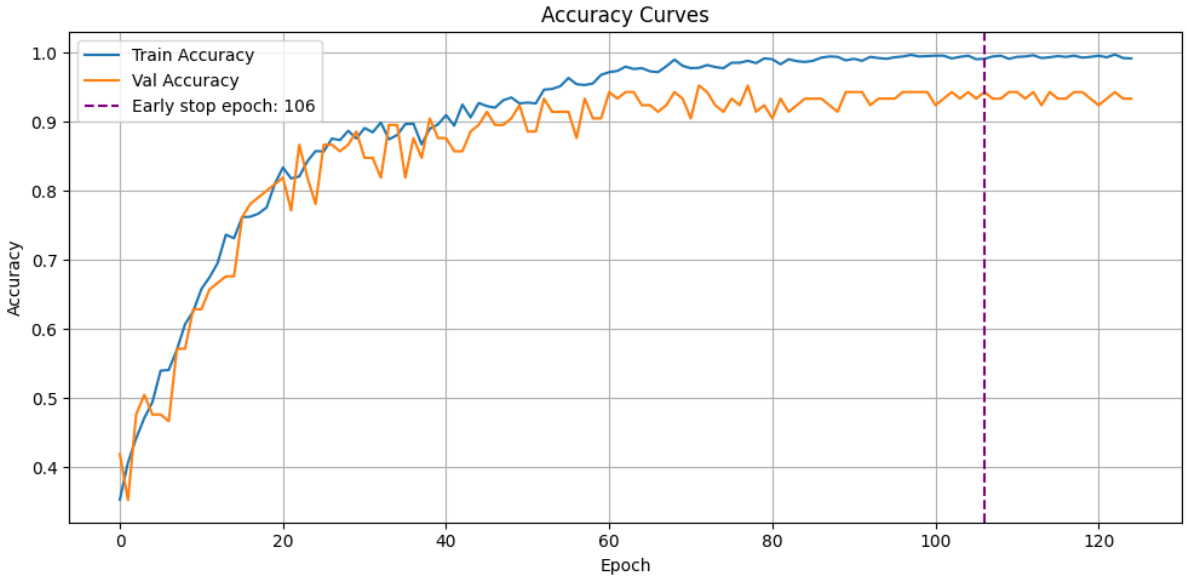

Figure 3.9: Accuracy curves for seed = 42.

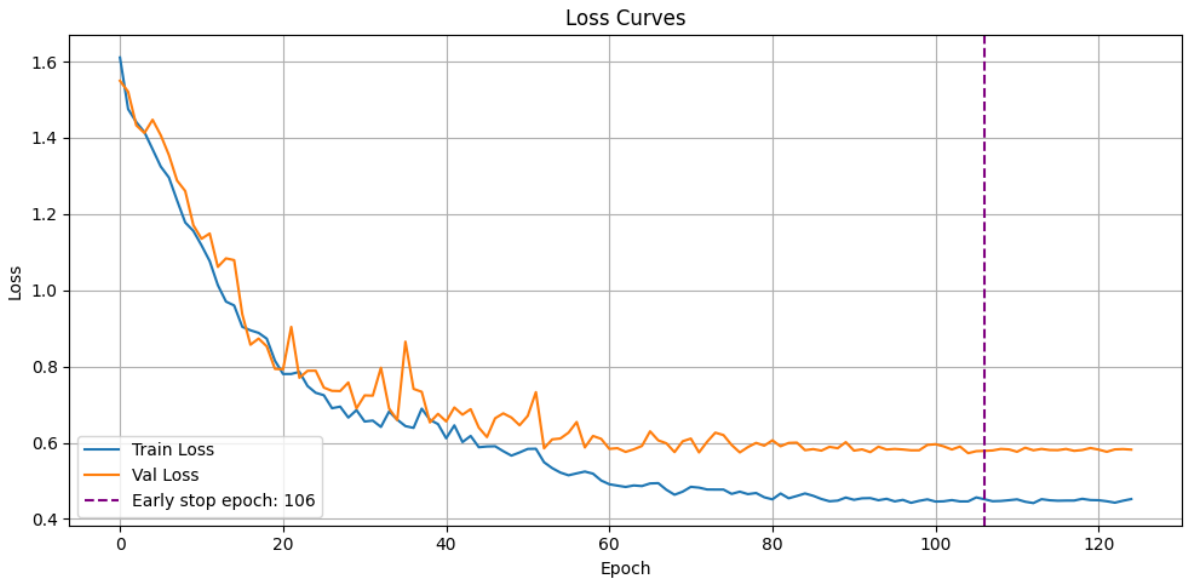

Figure 3.10: Loss curves for seed = 42.

3.4.1.6 Valence Model

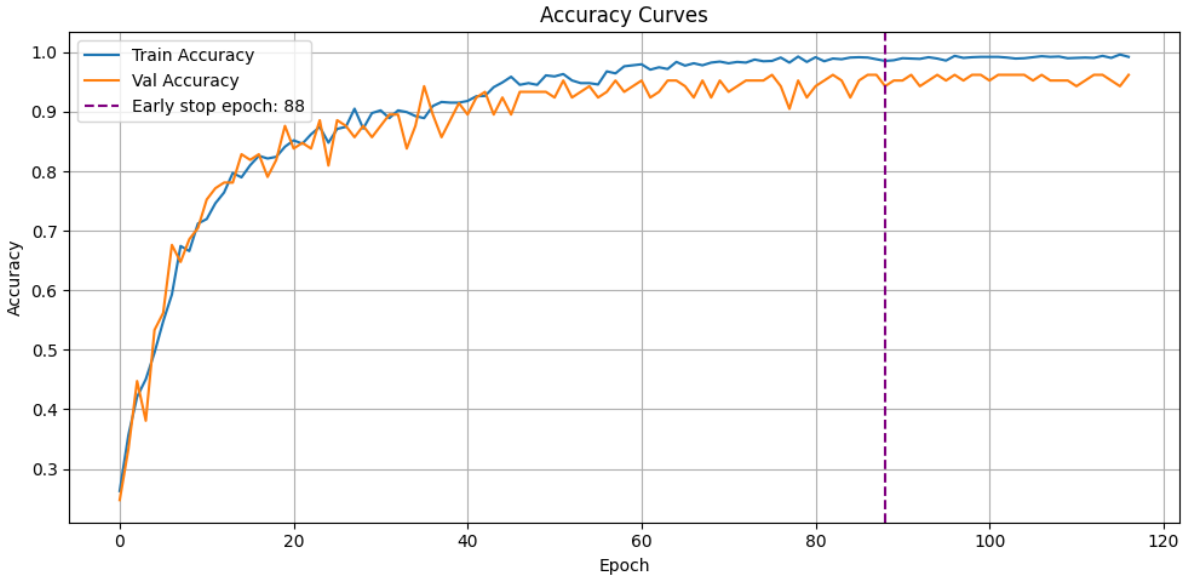

Figure 3.11: Accuracy curves for the Valence model.

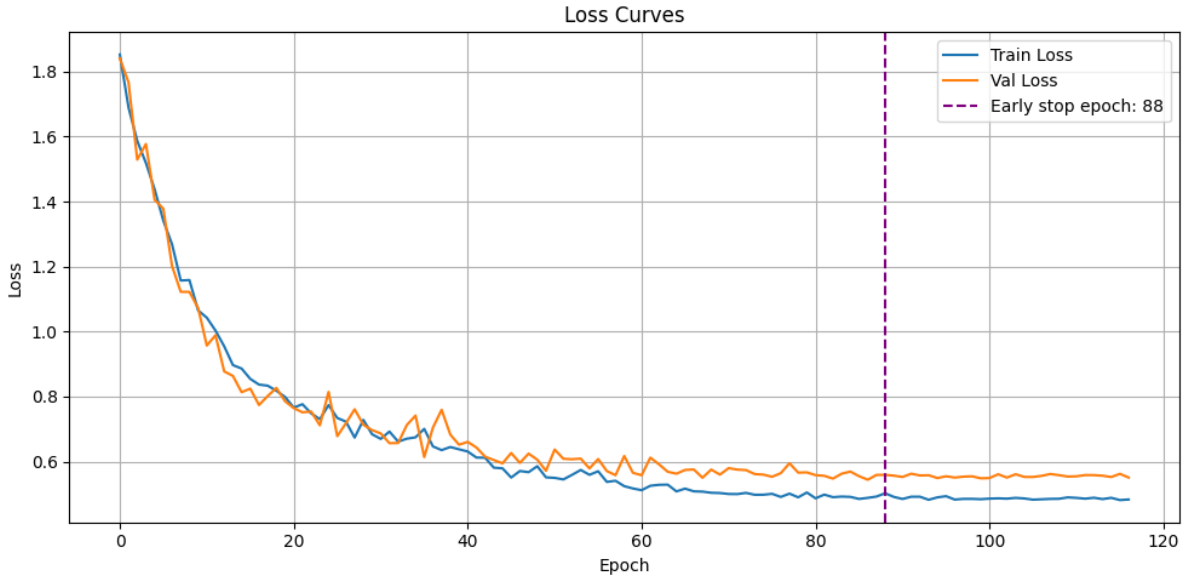

Figure 3.12: Loss curves for the Valence model.

3.4.1.7 Arousal Model

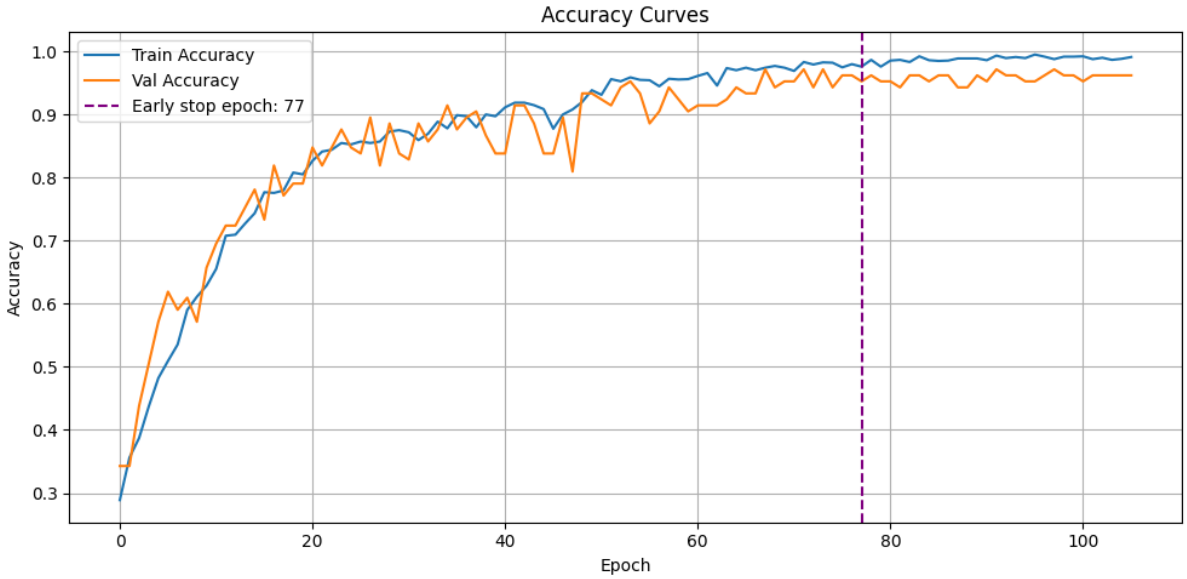

Figure 3.13: Accuracy curves for the Arousal model.

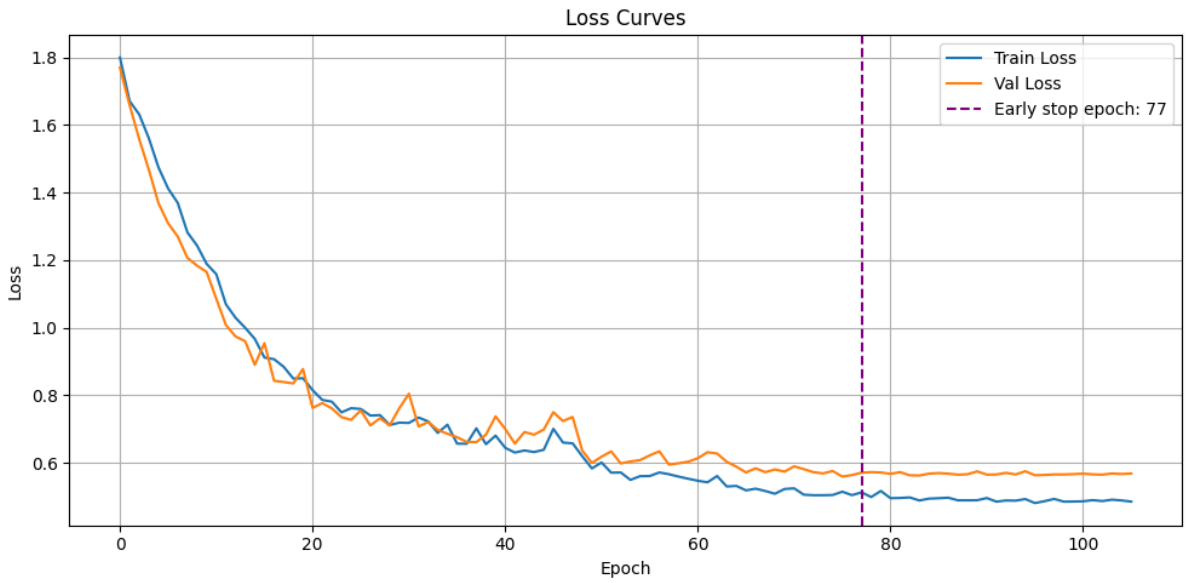

Figure 3.14: Loss curves for the Arousal model.

3.5 Interpretation of Results

These results confirm that early fusion of physiological modalities yields better performance than isolated unimodal processing. Furthermore, deep learning approaches leveraging raw signals outperform handcrafted feature-based classical models, aligning with the thesis' aim of moving toward a robust, real-world-ready emotion recognition system.

The bar chart below provides a visual summary of the average classification accuracy achieved by each methodological approach, Unimodal Classical ML, Early Fusion Classical ML, and the Deep Learning model, along with their respective standard deviations across models or training seeds, highlighting the performance gains enabled by fusion and learned feature representations:

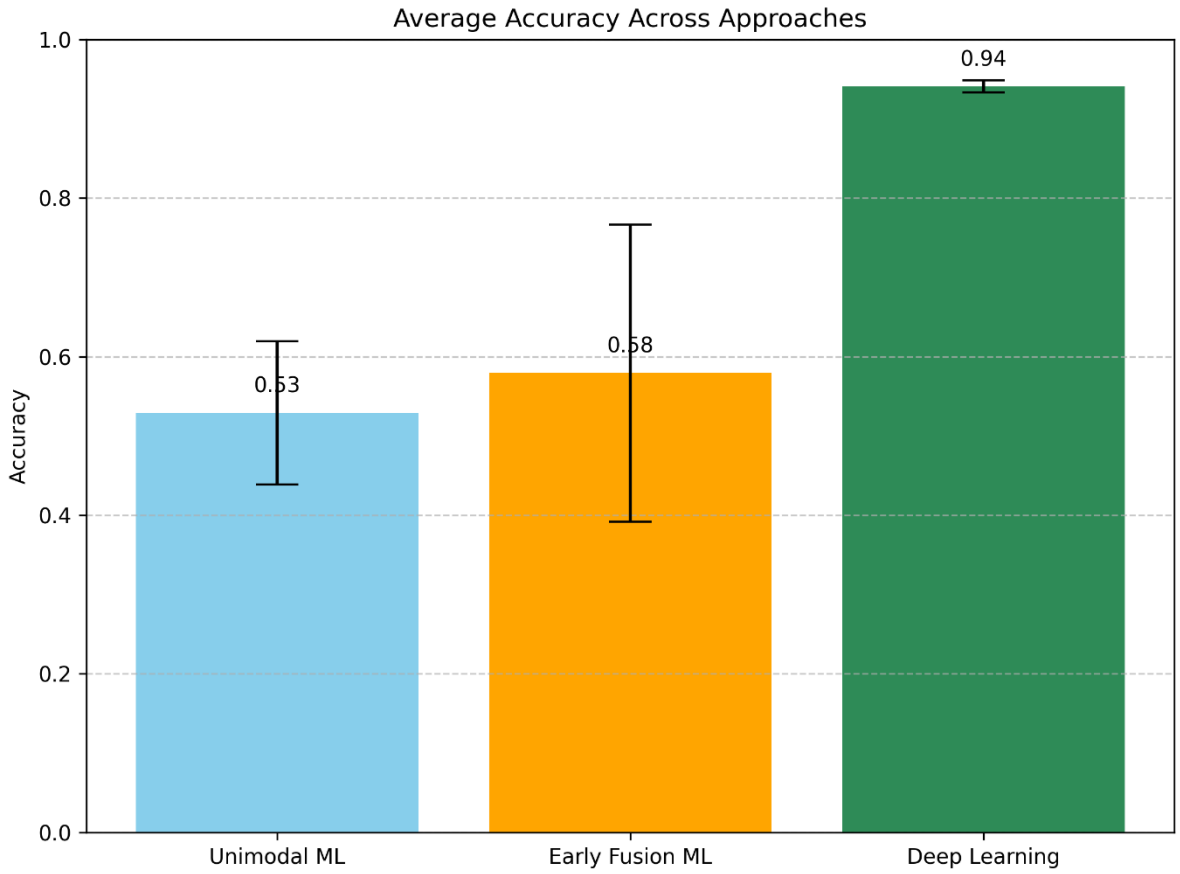

Figure 3.15: Bar chart highlighting the differences of the approaches in performance.

To complement the analysis at the model level, the following bar chart presents the accuracy and standard deviation of each individual model across both classical ML approaches, as well as the deep learning model. This detailed view highlights how early fusion impacts the performance of each classical model specifically, and emphasizes the consistent superiority of deep learning regardless of baseline comparisons.

The increase in Random Forest's performance under early fusion can be attributed to its ensemble nature and ability to model complex, high-dimensional feature interactions. By aggregating decision trees, RF naturally benefits from the richer and more diverse feature space provided by early fusion. It leverages complementary information from multiple modalities more effectively than simpler models.

In contrast, KNN and SVM are more sensitive to the structure and scaling of the input feature space. Early fusion leads to a significant increase in feature dimensionality, which can introduce noise or imbalance in feature contributions. For KNN, the higher the dimensionality, the more distance metrics become less meaningful. For SVM, high dimensionality can lead to poor generalization, especially when class boundaries become more complex due to feature fusion without sufficient samples to support hyperplane construction.

Moreover, both SVM and KNN lack internal mechanisms for feature selection or weighting unless explicitly added. Without preprocessing techniques like feature selection, their performance can degrade, unlike RF which implicitly handles feature importance during tree construction.

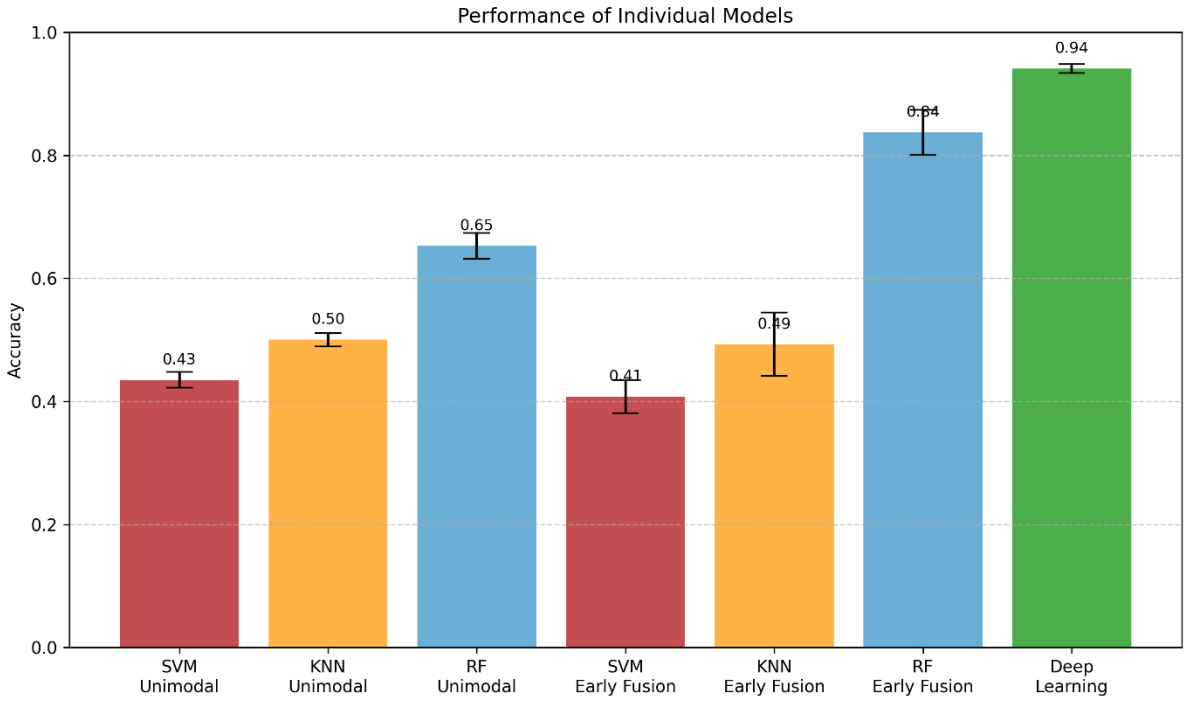

Figure 3.16: Bar chart highlighting the differences of the models in performance.

To complement the overall performance analysis, the class confusion matrix of the deep learning model for Seed 21 is presented below. This seed was selected as it produced one of the highest test accuracies among all training runs, and the deep learning model as a whole demonstrated the most consistent performance across seeds compared to classical baselines. This visualization highlights the distribution of correct and incorrect predictions across the target classes, offering a more detailed perspective beyond total accuracy.

As shown, the majority of predictions fall along the diagonal, indicating that the model correctly classifies most samples across all six classes. There are no significant off-diagonal concentrations, suggesting that no particular pair of classes is frequently confused.

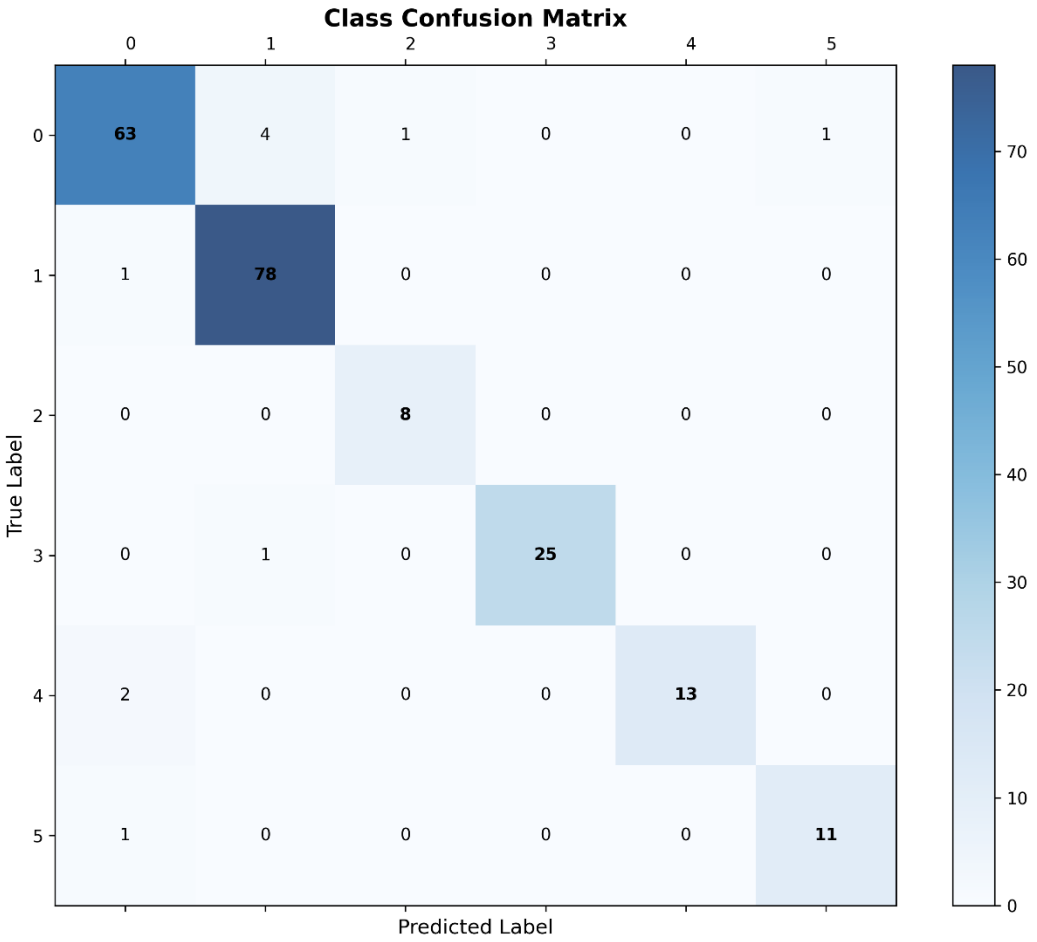

Figure 3.17: Class Confusion Matrix for the DL model with Seed 21.

Chapter 4: Conclusion and Suggestions

4.1 Thesis Recap

This thesis aimed to explore and evaluate the effectiveness of multimodal fusion techniques for emotion recognition using physiological signals, with a focus on robustness and suitability for real-world implementation. To that end, three methodological approaches were compared: unimodal classical machine learning models based on handcrafted features, early fusion classical ML models leveraging combined feature sets across modalities, and a deep learning model trained end-to-end on raw, aligned and filtered physiological signals.

The dataset consisted of five biosignals, three PPG channels (green, infrared, and red), EDA, and skin temperature, recorded in a valid setting. Emotion labeling followed the dimensional Valence-Arousal model (Figure 1.1). All models were trained and evaluated using a consistent data splitting strategy across five random seeds, ensuring fair comparisons and reproducibility. Classical models included Random Forest, K-Nearest Neighbors, and Support Vector Machines, while the DL model utilized a 1D CNN-based architecture with parallel branches for each modality. Performance was evaluated using accuracy, and results were aggregated across seeds to mitigate the effects of data variability and limited sample size.

4.2 Key Findings

The results clearly demonstrated the benefits of multimodal data integration and deep learning in the context of emotion recognition from physiological signals. First, early fusion of handcrafted features from all five modalities led to noticeable overall improvements in classification accuracy compared to unimodal models. This validates the hypothesis that emotional information is more effectively captured when complementary signals are processed together rather than in isolation. Notably, Random Forest showed the most significant gains in the early fusion setting, whereas KNN and SVM did not benefit to the same extent, likely due to their limited ability to capture interactions in higher-dimensional feature spaces.

The deep learning model outperformed both classical approaches by a substantial margin, achieving a mean accuracy of 0.9407 (±0.0072), compared to 0.5792 (±0.1857) for early fusion ML and 0.5292 (±0.0913) for unimodal ML. This reflects the power of representation learning from raw data, which bypasses the limitations of handcrafted features and leverages the full temporal and morphological structure of physiological signals.

Importantly, the DL model's superior performance was consistent across all five random seeds, underscoring both its stability and robustness. Additionally, the deep learning model had a training runtime of approximately 21 minutes and a saved model size of around 160 MB.

These findings collectively underscore the importance of multimodal integration and end-to-end learning for advancing affective computing systems, towards practical, real-time applications.

4.3 Implications for Real-World Deployment

The findings of this thesis have several important implications for real-world applications, particularly in scenarios involving wearable or embedded conditions. The results demonstrated that emotion recognition systems based on multimodal physiological signals can achieve high performance when

employing early fusion techniques and deep learning models. This supports the feasibility of implementing such systems in dynamic, real-world environments.

One key advantage is that the data sources used, such as photoplethysmography (PPG), electrodermal activity (EDA), and skin temperature, are all non-invasive, readily integrated into compact multi-sensor platforms, and already commonly collected by wearable devices. This strengthens the notion for embedding early-fusion or deep learning models into wearable sensors and devices aimed at affective computing.

Additionally, the use of raw, aligned signal inputs in the deep learning architecture eliminates the need for manually engineered features and domain-specific processing. This reduces development overhead and enables more adaptable and well-generalizable models, important attributes for consumer or healthcare applications where input conditions may vary. Furthermore, the consistency of performance across five random seeds indicates a stable and reproducible training process. While training such models on edge devices may not be feasible due to computational demands, real-time inference post-deployment remains a viable path, especially with further model optimization techniques.

These characteristics align well with needs in areas such as healthcare, for continuous and real-time stress and patient emotion monitoring during treatment, advanced human-computer interaction systems that seek to adapt to user emotional states in real time, and the advancement of the affective computing in general.

Nonetheless, key limitations remain. While high accuracy was achieved in this thesis, it was obtained using a constrained and relatively small dataset. Real-world deployment would involve additional challenges such as physiological noise, motion artifacts, and unseen environmental conditions. These concerns are further discussed in the following section.

4.4 Limitations

4.4.1 Dataset Constraints

The dataset used in this thesis resulted in approximately 2,000 samples after processing, a modest size, especially for training deep learning models that rely on large-scale data to generalize effectively. Furthermore, the dataset was collected under controlled laboratory conditions, limiting the natural variability present in real-world settings. Emotion labels were based on self-reports, which are inherently subjective and may not capture subtle or nuanced emotional states accurately. Additionally, emotion labeling presents challenges due to the complex and often ambiguous nature of emotional experiences. Inter-subject variability introduces further complications, as physiological responses can vary significantly across individuals. Finally, although the dataset exhibited class imbalance, it was preserved intentionally to reflect natural emotional distributions more realistically, which may have influenced model learning dynamics.

4.4.2 Generalizability to Real-World Scenarios

While promising results were achieved using data collected under controlled conditions, the model's performance in unconstrained, real-world environments remains untested. Real-world applications will likely involve additional challenges such as:

- Movement artifacts due to physical activity,
- Inconsistent sensor placement,
- Environmental noise, such as lighting and ambient temperature,

- Differences in hardware or sensor models.

The ability of the system to generalize to new users, unseen environments, and varied sensor conditions is currently uncertain.

4.4.3 Sensor and Signal Quality

The physiological sensors used, though non-invasive and suitable for wearable integration, show several technical limitations:

- PPG signals are highly susceptible to motion-induced noise and variations in skin contact.
- EDA and skin temperature probes are sensitive to environmental conditions and have relatively slow response times.
- Multimodal synchronization required careful preprocessing. Interpolating and aligning signals across modalities and time windows posed challenges, especially in preserving temporal integrity and dimensional consistency.

4.4.4 Methodological Trade-offs

Due to the small dataset size, only 5% was allocated to validation, which contributed to high variance across validation curves. While this approach preserved training data, it increased the instability of performance estimates. Additionally, classical machine learning models depended on handcrafted features, a process that introduces human bias and requires domain knowledge. Although deep learning partially mitigates this by learning representations directly from raw data, it also demands substantially more training data and compute.

Furthermore, extensive preprocessing was necessary across all modalities due to the inherently noisy nature of biosignals. This means that a dedicated preprocessing pipeline, external to the machine learning architecture, was required and implemented in Python in this case. Such pipelines may carry their own limitations, including maintainability, portability, and increased complexity in deployment scenarios.

4.4.5 Real-Time Deployment

Training time for each deep learning model was approximately 21 minutes, and the resulting model files were approximately 160 MB in size, a moderately large size for embedded or edge systems. While real-time inference after deployment may be feasible, on-device training remains impractical, especially as dataset size and diversity increase.

The model operates on 30-second data windows, meaning any real-time system would require buffering at least this duration before making a prediction. This limits temporal resolution and could cause the model to miss rapid emotional transitions. Shortening the window might allow more frequent predictions but would reduce the amount of input data per sample, potentially degrading accuracy.

Additionally, due to constrained computational resources during development, cross-validation could not be applied. Instead, a fixed-seed repeated train-test splitting approach was employed to simulate variability across runs. This strategy allowed assessment of the model's stability across different random partitions, serving as a practical substitute for cross-validation in this context.

4.4.6 Interpretability and Explainability

The deep learning model achieved high accuracy, but its inner workings are obscured, a common limitation of neural networks. In high-stakes domains such as healthcare, interpretability is essential. The lack of explainability may hinder trust, and future iterations should explore explainable AI (XAI) methods to make decisions more transparent.

4.4.7 Ethical Considerations

Emotions are inherently private, and systems that monitor affective states must be subject to strict privacy safeguards and user consent protocols. As emotion recognition technology matures, it will be essential to ensure responsible data collection, usage, and deployment in line with ethical and legal standards.

4.5 Suggestions

To build upon the foundation established in this thesis and move toward more robust, real-world emotion recognition systems, several directions for future work are proposed across data acquisition, model optimization, signal processing, and personalization:

4.5.1 Dataset Expansion and Diversity

Future research would benefit greatly from expanding both the size and diversity of the dataset. Increasing the number of participants, especially with a broad range of age groups, backgrounds, and emotional profiles, can enhance model generalizability, particularly for previously unseen users.

Furthermore, collecting physiological data in less controlled, semi-structured or real-world environments, for example through wearables in daily life, could provide more valid samples, supporting deployment beyond laboratory conditions. Cross-dataset evaluation, which suggests training on one dataset and testing on another, would also serve as a strong measure of model robustness. Finally, to address inherent subjectivity in emotional self-reporting, future datasets may benefit from combining self-assessment with more objective labeling techniques, such as third-party observation, facial expression analysis, or baseline physiological profiling.

4.5.2 Real-Time and Embedded Optimization

Although the current deep learning model demonstrated acceptable runtime and manageable size, further optimization is required for deployment in low-resource or real-time systems. One approach involves designing smaller and more efficient architectures optimized for edge deployment without a significant drop in performance. Post-training weight pruning may also help reduce the model's memory footprint and inference time, bringing it closer to real-time and embedded hardware platforms.

4.5.3 Improved Multimodal Fusion Strategies

The present thesis used early fusion and deep multimodal representation learning. Future research may explore hybrid fusion architectures that combine early and late fusion advantages, potentially guided by learnable attention mechanisms that dynamically assign weights to each modality based on contextual importance.

Additionally, current deep learning architectures treat all input branches identically. A more modality-aware approach could involve assigning unique architectures to each input stream (branch), tailored

specifically to the signal characteristics of PPG, EDA, skin temperature, thereby improving feature extraction efficiency per modality. Another promising direction is to develop models capable of handling missing modalities through techniques such as modality dropout or conditional networks, enhancing robustness in real-world applications where sensor failure is possible.

For classical machine learning pipelines, feature importance analysis and feature selection should be pursued to reduce dimensionality in early fusion approaches to improve model performance.

4.5.4 Signal Processing Enhancements

Given the noisy nature of biosignals, signal quality remains a key limitation. Future systems could integrate adaptive signal processing steps directly into the ML pipeline, such as learnable preprocessing layers. Additionally, time-aligned accelerometer data could be leveraged to detect motion artifacts and activate adaptive noise reduction algorithms that preserve the integrity of the underlying physiological signals. Such advances would further improve both model accuracy and real-world robustness.

4.5.5 Personalization and Transfer Learning

Inter-subject variability remains a challenge in affective computing. To improve model adaptability, future research may implement subject adaptation techniques such as lightweight fine-tuning on user-specific data. This would allow the model to quickly adjust to new users with minimal input.

In addition, integrating user context features, such as age, gender, activity level, health status, as inputs could provide valuable personalization cues, enhancing classification accuracy and stability. Transfer learning and meta-learning approaches could also be explored to enable efficient generalization across users, and use cases.

4.6 Final Thoughts

This thesis journey has been both a meticulous scientific attempt and an invaluable learning experience. Through the exploration of machine learning, deep learning, and multimodal fusion, the thesis provided not only technical insights but also a broader understanding of how artificial intelligence can be harnessed to interpret complex human phenomena such as emotion.

Multimodal fusion and deep learning emerged as particularly powerful principles, capable of capturing the richness and complexity of physiological data far beyond the capabilities of unimodal or hand-engineered methods. Their ability to process and combine diverse biosignals reinforces their role in enabling technologies in human-computer interfaces.

Emotion recognition, while challenging, holds immense potential. As a subtle and often unconscious form of user input, emotional state detection can pave the way for advanced affective computing systems that respond empathetically and intelligently. From stress monitoring to adaptive interfaces and healthcare applications, such systems may enhance human well-being in meaningful and impactful ways.

On a personal level, this thesis has been a constructive academic experience. It offered a hands-on education in machine learning, signal processing, neural architectures, and model design. More importantly, it fostered a deep appreciation for the role of intelligent systems in bridging the gap between technology and the nuances of human behavior.

It is the author's hope that this thesis not only contributes to the growing body of knowledge in affective computing but also serves as a useful foundation for fellow researchers who share a passion for machine learning and artificial intelligence.

REFERENCES

[1] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans Pattern Anal Mach Intell*, vol. 31, no. 1, pp. 39–58, 2009, doi: 10.1109/TPAMI.2008.52.

[2] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Trans Affect Comput*, vol. 1, no. 1, pp. 18–37, Jan. 2010, doi: 10.1109/T-AFFC.2010.1.

[3] R. W. . Picard, *Affective computing*. MIT Press, 2000.

[4] L. F. Barrett, B. Mesquita, K. N. Ochsner, and J. J. Gross, "The Experience of Emotion," *Annu Rev Psychol*, vol. 58, no. 1, pp. 373–403, Jan. 2007, doi: 10.1146/annurev.psych.58.110405.085709.

[5] R. W. Picard, "Affective computing: Challenges," *International Journal of Human Computer Studies*, vol. 59, no. 1–2, pp. 55–64, 2003, doi: 10.1016/S1071-5819(03)00052-1.

[6] J. Nam, H. Chung, Y. ah Seong, and H. Lee, "A New Terrain in HCI: Emotion Recognition Interface using Biometric Data for an Immersive VR Experience," Dec. 2019, [Online]. Available: http://arxiv.org/abs/1912.01177

[7] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition: Review of sensors and methods," Feb. 01, 2020, *MDPI AG*. doi: 10.3390/s20030592.

[8] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. R. Wróbel, "Emotion Recognition and Its Applications," 2014, *Springer Verlag*. doi: 10.1007/978-3-319-08491-6_5.

[9] J. Gerlings and A. Shollo, *Reviewing the Need for Explainable Artificial Intelligence (xAI)*. [Online]. Available: https://hdl.handle.net/10125/70768

[10] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, "Review on Psychological Stress Detection Using Biosignals," *IEEE Trans Affect Comput*, vol. 13, no. 1, pp. 440–460, Jan. 2022, doi: 10.1109/TAFFC.2019.2927337.

[11] J. A. Russell, "A circumplex model of affect.," *J Pers Soc Psychol*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980, doi: 10.1037/h0077714.

[12] P. Ekman, "An argument for basic emotions," *Cogn Emot*, vol. 6, no. 3–4, pp. 169–200, May 1992, doi: 10.1080/02699939208411068.

[13] P. S. Sreeja and G. S. Mahalakshmi, "Emotion Models: A Review," *International Journal of Control Theory and Applications*, vol. 10, no. 8, pp. 651–657, 2017.

[14] Y. Wang *et al.*, "A Systematic Review on Affective Computing: Emotion Models, Databases, and Recent Advances," Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.06935

[15] R. PLUTCHIK, "A GENERAL PSYCHOEVOLUTIONARY THEORY OF EMOTION," in *Theories of Emotion*, Elsevier, 1980, pp. 3–33. doi: 10.1016/B978-0-12-558701-3.50007-7.

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

[17] P. Ekman and W. V. Friesen, “Facial Action Coding System,” Jan. 14, 2019. doi: 10.1037/t27734-000.

[18] S. Basu, J. Chakraborty, A. Bag, and Md. Aftabuddin, “A review on emotion recognition using speech,” in *2017 International Conference on Inventive Communication and Computational Technologies (ICICCT)*, IEEE, Mar. 2017, pp. 109–114. doi: 10.1109/ICICCT.2017.7975169.

[19] S. G. Koolagudi and K. S. Rao, “Emotion recognition from speech: A review,” Jun. 2012. doi: 10.1007/s10772-011-9125-1.

[20] N. Alswaidan and M. E. B. Menai, “A survey of state-of-the-art approaches for emotion recognition in text,” *Knowl Inf Syst*, vol. 62, no. 8, pp. 2937–2987, Aug. 2020, doi: 10.1007/s10115-020-01449-0.

[21] Y. Athavale and S. Krishnan, “Biosignal monitoring using wearables: Observations and opportunities,” Sep. 01, 2017, *Elsevier Ltd*. doi: 10.1016/j.bspc.2017.03.011.

[22] B. A. Fallow, T. Tarumi, and H. Tanaka, “Influence of skin type and wavelength on light wave reflectance,” *J Clin Monit Comput*, vol. 27, no. 3, pp. 313–317, Jun. 2013, doi: 10.1007/s10877-013-9436-7.

[23] M. Rinkevičius *et al.*, “Photoplethysmogram Signal Morphology-Based Stress Assessment,” in *2019 Computing in Cardiology Conference (CinC)*, Computing in Cardiology, Dec. 2019. doi: 10.22489/cinc.2019.126.

[24] W. Cui, L. E. Ostrander, and B. Y. Lee, “In Vivo Reflectance of Blood and Tissue as a Function of Light Wavelength,” 1990.

[25] C. Park and B. Lee, “Real-time estimation of respiratory rate from a photoplethysmogram using an adaptive lattice notch filter,” *Biomed Eng Online*, vol. 13, no. 1, 2014, doi: 10.1186/1475-925X-13-170.

[26] F. Esgalhado, A. Batista, V. Vassilenko, S. Russo, and M. Ortigueira, “Peak Detection and HRV Feature Evaluation on ECG and PPG Signals,” *Symmetry (Basel)*, vol. 14, no. 6, Jun. 2022, doi: 10.3390/sym14061139.

[27] K. Kalinkov and V. Markova, “Preprocessing of PPG and EDA signals for detection of emotional and cognitive states via physiological signals”, doi: 10.29114/ajtuv.vol6.iss1.253.

[28] S. K. D’Mello and J. Kory, “A review and meta-analysis of multimodal affect detection systems,” Apr. 16, 2015, *Association for Computing Machinery*. doi: 10.1145/2682899.

[29] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, “A review of affective computing: From unimodal analysis to multimodal fusion,” *Information Fusion*, vol. 37, pp. 98–125, Sep. 2017, doi: 10.1016/j.inffus.2017.02.003.

[30] A. Gandhi, K. Adhvaryu, S. Poria, E. Cambria, and A. Hussain, “Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions,” *Information Fusion*, vol. 91, pp. 424–444, Mar. 2023, doi: 10.1016/j.inffus.2022.09.025.

[31] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal Machine Learning: A Survey and Taxonomy,” Aug. 2017, [Online]. Available: http://arxiv.org/abs/1705.09406

[32] P. J. Bota, C. Wang, A. L. N. Fred, and H. Placido Da Silva, “A Review, Current Challenges, and Future Possibilities on Emotion Recognition Using Machine Learning and Physiological Signals,” *IEEE Access*, vol. 7, pp. 140990–141020, 2019, doi: 10.1109/ACCESS.2019.2944001.

[33] J. Zhang, Z. Yin, P. Chen, and S. Nichele, “Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review,” *Information Fusion*, vol. 59, pp. 103–126, Jul. 2020, doi: 10.1016/j.inffus.2020.01.011.

[34] S. M. S. A. Abdullah, S. Y. A. Ameen, M. A. M. Sadeeq, and S. Zeebaree, “Multimodal Emotion Recognition using Deep Learning,” *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 73–79, May 2021, doi: 10.38094/jastt20291.

[35] J. Gao, P. Li, Z. Chen, and J. Zhang, “A survey on deep learning for multimodal data fusion,” May 01, 2020, *MIT Press Journals*. doi: 10.1162/neco_a_01273.

[36] H. Kaya, F. Gürpınar, and A. A. Salah, “Video-based emotion recognition in the wild using deep transfer learning and score fusion,” *Image Vis Comput*, vol. 65, pp. 66–75, Sep. 2017, doi: 10.1016/j.imavis.2017.01.012.

[37] A. Jet and H. J. O, “Supervised Machine Learning Algorithms: Classification and Comparison,” *International Journal of Computer Trends and Technology*, vol. 48, 2017, [Online]. Available: http://www.ijcttjournal.org

[38] Κ. Διαμαντάρας and Δ. Μπότσης, *ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ*. ΚΛΕΙΔΑΡΙΘΜΟΣ, 2019.

[39] B. Rim, N.-J. Sung, S. Min, and M. Hong, “Deep Learning in Physiological Signal Data: A Survey,” *Sensors*, vol. 20, no. 4, p. 969, Feb. 2020, doi: 10.3390/s20040969.

[40] S. Oh, J.-Y. Lee, and D. K. Kim, “The Design of CNN Architectures for Optimal Six Basic Emotion Classification Using Multiple Physiological Signals,” *Sensors*, vol. 20, no. 3, p. 866, Feb. 2020, doi: 10.3390/s20030866.

[41] 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, Aug. 2020, doi: 10.3390/s20164551.

[42] L. Tarantino, P. N. Garner, and A. Lazaridis, “Self-attention for speech emotion recognition,” in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association, 2019, pp. 2578–2582. doi: 10.21437/Interspeech.2019-2822.

[43] Y. Bengio, A. Courville, and P. Vincent, “Representation Learning: A Review and New Perspectives,” Apr. 2014, [Online]. Available: http://arxiv.org/abs/1206.5538

[44] S. M. Montgomery, N. Nair, P. Chen, and S. Dikker, “Introducing EmotiBit, an open-source multi-modal sensor for measuring research-grade physiological signals,” *Science Talks*, vol. 6, p. 100181, May 2023, doi: 10.1016/j.sctalk.2023.100181.

[45] S. M. Montgomery, N. Nair, P. Chen, and S. Dikker, “Validating EmotiBit, an open-source multi-modal sensor for capturing research-grade physiological signals from anywhere on the body,” *Measurement: Sensors*, vol. 32, p. 101075, Apr. 2024, doi: 10.1016/j.measen.2024.101075.

[46] P. D. Welch, “The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Aver. aging Over Short, Modified Periodograms Ik(fn) =-I A h (%) [a k,” 1967.

[47] C. Orphanidou, “Quality Assessment for the Photoplethysmogram (PPG),” 2018, pp. 41–63. doi: 10.1007/978-3-319-68415-4_3.

[48] J. Park, H. S. Seok, S. S. Kim, and H. Shin, “Photoplethysmogram Analysis and Applications: An Integrative Review,” Mar. 01, 2022, *Frontiers Media S.A.* doi: 10.3389/fphys.2021.808451.

[49] M. Wójcikowski and B. Pankiewicz, “Photoplethysmographic time-domain heart rate measurement algorithm for resource-constrained wearable devices and its implementation,” *Sensors (Switzerland)*, vol. 20, no. 6, Mar. 2020, doi: 10.3390/s20061783.

[50] D. Pollreisz and N. TaheriNejad, “Detection and Removal of Motion Artifacts in PPG Signals,” *Mobile Networks and Applications*, vol. 27, no. 2, pp. 728–738, Apr. 2022, doi: 10.1007/s11036-019-01323-6.

[51] J. Kim, J. W. Lee, and H. Shin, “Pre-processing of Photoplethysmographic Waveform for Amplitude Regularization,” *Journal of Electrical Engineering and Technology*, vol. 14, no. 4, pp. 1741–1748, Jul. 2019, doi: 10.1007/s42835-019-00185-y.

[52] R. C. Ontiveros, M. Elgendi, G. Missale, and C. Menon, “Evaluating RGB channels in remote photoplethysmography: a comparative study with contact-based PPG,” *Front Physiol*, vol. 14, 2023, doi: 10.3389/fphys.2023.1296277.

[53] P. Virtanen *et al.*, “SciPy 1.0: fundamental algorithms for scientific computing in Python,” *Nat Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020, doi: 10.1038/s41592-019-0686-2.

[54] V. Jindal, J. Birjandtalab, M. Baran Pouyan, and M. Nourani, *An Adaptive Deep Learning Approach for PPG-Based Identification*. 2016. doi: 10.0/Linux-x86_64.

[55] N. Pinheiro *et al.*, *Can PPG Be Used for HRV Analysis?* 2016. doi: 10.0/Linux-x86_64.

[56] I. O. Joudeh, A. M. Cretu, S. Guimond, and S. Bouchard, “Prediction of Emotional Measures via Electrodermal Activity (EDA) and Electrocardiogram (ECG) †,” *Engineering Proceedings*, vol. 27, no. 1, 2022, doi: 10.3390/ecsa-9-13358.

[57] A. Al-Nafjan and M. Aldayel, “Anxiety Detection System Based on Galvanic Skin Response Signals,” *Applied Sciences (Switzerland)*, vol. 14, no. 23, Dec. 2024, doi: 10.3390/app142310788.

[58] I. Jeong *et al.*, “Machine learning in biosignal analysis from wearable devices,” *Mater Horiz*, 2025, doi: 10.1039/D5MH00451A.

[59] C. R. Harris *et al.*, “Array programming with NumPy,” *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020, doi: 10.1038/s41586-020-2649-2.

[60] 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,” *Biomed Signal Process Control*, vol. 55, Jan. 2020, doi: 10.1016/j.bspc.2019.101646.

[61] Z. Zhu *et al.*, “An emotion recognition method based on frequency-domain features of PPG,” *Front Physiol*, vol. 16, 2025, doi: 10.3389/fphys.2025.1486763.

[62] M. A. Almarshad, M. S. Islam, S. Al-Ahmadi, and A. S. Bahammam, “Diagnostic Features and Potential Applications of PPG Signal in Healthcare: A Systematic Review,” Mar. 01, 2022, *MDPI*. doi: 10.3390/healthcare10030547.

[63] M. Brennan, M. Palaniswami, and P. Kamen, “Do existing measures of Poincare plot geometry reflect nonlinear features of heart rate variability?,” *IEEE Trans Biomed Eng*, vol. 48, no. 11, pp. 1342–1347, 2001, doi: 10.1109/10.959330.

[64] H. Posada-Quintero, “Electrodermal Activity: What it can Contribute to the Assessment of the Autonomic Nervous System,” Dec. 2016.

[65] P. van Gent, H. Farah, N. van Nes, and B. van Arem, “HeartPy: A novel heart rate algorithm for the analysis of noisy signals,” *Transp Res Part F Traffic Psychol Behav*, vol. 66, pp. 368–378, Oct. 2019, doi: 10.1016/j.trf.2019.09.015.

[66] F. Pedregosa FABIANPEDREGOSA *et al.*, “Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot,” 2011. [Online]. Available: http://scikit-learn.sourceforge.net.

[67] A. Paszke *et al.*, “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” 2019.

[68] J. Nickolls, I. Buck, M. Garland, and K. Skadron, “Scalable parallel programming with CUDA,” in *ACM SIGGRAPH 2008 classes*, New York, NY, USA: ACM, Aug. 2008, pp. 1–14. doi: 10.1145/1401132.1401152.

APPENDIX A : DEEP LEARNING MODEL AND SEED INITIALIZATION CODE

```
# Seed

SEED = 21
random.seed(SEED)
np.random.seed(SEED)
torch.manual_seed(SEED)
if torch.cuda.is_available():
    torch.cuda.manual_seed_all(SEED)
cudnn.deterministic = True
cudnn.benchmark = False
```

```
# Branch CNN class
class BranchCNN(nn.Module):
    def __init__(self, *args, **kwargs):
        super().__init__()

        self.conv1 = nn.Conv1d(in_channels=1, out_channels=320, kernel_size=3, padding='same')
        self.bn1 = nn.BatchNorm1d(320)
        self.pool1 = nn.MaxPool1d(kernel_size=2)

        self.conv2 = nn.Conv1d(in_channels=320, out_channels=480, kernel_size=3, padding='same')
        self.bn2 = nn.BatchNorm1d(480)
        self.pool2 = nn.MaxPool1d(kernel_size=2)

        self.conv3 = nn.Conv1d(in_channels=480, out_channels=640, kernel_size=3, padding='same')
        self.bn3 = nn.BatchNorm1d(640)
        self.pool3 = nn.MaxPool1d(kernel_size=2)

        self.conv4 = nn.Conv1d(in_channels=640, out_channels=960, kernel_size=3, padding='same')
        self.bn4 = nn.BatchNorm1d(960)
        self.pool4 = nn.MaxPool1d(kernel_size=2)

        self.conv5 = nn.Conv1d(in_channels=960, out_channels=1280, kernel_size=3, padding='same')
        self.bn5 = nn.BatchNorm1d(1280)
        self.global_pool = nn.AdaptiveAvgPool1d(output_size=1)


    def forward(self, x):
        x = x.unsqueeze(1)  # one initial "feature map"

        x = F.relu(self.bn1(self.conv1(x)))
        x = self.pool1(x)

        x = F.relu(self.bn2(self.conv2(x)))
        x = self.pool2(x)

        x = F.relu(self.bn3(self.conv3(x)))
        x = self.pool3(x)

        x = F.relu(self.bn4(self.conv4(x)))
        x = self.pool4(x)

        x = F.relu(self.bn5(self.conv5(x)))

        x = self.global_pool(x)

        x = x.squeeze(-1)  # return (batch, 64)
        return x


# Multimodal NN class
class MultimodalNN(nn.Module):
    def __init__(self, input_lengths, num_classes):
        super().__init__()

        self.branches = nn.ModuleList([
            BranchCNN()
            for _ in input_lengths
        ])

        total_features = len(input_lengths) * 1280

        self.fc1 = nn.Linear(total_features, 1024)

        self.fc2 = nn.Linear(1024, 768)

        self.fc3 = nn.Linear(768, 512)

        self.fc4 = nn.Linear(512, 256)

        self.fc5 = nn.Linear(256, 128)

        self.fc6 = nn.Linear(128, num_classes)

    def forward(self, inputs):
        features = []

        for branch, x in zip(self.branches, inputs):
            features.append(branch(x))  # -> (batch_size, branch_channels)

        x = torch.cat(features, dim=1)  # -> (batch_size, total_features)

        x = F.relu(self.fc1(x))

        x = F.relu(self.fc2(x))

        x = F.relu(self.fc3(x))

        x = F.relu(self.fc4(x))

        x = F.relu(self.fc5(x))

        logits = self.fc6(x)  # -> (batch_size, num_classes)
        return logits
```

APPENDIX B : CODE USER GUIDE

This codebase is designed to support an emotion recognition pipeline using both classical machine learning and deep learning models. It includes scripts for data preprocessing, feature extraction, model training, and evaluation.

The scripts test.py and visualization.py are used internally to store utility functions and visual routines. They are not intended to be executed directly. In particular, test.py served as a sandbox environment during development for testing of different components across the pipeline.

To begin using the pipeline, users must first configure the correct paths in the paths_config.py script. These include the path to the raw data folder, the folder containing the response labels, the output folder for handcrafted features, and the output folder for preprocessed raw signals. These paths must be updated before any processing or model training can occur.

Once the paths are correctly set, the script processing.py should be run. This script handles all preprocessing tasks, including interpolation, signal alignment, missing data handling, and feature extraction. Feature extraction is modularized into three files: ppg_methods.py for photoplethysmography signals, eda_methods.py for electrodermal activity, and therm_methods.py for temperature sensor data. After processing is complete, the system outputs aligned raw signal data and structured feature datasets suitable for classical models. Before executing any of the ML scripts, ensure that the feature files or aligned raw signals have been generated successfully by the processing step.

To run unimodal evaluations using classical machine learning models, the script unimodal.py can be executed. This script processes each modality individually, training three different models, Random Forest, K-Nearest Neighbors, and Support Vector Machines, on each. For early fusion evaluation, the feature_level.py script is used. It uses the concatenated features from all modalities and runs the same three models on this combined dataset to assess fusion performance.

The deep learning pipeline is handled through the cnn_model.py script. This script can be used to train a new model from scratch or test an existing model state file (.pth) on a fixed, seeded test set. Testing assumes a saved model file is found in the current directory. If set to train, a new model using predefined parameters will be trained and saved, along with the accuracy and loss curves.

To run the code, several Python libraries are required. The essential external dependencies include:

- torch, for building and training the deep learning model,
- sklearn, for classical machine learning algorithms, data splitting, and evaluation metrics,
- heartpy, for sampling rate detection,
- numpy and scipy, for numerical operations and signal processing routines,
- matplotlib, for plotting and visualizations.

These libraries should be installed in the Python environment before executing any scripts. A standard Python 3.x environment with pip can be used for installation.