

Course Code: Master of Engineering(Research)

Student ID:

January 2026

Multimodal-Informed Emotional Stress Recognition Based on ECG

Xinyang Zhang

School of Computer Science

Faculty of Engineering and Information Technology

University of Technology Sydney

NSW - 2007, Australia

Multimodal-Informed Emotional Stress Recognition Based on ECG

*A thesis submitted in fulfilment of the requirements
for the degree of*

Master of Engineering(Research)
in
Computer Science

by

Xinyang Zhang

Supervisor: Min Xu
Co-supervisor: Haimin Zhang

to

Electrical and Data Engineering
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW - 2007, Australia

January 2026

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, *Xinyang Zhang* declare that this thesis is submitted in fulfilment of the requirements for the award of Master of Engineering (Research), in the Faculty of Engineering and Information Technology at the University of Technology Sydney. This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. This document has not been submitted for qualifications at any other academic institution. This research was supported by an Australian Government Research Training Program (RTP) Scholarship doi.org/10.82133/C42F-K220.

SIGNATURE: Production Note:
 Signature removed prior to publication.

[Xinyang Zhang]

DATE: 02th February, 2026

PLACE: Sydney, Australia

PUBLICATIONS

PUBLICATIONS:

1. Zhang, X., Zhang, H., Xu, M. Multimodal Classification Algorithms in ECG-Based Emotional Stress Analysis: A Comprehensive Review. *Computers in Biology and Medicine*, under minor revision.
2. Zhang, X., Zhang, H., Xu, M. Multimodal-Informed Emotional Stress Recognition Based on ECG. To be submitted.

TABLE OF CONTENTS

Certificate of Original Authorship	i
Publications	iii
Table of Contents	v
List of Figures	vi
List of Tables	xi
Abstract	xiii
1 Introduction	1
1.1 Background and Motivation	1
1.2 Problem Statement	4
1.3 Research Objectives	5
1.4 Research Questions	7
1.5 Thesis Structure	9
2 Literature Review	11
2.1 Overview of Existing Research	11
2.2 Key Theories and Concepts	15
2.2.1 Physiological Basis of Emotional Stress	15
2.2.2 Modeling and Quantification of Emotional Stress	19
2.3 Gaps in the Literature	21
2.4 Relevance to the Current Study	25
3 Methodology	27
3.1 Research Design	28
3.2 Signal Processing and Model Architecture	29

3.2.1	Unified Signal Processing Pipeline	29
3.2.2	Overall Algorithm Framework	32
3.3	Model Training and Evaluation Protocol	46
3.3.1	Knowledge Distillation Framework	47
3.3.2	Training Procedure	52
3.3.3	Inference and Testing Protocol	58
3.3.4	Evaluation Metrics	64
3.4	Methodological Assumptions and Constraints	69
3.5	Chapter Summary	70
4	Datasets and Experimental Design	71
4.1	Public Dataset	71
4.2	Self-Collected Multimodal Stress Dataset	72
4.3	Data Organization and Usage Protocol	76
5	Results	77
5.1	Overview of Evaluation Tasks and Metrics	77
5.2	Overall Classification Performance	78
5.3	Effect of Knowledge Distillation	79
5.4	Confusion Matrix and Class-wise Performance	83
5.5	Training Dynamics and Convergence Behavior	85
6	Discussion	91
7	Conclusion	95
	Bibliography	99

LIST OF FIGURES

FIGURE	Page
2.1 Heart beat structure diagram.	16
2.2 Typical intramuscular EMG waveforms.	17

2.3	Example of a Skin Conductance Response (SCR) curve.	18
3.1	Unified signal processing pipeline diagram.	30
3.2	Overall framework of the proposed teacher student architecture with unified signal processing and knowledge distillation based training.	33
3.3	Architecture of the modality specific 1D Convolutional Neural Network encoder for physiological signal representation.	35
3.4	Transformer-based fusion architecture for multimodal physiological representations.	37
4.1	Overview of the experimental task design and trial structure. (a) Schematic illustration of the temporal structure within a single trial, including the pre-experimental rest, question presentation, answering phase, and post-experimental rest. (b) Organization of the training and testing phases, showing the distribution of mental arithmetic questions across different difficulty levels, including easy, medium, intermediate, and difficult conditions.	73
4.2	Illustration of multimodal physiological sensor placement.	74
5.1	Per subject classification accuracy comparison between the teacher model, the baseline student model, and the distilled student model on the WESADB two-class task.	80
5.2	Per subject classification accuracy comparison between the teacher model, the baseline student model, and the distilled student model on the WESADB four class task.	80
5.3	Per subject classification accuracy comparison between the teacher model, the baseline student model, and the distilled student model on the maus two-class task.	81
5.4	Per subject classification accuracy comparison between the teacher model, the baseline student model, and the distilled student model on the maus three-class task.	81
5.5	Per subject classification accuracy comparison between the teacher model, the baseline student model, and the distilled student model on the maus two-class task.	82
5.6	Per subject classification accuracy comparison between the teacher model, the baseline student model, and the distilled student model on the maus three-class task.	82

- 5.7 Row normalized confusion matrices illustrating class wise prediction performance under different classification settings on the WESAD dataset. (a) Confusion matrix of the distilled student model for the binary classification task, showing clear class separation with high diagonal dominance. (b) Confusion matrix of the distilled student model for the four class task, where increased off diagonal entries indicate higher classification difficulty under finer class granularity. (c) Confusion matrix of the baseline student model for the four class task, exhibiting more pronounced inter class confusion and reduced diagonal values compared with the distilled model. (d) Confusion matrix of the teacher model for the four class task, serving as a performance reference and demonstrating stronger class separation. All matrices are row normalized to highlight class specific error distributions and enable direct comparison of prediction behavior across models and task settings. 84
- 5.8 Row normalized confusion matrices of the distilled student model for multi class classification tasks across different datasets. (a) Confusion matrix for the four class task on the WESAD dataset, showing relatively strong diagonal dominance with moderate inter class confusion. (b) Confusion matrix for the four class task on the MyDB dataset, where increased off diagonal values indicate higher classification difficulty under more complex and noisy data conditions. (c) Confusion matrix for the three-class task on the MAUS dataset, illustrating intermediate class separation performance between WESAD and MyDB. All confusion matrices are row normalized to emphasize class wise error distribution and enable a direct comparison of prediction patterns across datasets with different characteristics and task definitions. 85

- 5.9 Training dynamics and convergence behavior of the evaluated models on the WESAD dataset under different classification settings. (a) Training loss curves for the binary classification task, illustrating the optimization process of the teacher, student, and distilled student models. (b) Validation accuracy curves for the binary classification task, showing the evolution of predictive performance across training epochs. (c) Training loss curves for the four class task, where increased task complexity leads to higher loss values and slower convergence. (d) Validation accuracy curves for the four class task, highlighting differences in convergence speed and final performance among the three models. In all cases, the distilled student model demonstrates smoother convergence and improved validation performance compared with the baseline student model, while the teacher model serves as a reference upper bound. 86
- 5.10 Validation accuracy curves for multi class classification tasks across different datasets. (a) Validation accuracy on the WESAD four class task, illustrating the convergence behavior of the teacher, student, and distilled student models under a public benchmark setting. (b) Validation accuracy on the MyDB four class task, where overall accuracy levels are reduced due to increased data complexity and noise. (c) Validation accuracy on the MAUS three-class task, showing intermediate convergence performance between WESAD and MyDB. Across all datasets, the distilled student model consistently achieves higher validation accuracy than the baseline student model throughout training, while the teacher model serves as a reference upper bound. 87
- 5.11 Training and validation loss curves for multi class classification tasks across different datasets. (a) Loss curves for the WESAD four class task, demonstrating smooth and stable optimization behavior for all evaluated models. (b) Loss curves for the MyDB four class task, where higher loss values and slower convergence reflect increased task difficulty. (c) Loss curves for the MAUS three-class task, exhibiting moderate convergence behavior between the WESAD and MyDB datasets. In all cases, the distilled student model shows consistently lower loss values compared with the baseline student model, indicating improved optimization stability during training. 88

LIST OF TABLES

TABLE	Page
2.1 Comparison between traditional and deep learning-based methods for multi-modal stress recognition.	13
2.2 Comparative summary of ECG, EMG, and EDA in emotional stress detection. . .	18
3.1 Comparative summary of ECG, EMG, and EDA signal processing and representation.	31
3.2 Structural summary of the multimodal teacher network.	38
3.3(a)Comparison of CNN encoders in Teacher and Student models (ECG branch). . .	43
3.3(b)Comparison of CNN encoders in Teacher and Student models (ECG branch) continued.	44
3.4 Compact structural comparison between Teacher and Student models.	45
3.5 Roles and data flow in the proposed knowledge distillation framework.	48
3.6 Dynamic mechanisms for stabilising knowledge distillation during student training.	50
3.7 Training settings and learning targets for different model configurations.	53
3.8 Dynamic mechanisms used to stabilize knowledge distillation during student training.	55
3.9 Unified inference and testing pipeline for teacher and student models.	61
3.10 Recorded output artifacts during inference and testing.	63
3.11 Evaluation procedure across training, validation, and testing stages.	68
5.1 Evaluation tasks overview.	78
5.2 Overall classification performance of the distilled model across datasets and tasks.	79
5.3 Statistical significance analysis of performance improvements achieved by knowledge distillation using paired t-tests.	89

LIST OF TABLES

ABSTRACT

Emotional stress recognition using physiological signals has attracted increasing attention due to its potential applications in mental health monitoring and wearable systems. Among various biosignals, electrocardiogram (ECG) offers strong practicality for continuous monitoring, but its discriminative capability is limited when used as a single modality under realistic conditions. While multimodal approaches can improve performance by leveraging complementary information, their reliance on multiple sensors constrains practical deployment.

This thesis proposes a training–inference paradigm that enhances ECG-based stress recognition while preserving deployability. Complementary physiological signals, including electrodermal activity (EDA) and electromyography (EMG), are utilized during training to provide enriched supervision, whereas ECG is retained as the sole modality during inference. A unified signal processing pipeline and a teacher–student knowledge distillation framework are developed to transfer multimodal knowledge into a compact ECG-based model.

The proposed approach is evaluated on multiple public datasets and a self-collected multimodal dataset under both binary and multi-class settings. Results demonstrate consistent improvements in classification performance, convergence stability, and robustness compared with ECG-only baselines.

Overall, this study shows that multimodal-informed training with knowledge distillation provides an effective solution to improve ECG-based stress recognition under practical deployment constraints.

INTRODUCTION

1.1 Background and Motivation

Emotional states and psychological stress are widely encountered in modern life and can affect individual well-being, cognitive performance, and decision-making processes [1]. Long-term or excessive stress is associated with a wide range of adverse outcomes, including mental health disorders, reduced work efficiency, and impaired physiological regulation [2]. As a result, reliable assessment and monitoring of emotional stress have attracted growing attention in various application domains, such as mental health support systems, human–computer interaction, and healthcare monitoring [3]. In these contexts, the ability to continuously and objectively capture stress-related information is considered essential to allow timely intervention and adaptive system responses.

Traditional approaches for assessing emotional states and psychological stress have mainly relied on self-report questionnaires, behavioral observations, or dialog analysis [4]. Although questionnaire-based methods are widely used in clinical and experimental settings, they are inherently subjective and are typically conducted offline, making them unsuitable for continuous or real-time monitoring. Behavioral and speech-based approaches, on the other hand, are highly sensitive to environmental conditions and contextual factors, such as lighting, noise, and social settings, which can significantly degrade their reliability in real-world applications. Moreover, these methods often require active user participation or explicit expression, limiting their applicability in scenarios where unobtrusive and passive monitoring is desired. These limitations highlight the need for more objective and

robust approaches to assessing emotional stress.

In contrast to subjective and behavior-dependent approaches, physiological signals provide a more objective and direct reflection of the internal states of the body [5]. Physiological measurements are closely regulated by the autonomic nervous system, which plays a crucial role in modulating emotional and stress responses [6]. As a result, changes in emotional states and psychological stress are often accompanied by measurable variations in physiological activities, and can be collected continuously and passively in daily settings. These characteristics make physiological signal-based approaches particularly suitable for long-term monitoring and real-world applications, and have motivated extensive research into the use of various biosignals, such as electrocardiogram (ECG), electrodermal activity (EDA), and electromyography (EMG), for the recognition of emotional stress [7].

Among the various physiological signals, ECG has received particular attention due to its practicality and widespread availability [8]. ECG acquisition technology has become increasingly mature and is now commonly integrated in wearable devices such as smart watches and chest straps, allowing unobtrusive and long-term monitoring in daily life [9]. Compared with other physiological measurements, ECG sensors are relatively low-cost and widely available, facilitating scalable data collection and deployment. These advantages make ECG a highly attractive modality for stress and emotion monitoring, especially in applications that require continuous assessment outside controlled laboratory environments.

Although multiple physiological signals such as EDA and EMG have been shown to provide complementary information for emotional stress recognition, this study adopts ECG as the sole modality at the inference stage due to considerations of practicality and deployment feasibility. ECG sensing technology is mature and widely available in wearable devices, enabling continuous and unobtrusive monitoring in real-world settings. In contrast, EDA and EMG measurements typically require stricter sensor placement and are more sensitive to motion and environmental disturbances, which limits their robustness and long-term usability outside controlled environments. At the same time, it is acknowledged that ECG provides only an indirect representation of emotional stress, which motivates the use of complementary physiological information during model training rather than during application. This design choice aims to balance physiological informativeness and practical feasibility by enhancing ECG-based stress recognition while maintaining a simple and deployable sensing configuration.

The relationship between ECG signals and emotional stress can be understood through the regulatory mechanisms of the autonomic nervous system. Cardiac activity is modu-

lated by the sympathetic and parasympathetic branches of the autonomic nervous system, both of which are highly sensitive to emotional and psychological stimuli [10]. Variations in stress and emotional states can alter autonomic balance, leading to observable changes in heart rate dynamics and cardiac rhythms that are reflected in ECG recordings [11]. Consequently, ECG signals provide indirect, yet meaningful physiological indicators of emotional stress. This physiological linkage forms the theoretical basis for the use of ECG data in emotion and stress recognition studies and has motivated extensive research into ECG-based affective computing.

However, there are some limitations on ECG signals in emotional stress recognition. Emotional information is not directly encoded in ECG signals. Instead, it is shown through subtle and indirect variations in cardiac activity, which can be easily influenced by confounding factors such as physical activity, individual fitness levels, and baseline cardiovascular conditions [12]. In addition, substantial inter-subject variability exists in ECG patterns, making it difficult for models trained on limited datasets to generalize effectively across individuals. These factors collectively constrain the discriminative power of single-modality ECG-based approaches, particularly in small scale or subject-independent scenarios.

To address the limitations of single ECG-based approaches, increasing attention has been directed to multimodal physiological signal analysis for emotional stress recognition [13]. By jointly considering multiple biosignals, such as ECG, EDA, and EMG, multimodal methods are able to capture complementary aspects of the physiological responses associated with emotional and stress states. Although the ECG primarily reflects cardiac dynamics, the EDA is closely related to sympathetic nervous system activity, and the EMG provides information on muscle tension and somatic responses. The integration of these heterogeneous signals enables a more comprehensive representation of emotional stress, and has been reported to improve recognition performance and robustness in various affective computing studies.

The above observations reveal a fundamental trade-off in current emotional stress recognition research. On the one hand, multimodal physiological approaches demonstrate clear advantages in capturing complementary stress-related information and improving recognition performance. On the other hand, the reliance on multiple sensors increases system complexity, data acquisition costs, and deployment difficulty, which limits their practicality in real-world and wearable applications. In contrast, single-modal ECG-based solutions offer superior feasibility, but suffer from constrained discriminative capability when used in isolation. This tension motivates the exploration of frameworks that can leverage the rich

information provided by multimodal physiological signals while retaining the simplicity and deployability of ECG-based inference. Addressing this challenge requires a careful reconsideration of how multimodal information can be effectively utilized without imposing additional burdens during practical use.

1.2 Problem Statement

In recent years, substantial progress has been made in emotional stress recognition using physiological signals, particularly through the adoption of multimodal approaches that integrate complementary biosignals [14]. By jointly modeling heterogeneous physiological responses, such methods have demonstrated improved recognition performance and robustness compared to single-modality solutions [15]. However, this performance gain is often achieved at the cost of increased system complexity, which requires multiple sensors, synchronized data acquisition, and more demanding computational resources [16]. As a consequence, there is a fundamental tension between recognition accuracy and practical deployability in current stress recognition systems, especially for real-world and wearable applications where simplicity and robustness are essential.

Despite its practicality and widespread adoption, the recognition of emotional stress based on ECG faces inherent limitations when relying on a single modality [17]. Emotional and stress-related information is not explicitly encoded in ECG signals but is reflected instead through subtle and indirect variations in cardiac dynamics, such as changes in rhythm and temporal patterns [11]. These variations are often weak and can easily be masked by confounding factors, including physical activity, individual fitness levels, baseline cardiovascular conditions, noise and motion artifacts [18, 19]. Moreover, substantial inter-subject variability exists in ECG characteristics, leading to highly subject-dependent patterns that limit the discriminative capability and generalization performance of single-modal ECG-based approaches, particularly in subject-independent or small-scale data scenarios.

To address the limited discriminative capability of single-modal ECG, many existing studies have turned to multimodal physiological signal analysis by incorporating additional biosignals such as EDA and EMG [20]. Using complementary information in different modalities, these approaches have reported improved recognition accuracy and robustness in controlled experimental settings. Nevertheless, reliance on multiple sensing modalities introduces practical challenges, including additional hardware requirements, precise sensor placement, reliable signal synchronization, and increased data acquisition

and computational costs [21–23]. These constraints significantly hinder large-scale data collection and limit the feasibility of deploying multimodal systems in continuous and unobtrusive wearable scenarios.

Beyond modality-related limitations, emotional stress recognition based on physiological signals is further constrained by limited data availability and substantial inter-subject variability [24–26]. Collecting high-quality labeled physiological data is time-consuming and resource-intensive, often resulting in datasets with relatively small sample sizes. When combined with subject-dependent physiological patterns, these factors increase the risk of overfitting and make it difficult for data-driven models to learn generalizable stress-related representations that remain reliable across unseen subjects or real-world conditions.

In addition to these challenges, there is a fundamental trade-off between recognition performance, model complexity, and practical deployability in emotional stress recognition systems. Recent advances in deep learning have enabled increasingly complex models to capture subtle temporal patterns from physiological signals; however, such improvements are often accompanied by increased computational cost, larger model size, and reduced interpretability. These characteristics are particularly problematic for wearable and real-time applications with strict constraints on computational resources, energy consumption, and system stability.

In summary, the limitations of single-modal ECG-based methods, the practical constraints of multimodal sensing systems, and the challenges posed by limited data availability and inter-subject variability form a tightly coupled set of obstacles in emotional stress recognition [27, 28]. Improvements in recognition performance are frequently achieved at the expense of increased system complexity, while approaches that prioritize simplicity and deployability tend to suffer from reduced discriminative capability. Addressing these intertwined challenges in a unified manner remains an open research problem and motivates the research objectives presented in the following section.

1.3 Research Objectives

The general objective of this research is to investigate effective strategies for emotional stress recognition using ECG signals under practical and constrained conditions. This study aims to improve the discriminative capability and robustness of ECG based stress recognition while maintaining suitability for wearable and real time applications. In particular, this research emphasizes a training and inference paradigm that allows richer physiological information to be exploited during model development, while ensuring that the recognition

process remains simplified and efficient when applied in real world scenarios. The focus is placed on methodological strategies that explicitly account for the gap between training conditions and practical use, rather than assuming identical sensing configurations across all stages [29]. Considering the limitations of single-mode ECG signals, the availability of complementary physiological information, and the constraints imposed by the complexity and scale of the system, this research seeks to improve recognition performance without increasing the sensing requirements or computational burden during application.

The first objective of this research is to examine how effectively emotional stress related information can be captured from ECG signals under realistic experimental conditions. This objective focuses on analyzing the extent to which meaningful representations can be derived from ECG signals despite their indirect association with psychological states and their susceptibility to noise and variability [30]. Particular attention is given to the characteristics of ECG representations that are relevant to stress recognition, including their consistency and sensitivity across different task conditions. By systematically studying ECG based representations and recognition performance, this research aims to establish a clear understanding of the factors that limit the stability and reliability of ECG based emotional stress recognition, thereby providing a principled foundation for subsequent modeling strategies.

The second objective of this research is to investigate the role of complementary physiological information in enhancing ECG based emotional stress recognition without increasing system complexity at the inference stage. Rather than constructing a fully multimodal recognition system, this objective focuses on examining how information from EDA and EMG can be utilized during model development to support and strengthen ECG based representations [31]. The emphasis is placed on understanding how complementary modalities can provide informative guidance during training, while remaining absent from the final recognition pipeline used in application scenarios. The aim is to study strategies that benefit from multimodal physiological signals while preserving the simplicity and practicality of ECG only recognition, which is critical for real world and wearable use.

The third objective of this research is to investigate the stability and reliability of ECG based emotional stress recognition under limited data conditions. Given the difficulty of collecting large scale labeled physiological datasets, this objective focuses on examining how structured training strategies and informative supervision can reduce overfitting and maintain consistent recognition performance when training data are constrained [32]. This objective reflects common practical conditions in physiological computing, where data availability is often limited by experimental cost and participant burden. By studying learn-

ing behavior and performance trends in data limited settings, this research aims to identify factors that contribute to stable ECG based stress recognition without relying on increased data volume or complex sensing configurations.

The fourth objective of this research is to examine the trade off between recognition performance and model complexity in ECG based emotional stress recognition. Although advanced modeling techniques can improve recognition accuracy, increased model complexity often leads to higher computational cost and reduced suitability for wearable and real time applications [33]. This objective focuses on analyzing how efficient modeling strategies can balance recognition performance and computational efficiency under practical constraints. The goal is not to maximize performance in isolation, but to understand how design choices influence the feasibility and reliability of stress recognition systems intended for realistic use.

Together, these objectives establish a coherent research framework that addresses the limitations of ECG based emotional stress recognition from the perspectives of discriminative capability, utilization of complementary physiological information, learning stability under constrained data conditions, and practical efficiency. Based on these objectives, a set of research questions is formulated to guide the subsequent investigation and structure the experimental analysis presented in later chapters.

1.4 Research Questions

Based on the research objectives defined in the previous section, this study formulates a set of research questions to guide the subsequent investigation. These research questions are designed to translate the overall research objectives into specific and examinable problems that clearly reflect the scope and focus of this work. By structuring the experimental design and analysis around these questions, this study aims to ensure a coherent and systematic evaluation of ECG based emotional stress recognition under practical constraints [34].

The first research question examines the extent to which ECG signals can support reliable emotional stress recognition under realistic experimental conditions. This question focuses on evaluating the discriminative capacity of ECG signals when there are practical constraints such as noise, physiological variability, and limited experimental control [35]. Rather than assuming an idealized data acquisition setting, this study seeks to understand how effectively stress related information can be represented and distinguished using ECG signals in conditions that more closely reflect real world applications. Particular attention is given to the consistency and robustness of ECG based representations across

different stress inducing task conditions, as well as to the limitations arising from the indirect relationship between cardiac activity and psychological stress states [36]. Addressing this question establishes a foundational understanding of both the potential and inherent constraints of ECG based emotional stress recognition, which provides an essential reference for subsequent investigations in this thesis.

The second research question investigates the role of complementary physiological information in enhancing ECG based emotional stress recognition when system complexity at the inference stage is constrained [37]. This question focuses on understanding whether and how the information from additional physiological signals can provide meaningful guidance during model development while ensuring that the final recognition process relies solely on ECG signals in application scenarios [38]. Rather than evaluating fully multimodal recognition systems, this study examines the contribution of multimodal information at the training stage and its influence on the quality and robustness of ECG based representations. Particular attention is given to the extent to which complementary physiological signals can support learning more informative and stable representations of stress related patterns without introducing additional sensing requirements during deployment. By addressing this question, the study aims to clarify the practical value of multimodal physiological information in a training and application setting that prioritizes simplicity, efficiency, and feasibility for use in the real world [39].

The third research question examines the stability and reliability of ECG based emotional stress recognition under limited data conditions that commonly arise in physiological data collection [40]. This question focuses on understanding how constrained training data influence learning behavior, performance consistency, and susceptibility to overfitting in ECG based stress recognition tasks. Rather than assuming access to large scale datasets, this study considers realistic experimental scenarios in which data availability is restricted by practical factors such as collection cost, participant burden, and labeling effort [41]. Particular attention is given to how different training strategies and forms of informative supervision affect the consistency of recognition performance in repeated experiments and task conditions [42]. By addressing this question, the study seeks to clarify the extent to which stable and reliable ECG based emotional stress recognition can be achieved in data limited settings, thereby providing insight into the robustness of learning approaches under realistic constraints.

The fourth research question explores the balance between recognition performance and model complexity in ECG based emotional stress recognition under practical constraints. This question focuses on understanding how increases in model complexity in-

fluence recognition accuracy, computational cost, and overall feasibility for wearable and real time applications [43]. Rather than pursuing performance improvements in isolation, this study examines the trade offs that arise when designing models intended for realistic use, where efficiency and reliability are critical considerations [44]. Particular attention is given to identifying whether competitive recognition performance can be achieved without incurring excessive computational overhead or sacrificing robustness. By addressing this question, the study aims to provide insight into how design choices affect the practicality of ECG based stress recognition systems in real world scenarios.

In summary, these research questions define the key aspects of ECG based emotional stress recognition that are examined in this thesis, including discriminative capability, the use of complementary physiological information, the stability of learning under limited data conditions and practical efficiency. The subsequent chapters are organized to systematically address these questions through methodological development and experimental analysis. An overview of the thesis structure is presented in the following section.

1.5 Thesis Structure

This thesis is organized to present a coherent progression from problem formulation to methodological development and experimental evaluation. The overall structure is designed to systematically address the research objectives and research questions defined in the preceding sections, while maintaining a clear separation between conceptual background, methodological design, data description, and analytical findings. Through this organization, the thesis aims to provide a transparent and logical framework for investigating ECG based emotional stress recognition under practical constraints.

Chapter 2 reviews existing research on emotional stress recognition using physiological signals, with particular emphasis on ECG based approaches, multimodal learning strategies, and their practical limitations. Chapter 3 presents the methodological framework of this study, describing the overall modeling approach and training strategy adopted to support ECG based stress recognition under realistic constraints. Chapter 4 introduces the datasets used in this research, including publicly available datasets and a self collected task induced stress dataset, and explains the data collection procedures, preprocessing steps, and experimental usage. Chapter 5 reports the experimental results and analysis, addressing the research questions through systematic evaluation and comparison. Finally, Chapter 6 discusses the findings of the study, examines their implications in relation to the existing literature, and outlines the limitations and potential directions for future work.

The organization of the thesis is closely aligned with the research questions introduced in the preceding section, ensuring that each stage of the investigation directly contributes to addressing the defined research focus. The overall research workflow follows a structured progression from problem formulation and methodological development to data preparation, experimental evaluation, and subsequent discussion. Within this structure, the methodological framework and experimental analysis are explicitly designed to provide systematic responses to the research questions, each chapter contributing to specific aspects of the investigation. This organization enables a clear and traceable connection between the research questions, the experimental design, and the analytical outcomes presented throughout the thesis.

LITERATURE REVIEW

This chapter reviews existing research related to emotional stress recognition using physiological signals, with a particular focus on ECG based approaches and their practical applicability. The purpose of this review is to provide a structured overview of the current research landscape, clarify key theoretical concepts, and identify the limitations and open challenges relevant to this study. Rather than offering an exhaustive survey of all available methods, this chapter emphasizes representative studies and research trends that are most closely related to the objectives and research questions defined in Chapter 1. Through this review, the chapter establishes the conceptual and empirical foundation of the methodological framework presented in subsequent chapters.

2.1 Overview of Existing Research

Emotional stress recognition using physiological signals has become an active research area due to its relevance to mental health monitoring, human computer interaction, and wearable healthcare applications [45]. Physiological signals provide an objective means of capturing stress related responses, complementing subjective assessments such as questionnaires and self reports [46]. As a result, a wide range of studies have explored the use of biosignals to infer emotional and stress states, with increasing attention given to non-invasive and wearable friendly sensing modalities. Within this broader context, research has progressively shifted toward developing automated and data driven approaches that aim to detect stress under controlled and semi realistic experimental conditions.

Early research on emotional stress recognition using physiological signals has often focused on single modality approaches, motivated by considerations of sensing simplicity and practical feasibility. Among various biosignals, ECG has received particular attention due to its non invasive nature, relatively low acquisition cost, and suitability for long term wearable monitoring. Representative studies have examined the relationship between stress related physiological responses and ECG derived characteristics, typically aiming to distinguish stress from non stress states or to classify multiple stress levels under controlled experimental conditions. For example, He et al. used a CNN with Lomb–periodogram representations derived from ECG R-peak sequences for real-time acute cognitive-stress detection and demonstrated advantages over conventional HRV feature-based classifiers [47]. In addition, Cheema et al. proposed a psychological stress detection framework based on cardiac sound signals, combining EMD-based feature extraction with phase space reconstruction and an RBF-SVM classifier, and reported promising results in an exam-stress dataset [48]. In general, this line of work has established the ECG as a widely adopted modality in stress recognition, providing an important foundation for subsequent developments in both modeling techniques and experimental design.

Before the widespread adoption of deep learning techniques, a large body of research on emotional stress recognition relied on manually designed features extracted from physiological signals combined with classical machine learning classifiers. In the context of ECG based stress recognition, commonly used features include time domain and frequency domain cardiac descriptors, such as heart rate related measures and heart rate variability indicators, which are derived from domain knowledge to reflect stress induced changes in autonomic nervous system activity. Representative studies typically followed a pipeline of feature extraction, feature selection, and conventional classification using methods such as SVM, Naïve Bayes, or other classical models. For example, Cheema et al. proposed a framework for the detection of psychological stress based on cardiac-sound, using EMD with phase space reconstruction for feature extraction and an RBF-SVM classifier, and reported improved performance over a baseline ECG method LF/HF [49]. Similarly, Tanev et al. extracted linear and non-linear HRV features from the sternal ECG, selected informative features via forward sequential search, and used a leave-one-out Naïve Bayesian classifier for multi-stage acute-stress classification [50]. Overall, feature based approaches played an important role in early stress recognition studies by providing interpretable representations and relatively low computational complexity, although their effectiveness is often influenced by feature design choices, preprocessing quality, and inter-subject variability across different experimental conditions.

With the increasing availability of physiological datasets and advances in computational resources, deep learning based approaches have been progressively adopted for emotional stress recognition [51]. Unlike feature based methods that rely on manually designed descriptors, deep learning models aim to learn discriminative representations directly from raw or minimally processed physiological signals. In ECG based stress recognition, various neural network architectures have been explored to capture temporal patterns and nonlinear relationships associated with stress induced physiological changes. These approaches have demonstrated improved modeling flexibility and reduced dependence on handcrafted feature design, enabling end to end learning frameworks under controlled experimental settings [52]. As a result, deep learning has become a prominent methodological direction in recent stress recognition research, particularly in scenarios where complex temporal dynamics and subtle signal variations need to be modeled.

Table 2.1: Comparison between traditional and deep learning-based methods for multi-modal stress recognition.

Approach	Feature method	Data scale	Interpretability	Fusion type	Alignment	Deployment cost	Generalization	Sensitivity
Machine Learning	Manual feature extraction	Can work with small datasets	High	Mostly early or decision-level fusion	Requires manual synchronization or preprocessing	Low computational resources required; suitable for edge devices	Relatively robust in low-variance tasks	More reliant on individual feature quality
Deep Learning	Automatic feature learning via deep networks	Requires large-scale, labeled data for optimal performance	Low; needs XAI tools such as SHAP or Grad-CAM	Supports mid-level fusion via attention or KAN mechanisms	Can model inter-modal dynamics and delays	High computational cost; lightweight models needed for deployment	Sensitive to overfitting without regularization or transfer learning	Better at capturing high-order correlations between modalities

To provide a clearer and more structured overview of the methodological characteristics discussed above, Table 2.1 summarizes key differences between traditional feature based machine learning approaches and deep learning based methods in physiological stress recognition. The comparison is organized around aspects that are frequently considered in the design and evaluation of stress recognition systems, including feature extraction strategy, data requirements, interpretability, fusion characteristics, deployment cost,

and sensitivity to generalization. These dimensions reflect not only differences in modeling techniques, but also practical considerations related to data availability, system complexity, and applicability across varying experimental conditions. Rather than evaluating performance outcomes, the table highlights representative methodological properties commonly reported in the existing literature, serving as a structured reference for understanding prevailing research trends.

To address the limited discriminative capability of single modality physiological signals, a substantial body of research has explored multimodal approaches for emotional stress recognition. By combining complementary biosignals, such as ECG, EDA, and EMG, these methods aim to capture different aspects of stress related physiological responses that may not be fully observable from a single signal source. Multimodal studies typically report improved robustness and representation richness under controlled experimental conditions, reflecting the potential benefits of integrating heterogeneous physiological information. For example, Zhang et al. proposed an end-to-end cross-modal attention based fusion model that aligns and integrates multimodal physiological signals while suppressing redundant or noisy information, and reported improved stress recognition performance on benchmark datasets [53]. More broadly, Mahato et al. reviewed advances in multimodal hybrid wearable sensors that simultaneously track biophysical and biochemical signals, and highlighted the importance of integrating AI driven fusion and predictive algorithms for real time and comprehensive health assessment [54]. However, the adoption of multimodal sensing also introduces additional system level considerations, including increased sensor requirements, signal alignment complexity, and higher computational demands. As a result, while multimodal approaches represent an important research direction in physiological stress recognition, their practical feasibility and deployment constraints remain closely tied to the specific application context.

Despite the methodological advances reported in existing studies, the translation of physiological stress recognition methods from controlled laboratory settings to real world applications remains a challenge. Many approaches are evaluated under carefully designed experimental protocols with limited environmental variability, which may not fully reflect the conditions encountered in daily life [55]. In practical scenarios, physiological signals are often affected by motion artifacts, sensor placement variability, and individual differences, leading to increased uncertainty in stress related patterns [56]. Furthermore, system level constraints, such as sensor availability, computational resources, and energy consumption, impose additional limitations on the deployment of complex recognition models in wearable and real time settings [57]. These practical considerations highlight the gap between

experimental performance and real world applicability in physiological stress recognition research.

In summary, existing research on physiological stress recognition has explored a wide range of methodological approaches, spanning single modality and multimodal signal analysis, feature based modeling, and deep learning frameworks. While these studies demonstrate the feasibility of using physiological signals, particularly ECG, for stress recognition, they also reveal notable challenges related to discriminative capability, data dependency, system complexity, and practical deployment. To better understand these challenges and to clarify the assumptions underlying different research directions, it is necessary to examine the key theoretical foundations and conceptual frameworks that support physiological stress recognition. Accordingly, the next section introduces the essential theories and concepts that underpin the methods and experimental designs discussed in the literature.

2.2 Key Theories and Concepts

Building upon the overview of existing research presented in the previous section, this section introduces key theories and concepts that underpin physiological stress recognition. The purpose of this section is not to review specific algorithms, but to clarify the foundational assumptions and conceptual frameworks that guide the interpretation, modeling, quantification, and evaluation of stress related physiological signals. To this end, the section is organized into two main parts. First, the physiological basis of emotional stress and the fundamental characteristics of physiological signals are discussed to explain how stress related responses are indirectly reflected in biosignals. Second, common modeling paradigms and approaches to stress formulation and quantification adopted in stress recognition research are introduced to outline how stress related information is typically represented, inferred, and assessed. Together, these concepts provide the theoretical foundation for the methodological framework developed in the following chapter.

2.2.1 Physiological Basis of Emotional Stress

Emotional stress activates the autonomic nervous system, leading to measurable physiological changes in multiple bodily systems, including cardiovascular regulation, muscle activity, and sweat gland secretion. These responses can be captured through non-invasive physiological signals such as ECG, EMG, and EDA, which reflect complementary aspects of sympathetic and parasympathetic modulation. Compared to subjective assessments,

physiological signals provide continuous, real-time, and objective observations of stress-related responses. However, these signals do not directly encode psychological states; instead, they serve as indirect proxies whose interpretation depends on the underlying physiological mechanisms and contextual factors. Understanding the physiological basis and signal characteristics of these modalities is, therefore, essential for interpreting stress-related patterns and for motivating subsequent modeling and quantification strategies.

ECG is one of the most widely used physiological modalities for stress assessment due to its sensitivity to autonomic nervous system regulation and its compatibility with wearable monitoring. Emotional stress is commonly associated with increased sympathetic activation and reduced parasympathetic influence, resulting in changes in heart rate dynamics and cardiac waveform morphology. Such effects have been reflected in alterations in heart rate variability indices as well as variations in PQRST components, including QRS duration and QT interval [11, 58, 59]. Although certain frequency-domain measures remain subject to methodological debate [60], they continue to be widely applied as non-invasive indicators of stress-related autonomic shifts [61]. Representative ECG waveform structures and stress-related variations are illustrated in Figure 2.1, highlighting the physiological rel-

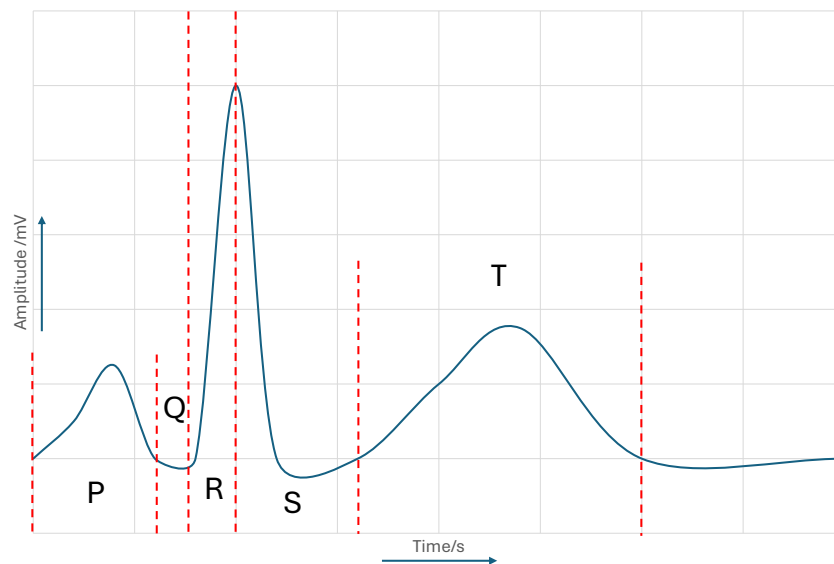


Figure 2.1: Heart beat structure diagram.

EMG captures the electrical activity of skeletal muscles and reflects neuromuscular responses influenced by autonomic and central nervous system activation during emotional stress. Increased sympathetic arousal is often associated with elevated muscle tension and altered motor unit firing behavior, which can be observed as changes in EMG amplitude

and spectral characteristics, particularly in facial and upper limb muscles [62, 63]. These responses provide useful indicators of sustained physiological arousal and have been linked to stress and anxiety-related conditions [64]. Typical EMG waveform patterns are shown in Figure 2.2. Nevertheless, EMG signals are highly sensitive to motion artifacts, electrode placement, and inter-individual variability, which can limit their standalone reliability in stress recognition tasks [65].

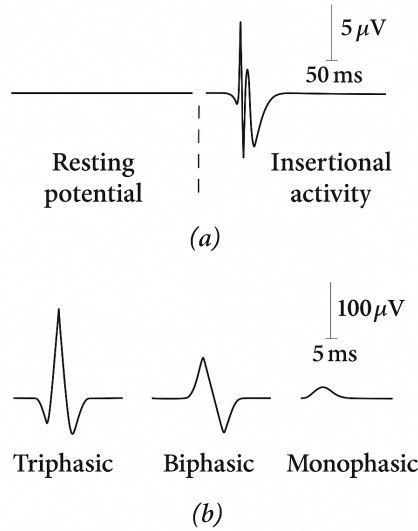


Figure 2.2: Typical intramuscular EMG waveforms.

EDA reflects changes in skin conductance driven by sweat gland activity under sympathetic nervous system control and is particularly sensitive to emotional and cognitive arousal [66]. Stress-related activation typically manifests itself as variations in the level of tonic skin conductance and transient phasic responses, characterized by features such as amplitude, rise time, and recovery rate [67]. An example of a skin conductance response is shown in Figure 2.3. Although EDA provides rapid and sensitive indicators of sympathetic activation, it is also affected by environmental conditions and individual differences in skin properties, which can reduce robustness when used in isolation [68].

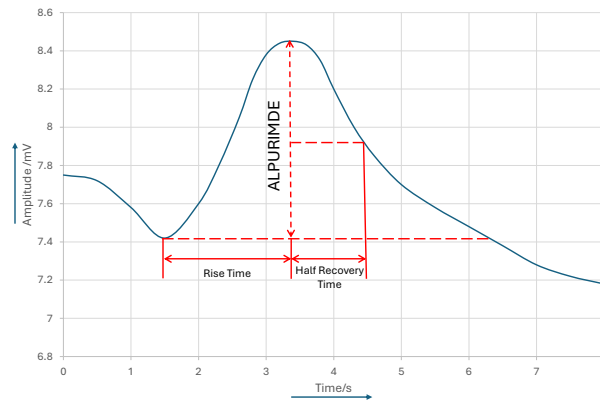


Figure 2.3: Example of a Skin Conductance Response (SCR) curve.

Although ECG, EMG, and EDA capture different physiological responses to emotional stress, none of these modalities alone can fully represent the complex relationship between psychological states and bodily reactions. ECG primarily reflects cardiovascular regulation, EMG captures muscle activation patterns, and EDA indexes sympathetic arousal, yet each modality is constrained by noise, non-stationarity, and sensitivity to contextual and individual factors [69, 70]. Table 2.2 provides a comparative summary of the key characteristics, advantages, and limitations of these signals in stress recognition contexts, illustrating their complementary functions and motivating their combined use in multimodal systems.

Table 2.2: Comparative summary of ECG, EMG, and EDA in emotional stress detection.

Dimension	Electrocardiogram (ECG)	Electromyography (EMG)	Electrodermal Activity (EDA)
Key Feature Indicators	HRV (SDNN, LF/HF), QTc, PQRST morphology	RMS, mean frequency, facial/limb MUAP waveforms	SCL, SCR amplitude, rise time, recovery time
Stress-Related Changes	HRV reduction, waveform variation	Increased muscle tension, amplitude changes	Increased SCR frequency and amplitude
Advantages	Clinically mature, wearable compatible	Direct reflection of muscle activation	High sensitivity to sympathetic activity
Limitations	Motion artifacts, inter-individual variability	Electrode sensitivity, noise	Environmental and skin-property dependence

In general, the physiological signals discussed above provide the empirical basis for emotional stress recognition, but their indirect nature and inherent variability impose fundamental constraints on how stress-related information can be interpreted and quantified.

These characteristics underscore the need for appropriate modeling assumptions, stress formulations, and evaluation strategies beyond raw signal observation. Consequently, the next subsection introduces common modeling and quantification paradigms that formalize emotional stress for computational inference and performance assessment.

2.2.2 Modeling and Quantification of Emotional Stress

Building on the physiological foundations discussed in the previous subsection, this subsection focuses on how emotional stress is modeled and quantified in computational stress recognition research. As emotional stress cannot be observed or measured directly, modeling choices play a central role in translating physiological responses into interpretable output. Rather than describing specific algorithms, this subsection introduces common conceptual frameworks that underlie stress formulation, supervised learning assumptions, label construction, and performance evaluation. These concepts establish a shared understanding of how stress related information is represented and assessed, and they provide the conceptual basis for the methodological design presented in the following chapter.

In computational stress recognition, emotional stress is generally regarded as a latent construct rather than a directly observable quantity [71]. Unlike physiological signals that can be measured through sensors, stress represents an internal psychological state that must be inferred from indirect observations [72]. Physiological responses such as changes in cardiac activity, muscle tension, or skin conductance reflect autonomic and neurophysiological processes associated with stress, but they do not constitute explicit measurements of stress itself. As a result, stress recognition relies on the assumption that consistent and informative relationships exist between underlying psychological states and observable physiological patterns. This indirect nature highlights the importance of clearly defining how stress is conceptualized before it is modeled or quantified.

Given that emotional stress is treated as a latent construct, computational stress recognition systems must adopt explicit formulations to represent stress in a form that can be learned and inferred. In existing research, stress is commonly formulated using discrete categorical labels, ordinal levels, or continuous-valued representations, depending on the experimental design and application context [73, 74]. Binary formulations often distinguish between stress and non-stress conditions, providing a simplified abstraction that facilitates robust classification under limited data conditions [75]. Multi-level or ordinal formulations aim to capture graded stress intensity by defining ordered categories, such as low, moderate, and high stress, while continuous formulations seek to model stress as a continuous variable that reflects gradual variations in psychological or physiological load [76].

Each formulation entails different modeling assumptions regarding the nature of stress responses, the separability of physiological patterns, and the interpretability of model output. Consequently, the choice of stress formulation directly influences how physiological signals are mapped to model predictions, how performance is quantified, and how results are interpreted within a given research setting.

Within most computational stress recognition studies, emotional stress modeling is commonly approached from a supervised learning perspective, in which physiological observations are mapped to predefined stress labels or scores [77]. In this paradigm, learning models aim to identify statistical regularities and discriminative patterns that differentiate stress-related states based on extracted physiological features. This formulation implicitly assumes that the available labels provide a consistent representation of the underlying psychological condition and that the observed physiological signals contain sufficient information to support a meaningful separation between stress categories or levels. In practice, supervised modeling further assumes a degree of stability in the relationship between physiological responses and stress states over time and experimental conditions [78]. Although these assumptions enable tractable learning and evaluation, they also shape how model outputs should be interpreted, as predictions primarily reflect learned statistical associations rather than direct measurements of emotional stress [79].

In supervised stress recognition, the construction of labels plays a central role in defining what the model is expected to learn and predict [80]. As emotional stress cannot be directly measured, ground truth labels are typically derived from indirect sources, such as experimental task design, self reported assessments, or protocol based segmentation of experimental sessions. These labels serve as operational representations of stress states and provide a reference against which model outputs are evaluated. Importantly, such labels reflect the assumed associations between experimental conditions, subjective experience, and physiological responses, rather than objective measurements of stress itself [81]. Consequently, label construction inherently embeds modeling assumptions regarding temporal alignment, stress intensity, and inter subject consistency, which influence both the learning process and the interpretation of recognition results.

Once emotional stress is formulated within a supervised learning framework and represented through constructed labels, model performance must be quantified using appropriate evaluation measures. In physiological stress recognition research, performance is commonly assessed through classification based metrics that summarize the degree of agreement between model predictions and reference labels [82]. Such metrics provide a standardized means of comparing different modeling approaches and experimental settings,

allowing researchers to evaluate whether learned representations can consistently distinguish between predefined stress states or levels. Importantly, these quantitative measures characterize model behavior with respect to the adopted stress formulation and labeling scheme, rather than offering direct validation of the underlying psychological construct [83]. As a result, performance evaluation should be understood as an assessment of predictive consistency and discriminative capability under given assumptions, rather than as a direct measure of stress intensity or physiological validity.

In general, modeling and quantification of emotional stress involve a series of conceptual choices that shape how physiological signals are interpreted and evaluated in computational stress recognition research. From treating stress as a latent construct, to selecting appropriate formulation strategies, adopting supervised learning assumptions, constructing reference labels, and quantifying predictive performance, each step introduces assumptions that influence both model behavior and result interpretation. These considerations establish a coherent conceptual framework for understanding how stress related information is represented and assessed prior to methodological implementation. Building on this framework, the following section examines gaps and open challenges in the existing literature, highlighting unresolved issues that motivate the methodological developments presented in this thesis.

2.3 Gaps in the Literature

Despite substantial progress in physiological stress recognition, existing studies continue to face unresolved challenges in balancing recognition performance and practical deployability. Previous research has demonstrated that both single-modality and multimodal approaches can achieve promising results in controlled experimental settings. However, improvements in recognition accuracy are often accompanied by increased system complexity, stronger sensing requirements, or heavier computational demands, which limit applicability in real-world and wearable scenarios. As a result, there remains a persistent gap between methodological advances reported in the literature and the practical constraints imposed by realistic stress monitoring applications. The following discussion identifies key gaps that hinder the effective translation of current stress recognition methods into deployable systems, with particular emphasis on ECG-based recognition under constrained inference conditions.

Although ECG-based stress recognition has demonstrated promising performance in several studies, its discriminative capability and robustness remain limited under realis-

tic conditions. Existing work has shown that combining handcrafted HRV features with deep representations extracted from ECG signals can yield high recognition accuracy in controlled settings. For example, Zhou et al. integrated GMM-processed HRV features with CNN-learned ECG representations and achieved high recognition performance in different stress states [84]. However, such results are often obtained using relatively small datasets and protocol-specific stress induction paradigms, which can introduce confounding factors and restrict generalizability. Similarly, Ramteke et al. reported a strong performance using an attention-guided dual-stream BiLSTM framework based on ECG-derived HRV and time-spectral sequences [85]. Despite the reported accuracy gains, this approach relies on pre-computed features and a relatively complex model design, raising concerns regarding end-to-end learning capability, overfitting under limited data, and suitability for real-time deployment. These examples illustrate while single-modality ECG-based methods can achieve high performance under favorable conditions, their stability, generalization, and practicality remain insufficiently addressed in the existing literature.

Multimodal physiological approaches have been widely explored to enhance stress recognition performance by leveraging complementary information across different modalities. Recent studies demonstrate that integrating heterogeneous signals and advanced model architectures can lead to notable accuracy gains in benchmark datasets. For instance, Yang et al. transformed multimodal physiological signals into image-based representations using time series encoding techniques and trained a two-dimensional convolutional network for multi-class stress recognition [86]. While this strategy achieved competitive performance, it relies on a multi-stage, non end-to-end pipeline that introduces additional processing complexity and may limit scalability and real-time applicability. Moreover, the integration of multiple modalities can introduce redundancy and noise, potentially leading to performance saturation when class imbalance and deployment constraints are not explicitly addressed. Similarly, Bhatt et al. proposed a multimodal fusion framework combining physiological parameters with visual information and ensemble-based classification to achieve high stress recognition accuracy [87]. Despite strong reported results, the approach depends heavily on complex sensing configurations and large model ensembles, while limited data availability and observed confusion between adjacent stress levels raise concerns regarding reproducibility, generalization, and practical deployment. These studies highlight that although multimodal methods can improve recognition performance, their increased algorithmic complexity and sensing requirements often hinder applicability in resource-constrained and real-world stress monitoring scenarios.

Beyond model architecture and modality selection, a notable gap in the literature lies

in how supervision is constructed and transferred during training, particularly when multimodal information is involved. Several studies attempt to exploit multimodal physiological signals to derive richer supervision signals; however, the resulting representations are not always robust or well aligned with fine-grained stress inference. For example, He et al. used baseline-normalized multimodal physiological features and unsupervised clustering to generate pseudo-labels for anxiety recognition, which were then used to train subject-independent classifiers [88]. Although this approach demonstrated moderate classification performance, the limited cluster separability and overlapping stress levels suggest that noisy supervision can constrain reliable discrimination and generalization. Similarly, Wang et al. proposed a multimodal fusion framework augmented with a symbolic regression based analyzer to enhance interpretability in stress prediction [89]. Despite the high reported accuracy, the method exhibits sensitivity to temporal window length and model complexity, and the inclusion of computationally expensive symbolic components raises concerns about scalability and real-time deployment. These studies indicate that although multimodal supervision can enrich training signals, existing approaches often lack a principled mechanism to decouple complex training supervision from efficient and stable inference objectives.

Recent studies have begun to explore cross-modal knowledge transfer as a means of leveraging rich multimodal information to improve stress recognition performance. Such approaches aim to align or distill representations across modalities in order to enhance learning under limited or heterogeneous data conditions. For instance, Gao et al. proposed a multimodal framework that integrates facial, EEG, and ECG signals through feature disentanglement, cross-modal Transformer alignment, and graph-based knowledge distillation, achieving strong recognition results on benchmark datasets [90]. However, the effectiveness of this approach is closely tied to a highly complex training pipeline involving pre-trained feature extractors and multiple loss functions, and its performance is strongly influenced by the dominant modality, particularly EEG. Moreover, reported ablation results suggest that the stability of cross-modal distillation may degrade under domain discrepancies, while the reliance on continuous facial video and multi-channel EEG acquisition limits practical wearable deployment. Similarly, Aravinth et al. introduced a dynamic cross domain transfer learning framework that combines multimodal signals with domain-adversarial learning and online personalization to improve stress recognition across domains [91]. Although promising results have been reported, the proposed architecture involves substantial implementation overhead and depends on high-quality multimodal sensing, while the inclusion of online fine-tuning raises concerns about stability, controllabil-

ity, and suitability for real-world applications. These studies indicate that, despite growing interest in cross-modal knowledge transfer, existing methods remain constrained by complexity and modality dependence, and lack a systematic pathway for translating multimodal training advantages into efficient and reliable ECG-only inference.

Another persistent gap in the literature concerns the lack of a unified approach to balance representation power, computational efficiency, and practical deployability in stress recognition models. Recent work has shown that compact deep architectures can achieve high recognition accuracy while meeting edge deployment constraints. For example, Cvetkovic et al. proposed a lightweight residual convolutional network optimized for edge devices through architectural simplifications and post-training quantization, achieving real-time inference with strong performance [92]. However, the reported results rely heavily on highly overlapped input windows, and performance degradation under aggressive quantization raises concerns regarding strict generalization validity and ultra-low-power deployment. In parallel, system-level frameworks have been proposed to standardize multimodal analysis and deployment pipelines. Li et al. introduced a low-code framework that enables reproducible multimodal modeling and executable deployment across multiple physiological tasks [93]. Despite its practical utility, the framework primarily supports offline processing and limited fusion strategies, and its modality coverage remains incomplete for real-time wearable stress monitoring. These studies illustrate a broader disconnect in the literature, where models that emphasize efficiency often sacrifice robustness or generalization, while systems that support complex modeling remain difficult to deploy in resource-constrained settings. As a result, a cohesive framework that jointly addresses model expressiveness, efficiency, and practical usability remains insufficiently explored.

In summary, the existing literature on physiological stress recognition reveals a set of interrelated gaps that limit the practical translation of current methods. Single-modality ECG-based approaches continue to face challenges in robustness and generalization under realistic conditions, while multimodal systems, although capable of improving recognition performance, often incur substantial sensing and computational overhead. Furthermore, current studies rarely provide a systematic mechanism to decouple complex multimodal training from efficient and stable inference, and cross-modal knowledge transfer strategies remain constrained by model complexity, modality dependence, and deployment feasibility. At the system level, the lack of unified treatment for model expressiveness, efficiency, and usability further exacerbates the gap between methodological advances and real-world application requirements. These unresolved issues collectively highlight the need for stress recognition frameworks that can effectively leverage rich multimodal information during

training while maintaining practical, efficient, and reliable inference. The following section outlines how the current study addresses these gaps and positions its contributions within the broader research landscape.

2.4 Relevance to the Current Study

Building upon the gaps identified in the preceding section, the current study is designed to address key challenges that limit the practical applicability of existing physiological stress recognition methods [94]. In particular, this research focuses on the tension between recognition performance and deployability, with an emphasis on ECG-based stress recognition under constrained inference conditions [95]. Rather than pursuing increased sensing complexity or fully multimodal inference, the study adopts a research perspective that prioritizes practical deployment while seeking to enhance representation quality through informed training strategies. This positioning directly responds to the limitations observed in previous work and establishes the rationale for the methodological choices presented in subsequent chapters.

The emphasis on ECG as the core modality during inference is motivated by practical considerations associated with real-world and wearable stress monitoring. Compared with other physiological signals, ECG offers a favorable balance between physiological relevance, sensing stability, and long-term wearability, making it suitable for continuous monitoring outside controlled laboratory environments [96]. Although additional modalities such as EMG or EDA can provide complementary information, their integration at the inference stage often introduces increased sensing burden, synchronization requirements, and susceptibility to motion or environmental interference. By focusing on ECG inference, the current study aligns its design with deployment constraints commonly encountered in wearable applications, where simplicity, robustness, and reliability are essential. This choice reflects an application-driven perspective that prioritizes feasible stress recognition under realistic operating conditions.

Given the limited discriminative capacity of single-modality ECG signals under realistic conditions, the current study adopts a research paradigm that differentiates between the roles of training and inference. Although ECG is retained as the sole modality in the inference stage, richer physiological information from complementary signals can provide valuable supervision during model development [97]. This approach enables the learning process to take advantage of cross-modal physiological signals and shared stress-related patterns without imposing additional sensing requirements during the application. This

perspective addresses a recurring limitation in existing studies, where multimodal information is typically coupled with both training and inference, leading to increased system complexity and reduced deployability. By conceptually separating multimodal supervision from practical inference, the study positions multimodal information as a training resource rather than a deployment dependency, thus reconciling the benefits of multimodal learning with the constraints of real-world stress monitoring.

Although this study includes the construction and use of a task-driven physiological dataset, its primary contribution lies in the development and examination of an algorithmic framework rather than in the expansion of the dataset. The role of the collected data is to support methodological validation and to provide complementary evidence alongside public benchmarks, rather than to serve as a large-scale or comprehensive resource. Given the practical constraints associated with physiological data acquisition, the study does not aim to exhaustively explore data diversity or population coverage. Instead, the experimental design is structured to assess how the proposed research perspective performs under controlled and reproducible conditions. This emphasis ensures that the study findings are interpreted in relation to algorithmic effectiveness and design choices, rather than being conflated with the scale or breadth of the collection of the dataset.

In summary, the design considerations outlined in this section establish a coherent research perspective to address practical challenges in ECG-based emotional stress recognition. By focusing on inference on ECG signals, leveraging complementary physiological information during training, and highlighting algorithmic design under realistic constraints, the current study aims to bridge the gap between methodological effectiveness and real-world applicability. This perspective directly informs the methodological choices adopted throughout the study and provides a structured rationale for the proposed framework. Based on these considerations, the following chapter presents the methodological approach developed to operationalize this research perspective, including model formulation, training strategy, and evaluation protocol.

METHODOLOGY

This chapter presents the methodological framework adopted in this study for ECG based stress recognition under practical deployment constraints [98]. The proposed methodology is designed to balance recognition performance with real world applicability by prioritizing robustness simplicity and inference efficiency. Rather than pursuing increasingly complex sensing configurations or fully multimodal inference pipelines, the method focuses on ECG as the primary modality during inference while leveraging complementary physiological signals during training to enhance representation learning [99].

The overall design follows a unified pipeline across multiple datasets including public benchmarks and a self collected database while allowing dataset specific adaptations where required by differences in acquisition protocols and temporal structure. Signal processing strategies model architectures and the training procedures are jointly designed to support this unified framework. In particular, a teacher student learning paradigm based on knowledge distillation is employed to transfer information learned from multimodal training data to a lightweight ECG only inference model [100]. This design enables an improved discriminative capability without increasing the sensing burden or inference complexity.

The remainder of this chapter details the research design signal processing and model architecture training and evaluation protocol as well as the methodological assumptions and constraints underlying the proposed approach.

3.1 Research Design

This study adopts an engineering oriented research design that focuses on methodological effectiveness under practical deployment constraints [101]. Rather than pursuing increasingly complex sensing configurations or laboratory specific solutions, the proposed approach emphasizes robustness simplicity and inference efficiency in real world wearable scenarios. The research problem is formulated around ECG based stress recognition, where the primary objective is to achieve reliable discrimination performance while maintaining a minimal sensing and inference footprint [102]. Consequently, the methodological design prioritizes algorithmic strategies that improve the quality of the representation without increasing the complexity of the system at the inference stage.

A key aspect of the research design is the deliberate distinction between the training stage and the inference stage in terms of modality usage [103]. During inference the method relies exclusively on ECG signals due to their favorable balance between physiological relevance sensing stability and long term wearability. This choice reflects practical considerations in real world deployment where minimizing sensor count and system complexity is essential. In contrast, additional physiological signals are incorporated only during training to provide complementary information that supports more informative representation learning [104]. By separating the roles of training and inference, the proposed design leverages multimodal information without imposing additional sensing or computational requirements at deployment time.

To support a robust evaluation of the proposed methodology, the research design incorporates multiple datasets drawn from both public benchmarks and a self collected database. These datasets differ in acquisition protocols temporal organization and signal characteristics, which reflects the diversity commonly encountered in practical applications. Rather than tailoring the method to a specific dataset a unified methodological framework is adopted across all data sources [105]. Dataset specific adaptations are introduced only where necessary to accommodate differences in signal structure and annotation schemes while the core processing pipeline model architecture and training strategy remain consistent. This design enables a systematic assessment of the proposed approach under varying data conditions while maintaining methodological coherence.

Based on the research design outlined above, the remainder of this chapter details the methodological components of the proposed approach. Section 3.2 describes the signal processing strategy and the overall model architecture. Section 3.3 presents the training procedure and evaluation protocol used to assess model performance. Section 3.4 dis-

cusses the methodological assumptions and constraints that define the scope of the study. This chapter concludes with a brief summary of the proposed methodology.

3.2 Signal Processing and Model Architecture

This section describes the signal processing strategy and model architecture that together form the core of the proposed ECG based stress recognition framework [106]. The design of this section emphasizes a clear separation between signal representation and learning architecture while maintaining a unified interface across all datasets [107]. Signal processing is formulated to produce standardized window level inputs that are compatible with the downstream models, whereas the model architecture is designed to support effective representation learning under constrained inference conditions. By jointly considering signal processing and architectural design, the proposed framework establishes a consistent and deployment oriented foundation for subsequent training and evaluation procedures.

3.2.1 Unified Signal Processing Pipeline

This subsection presents a unified signal processing pipeline that serves as the front end of the proposed ECG based stress recognition framework. The primary objective of this pipeline is to transform heterogeneous raw physiological recordings into a standardized representation that can be processed consistently by subsequent learning models. Given the diversity of data sources considered in this study, including datasets with different acquisition protocols temporal structures and annotation schemes, the signal processing stage is designed to abstract away dataset specific irregularities while preserving physiologically meaningful information. Rather than relying on dataset dependent heuristics, the pipeline emphasizes a principled organization of signals in the temporal domain to establish a common input interface. Through this design, the signal processing stage functions not only as a preparatory step but also as an integral methodological component that ensures coherence between data representation and model architecture across all experiments [108].

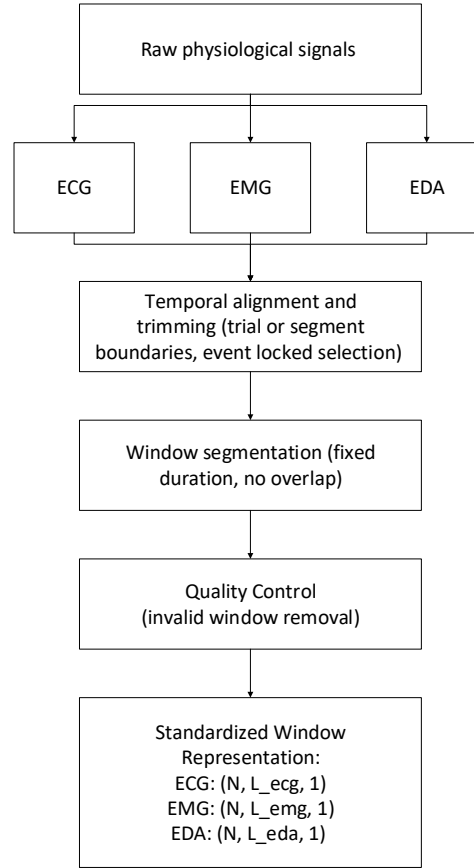


Figure 3.1: Unified signal processing pipeline diagram.

An overview of the unified signal processing pipeline is illustrated in Figure 3.1, which summarizes the sequential operations applied to raw physiological recordings prior to the input of the model. As shown in the diagram, raw signals are first organized according to their temporal alignment, with boundaries defined by trial or segment information and event locked selection used where applicable. The aligned signals are then segmented into fixed duration windows without overlap, forming the basic units for subsequent analysis. To ensure data reliability, a quality control step is applied to remove windows that contain invalid samples or violate predefined stability criteria. Following these steps, all retained windows are represented in a standardized window level format, which serves as the direct input to the downstream learning models. This structured processing flow establishes a consistent transformation from raw signals to model ready representations, enabling coherent integration of heterogeneous data sources within a unified framework.

Within the unified pipeline, window level representation is adopted as the fundamental unit for subsequent modeling. Each retained signal segment is expressed as a fixed length temporal window, which provides a consistent and bounded input structure for learning

based models [109]. This design enables efficient batch processing and stable optimization while avoiding the variability associated with variable length sequences. In addition, the use of non overlapping windows reduces redundancy between samples and limits temporal dependency across adjacent inputs, which is particularly important for preventing information leakage during subject level evaluation [110]. By representing signals at the window level, the pipeline establishes a balance between preserving local temporal dynamics and maintaining computational tractability, making the resulting representations suitable for both training and inference under constrained deployment settings.

Table 3.1: Comparative summary of ECG, EMG, and EDA signal processing and representation.

Dimension	ECG	EMG	EDA
Signal representation	Continuous ECG waveform segmented into fixed-length windows	Multi-channel EMG signals averaged and segmented into fixed-length windows	Continuous EDA (GSR) signal segmented into fixed-length windows
Sampling & alignment	Resampled and aligned to a unified temporal grid	Resampled and temporally aligned with ECG	Resampled and temporally aligned with ECG
Windowing strategy	Non-overlapping windows with enforced temporal boundaries	Same non-overlapping windowing strategy as ECG	Same non-overlapping windowing strategy as ECG
Temporal boundary control	Windows constrained within valid task phases and trials	Windows constrained to match ECG temporal boundaries	Windows constrained to match ECG temporal boundaries
Label assignment	Window-level labels derived from task phase annotations	Inherits window-level labels aligned with ECG	Inherits window-level labels aligned with ECG
Noise and artifact handling	Structural rejection of invalid samples; no explicit filtering	Structural rejection of invalid samples; no explicit filtering	Structural rejection of invalid samples; no explicit filtering
Missing data policy	Discarded if containing NaN, inf, or corrupted packets	Discarded if containing NaN or invalid values	Discarded if containing NaN or invalid values
Output format	Saved as unified window-level tensors with metadata	Saved as unified window-level tensors with metadata	Saved as unified window-level tensors with metadata

To further clarify the unified design principles across different signal modalities, Table 3.1 provides a comparative summary of the signal processing and representation strategies adopted for ECG, EMG, and EDA. As illustrated in the table, although these modalities differ in their raw signal characteristics and channel configurations, they are processed according to a shared set of structural constraints, including window level representation, non overlapping segmentation, temporal boundary enforcement, and standardized output formatting. Modality specific handling is limited to the initial organization and alignment stages, ensuring that all signals ultimately conform to a common window based interface. This design allows multimodal information to be incorporated during training without introducing heterogeneity at the model input level. By explicitly distinguishing unified processing principles from modality dependent adaptations, the proposed pipeline maintains methodological consistency while supporting flexible integration of heterogeneous physiological signals.

In summary, the unified signal processing pipeline establishes a consistent transformation from raw physiological recordings to standardized window level representations that are suitable for downstream learning models. By emphasizing temporal organization structural constraints and controlled adaptation, the pipeline ensures that variations across datasets and signal modalities are handled prior to model input without affecting the learning architecture. This design provides a stable and coherent interface between data and models, allowing subsequent algorithmic components to focus on representation learning rather than dataset specific irregularities. Building upon this unified input foundation, the next subsection introduces the overall algorithm framework that operates on these standardized representations.

3.2.2 Overall Algorithm Framework

The overall algorithm framework of the proposed approach is illustrated in Figure 3.2, which depicts the end to end information flow from preprocessed physiological signals to final classification outputs. As shown in the figure, the standardized window level representations produced by the unified signal processing pipeline serve as the common input interface for all subsequent components. The framework explicitly separates the training stage from the inference stage, reflecting the different functional requirements in each phase. During training, multimodal physiological signals are utilized to support richer representation learning, while during inference the system operates exclusively on ECG signals to ensure the feasibility of deployment. This system level organization clarifies how signal processing, model architecture, and stage specific operation are integrated into a coher-

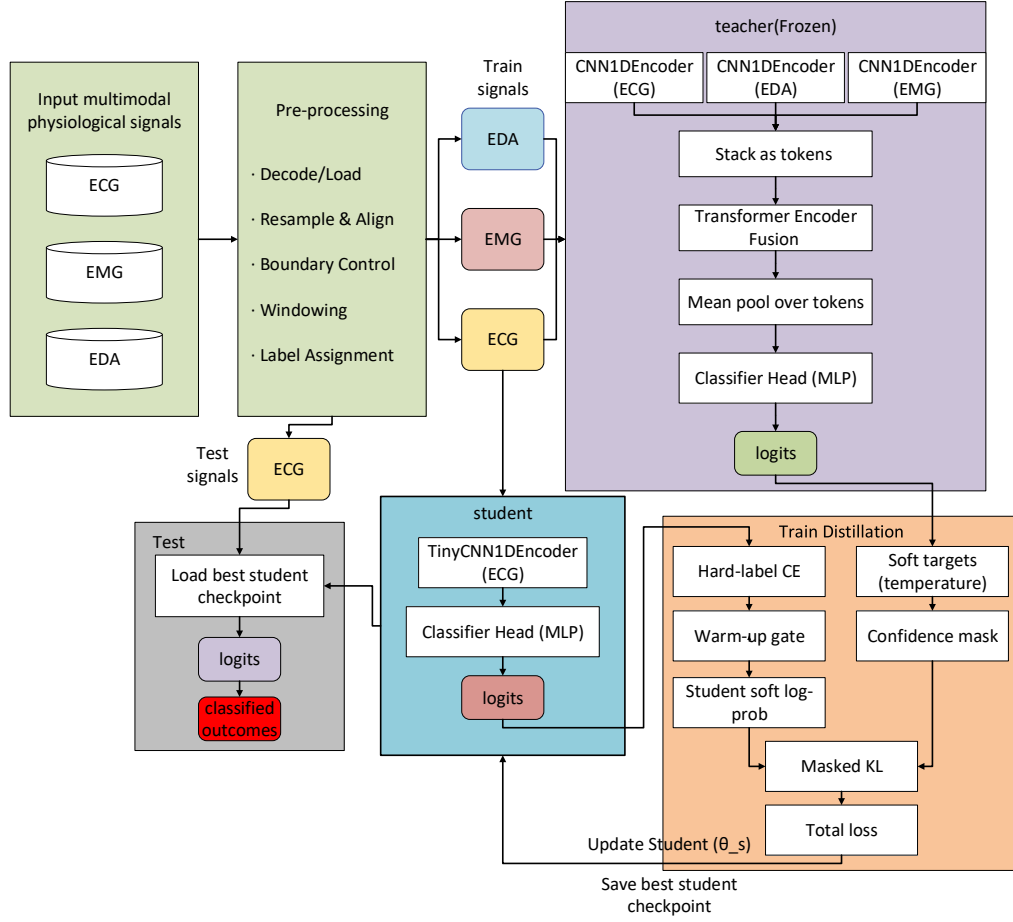


Figure 3.2: Overall framework of the proposed teacher student architecture with unified signal processing and knowledge distillation based training.

ent pipeline, providing the structural context for the detailed network designs described in the following subsections.

Within the proposed framework, teacher and student networks are designed to fulfill complementary roles that together balance representational capacity and deployment constraints. The teacher network operates exclusively during the training stage and is allowed to leverage multimodal physiological input to capture richer and more discriminative patterns that may not be fully observable from a single modality. By contrast, the student network is designed as a compact model that relies solely on ECG signals and serves as the final inference engine of the system. Although the two networks differ in input modalities and architectural complexity, they are aligned at the output level through a shared prediction space, which enables effective transfer of supervisory information during training. This explicit separation of roles allows the framework to exploit multimodal information when available while preserving a lightweight and practical configuration for real world in-

ference, thereby addressing the inherent trade off between performance and deployability in physiological stress recognition systems.

3.2.2.1 Teacher Network

The teacher network is designed as a high capacity multimodal model that operates during the training stage of the proposed framework. Its primary function is to learn a strong and informative representation of the physiological state by jointly leveraging complementary information from ECG, EDA, and EMG signals, all of which are provided in the standardized window level format produced by the unified signal processing pipeline. Compared with a single modality input setting, this multimodal configuration enables the teacher to capture richer discriminative patterns that may be weak or ambiguous when observed from ECG alone, particularly when physiological responses are subtle or partially confounded by noise and individual variability [111]. Structurally, the teacher is formulated as a sequence of modality specific feature encoders followed by a token based fusion module and a classification head, producing logits in the task prediction space. In the overall system, the teacher serves as a reference model that establishes a strong supervisory signal for subsequent learning stages, while the final deployment objective remains centered on an ECG only student model. The detailed architectural components of the teacher network are described in the following paragraphs, focusing on its input representation, modality encoding, fusion mechanism, and output formulation.

Formally, the input to the teacher network is a multimodal window tensor

$$\mathbf{X} \in \mathbb{R}^{B \times 3 \times L}, \quad (3.1)$$

where B denotes the batch size, L is the fixed window length, and the three channels correspond to ECG, EDA, and EMG signals, respectively. Each modality is processed independently by a modality specific one dimensional convolutional encoder to extract a compact representation. For a given modality m , this process can be expressed as

$$\mathbf{z}^{(m)} = f_m(\mathbf{x}^{(m)}), \quad m \in \{\text{ECG, EDA, EMG}\}, \quad (3.2)$$

where $f_m(\cdot)$ denotes a CNN based feature encoder with shared architecture but independent parameters across modalities. The resulting modality level embeddings are treated as a short sequence of tokens and serve as the input to a Transformer based fusion module, which explicitly models cross modal interactions at the representation level. This design enables the teacher network to integrate complementary physiological information while

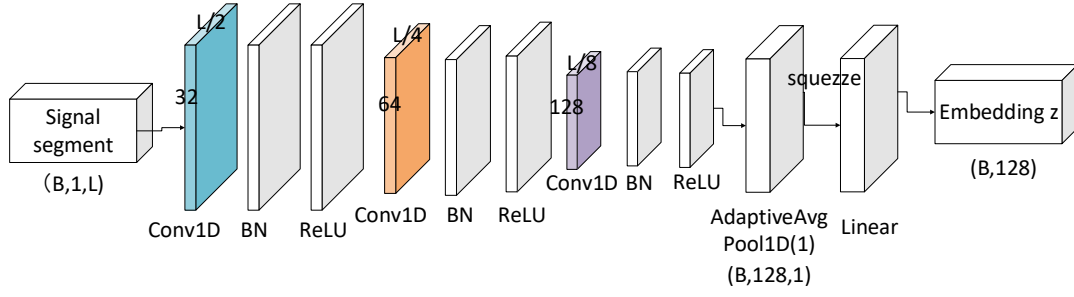


Figure 3.3: Architecture of the modality specific 1D Convolutional Neural Network encoder for physiological signal representation.

maintaining a compact and fixed length intermediate representation suitable for downstream classification and knowledge distillation.

Each modality stream is encoded by an independent 1DCNN encoder with an identical architecture but modality specific parameters. The encoder operates directly on fixed length raw signal windows and follows a hierarchical temporal abstraction strategy, as illustrated in Figure 3.3. Specifically, the input signal $x^{(m)} \in \mathbb{R}^{B \times 1 \times L}$ is processed by a sequence of strided one dimensional convolutional layers, each followed by batch normalization and nonlinear activation, progressively reducing the temporal resolution while increasing the feature dimensionality [112]. This design enables the encoder to capture both short range morphological patterns and longer range temporal characteristics within each signal window.

After convolutional feature extraction, a global average pooling operation is applied along the temporal dimension to obtain a fixed length representation that is invariant to minor temporal shifts within the window [113]. The pooled feature is then projected into the embedding space through a linear layer, resulting in a modality specific embedding vector $z^{(m)} \in \mathbb{R}^{B \times D}$. This process can be summarized as follows.

$$\mathbf{z}^{(m)} = \mathbf{W}_p \text{GAP}(\phi(\text{CNN}_{1D}(\mathbf{x}^{(m)}))), \quad \mathbf{z}^{(m)} \in \mathbb{R}^{B \times D} \quad (3.3)$$

where $\text{CNN}_{1D}(\cdot)$ denotes the stacked convolutional blocks shown in Figure 3.3, $\phi(\cdot)$ represents the nonlinear activation, GAP indicates the global average pooling, and $\mathbf{W}_p$ is the linear projection matrix. As visualized in Figure 3.3, the encoder transforms each input signal segment into a compact embedding of dimension D , providing a standardized and modality aware representation for subsequent token level fusion. By maintaining independent encoders for each modality, the teacher network avoids parameter interference across heterogeneous physiological signals while preserving their distinct temporal dynamics.

In summary, the 1DCNN encoder serves as a modality specific feature extraction module that transforms raw physiological signal windows into compact and fixed dimensional embeddings. By employing a hierarchical convolutional structure with progressive temporal downsampling and global average pooling, the encoder captures discriminative temporal patterns while remaining robust to minor temporal misalignment within each window. The use of identical architectures with independent parameters across modalities ensures that heterogeneous signal characteristics are modeled without cross modality interference. As a result, each encoder produces a standardized embedding that preserves modality specific information while being directly comparable across modalities. These embeddings form the fundamental building blocks for the subsequent token level fusion stage, where cross modal dependencies are explicitly modeled by the Transformer based fusion module.

After modality specific feature extraction, the teacher network converts the set of independent modality embeddings into a tokenized representation suitable for sequence based modeling. Specifically, the embedding vectors obtained from the ECG, EDA, and EMG encoders are treated as discrete tokens, where each token corresponds to one physiological modality. Given the modality embeddings $\mathbf{z}^{ECG}$, $\mathbf{z}^{EDA}$, $\mathbf{z}^{EMG}$ they are stacked along a newly introduced sequence dimension to form a short token sequence

$$\mathbf{T} = \text{stack}(\mathbf{z}^{(ECG)}, \mathbf{z}^{(EDA)}, \mathbf{z}^{(EMG)}) \in \mathbb{R}^{B \times 3 \times D}, \quad (3.4)$$

In this formulation, the token length is fixed to three, corresponding to the three physiological modalities, and each token shares the same embedding dimensionality D . This explicit tokenization step converts multimodal fusion from a feature concatenation problem into a structured sequence modeling problem, enabling modality interactions to be expressed through token level operations rather than implicit parameter mixing. To preserve modality identity within the sequence, a learnable positional embedding is added to the token sequence, ensuring that each token maintains a consistent semantic role across training iterations. The resulting tokenized representation provides a compact and interpretable interface for subsequent fusion, which is performed by a Transformer encoder as described in the next subsection.

Following the tokenization of modality specific embeddings, cross modal interaction is explicitly modeled using a Transformer encoder operating at the token level. The overall fusion process is illustrated in Figure 3.4, where each token represents a distinct physiological modality and is treated as an element in a short input sequence. In contrast to feature concatenation or weighted averaging, this design enables the model to learn adaptive relationships among modalities through self attention mechanisms, allowing each modal-

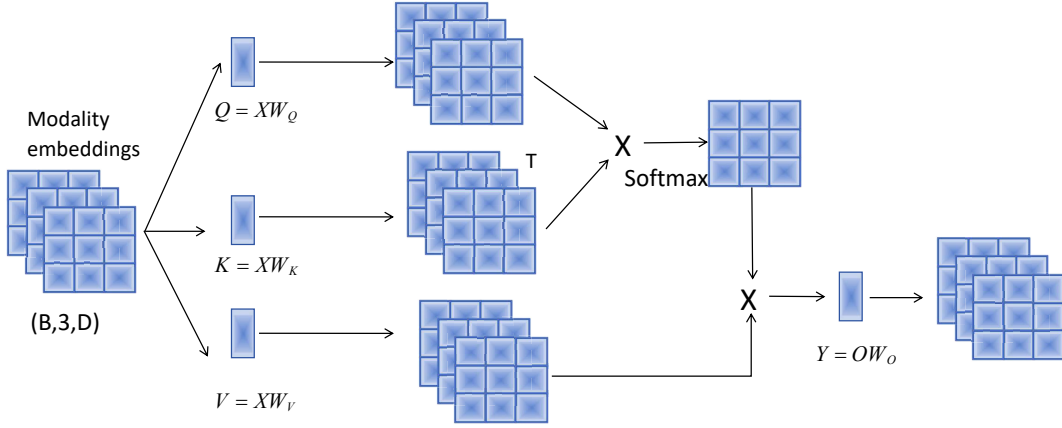


Figure 3.4: Transformer-based fusion architecture for multimodal physiological representations.

ity representation to selectively attend to others based on task relevance [114]. Given the tokenized representation $\tilde{\mathbf{T}} \in \mathbb{R}^{B \times 3 \times D}$, the Transformer encoder applies multi head self attention followed by position wise feedforward layers to produce a fused token sequence,

$$\mathbf{H} = \text{TransformerEnc}(\tilde{\mathbf{T}}) \in \mathbb{R}^{B \times 3 \times D}, \quad (3.5)$$

Through self attention, each output token in $\mathbf{H}$ encodes not only modality specific information but also contextual dependencies learned from other modalities within the same window. This formulation is particularly suitable for multimodal physiological signals, where the relative importance of each modality can vary between subjects, time segments, or experimental conditions. As shown in Figure 3.4, the Transformer fusion module preserves the original token structure while enriching each token with cross modal context, forming the basis for subsequent pooling and classification.

Following token level fusion, the teacher network aggregates the enriched token representations into a single global feature vector for classification. Given the fused token sequence $\mathbf{H} \in \mathbb{R}^{B \times 3 \times D}$ produced by the Transformer encoder, a mean pooling operation is applied along the token dimension to obtain a unified representation,

$$\mathbf{h} = \frac{1}{3} \sum_{i=1}^3 \mathbf{H}_{:,i,:} \in \mathbb{R}^{B \times D}, \quad (3.6)$$

This pooling strategy is chosen due to the fixed and small number of tokens, ensuring a parameter free and stable aggregation that treats all modalities symmetrically. Compared with attention based pooling, mean pooling avoids introducing additional learnable para-

meters and reduces the risk of overfitting, while still preserving the cross modal interactions already encoded by the Transformer.

The aggregated feature vector h is then passed to a multilayer perceptron classification head to produce task level predictions. The classification head consists of a normalization layer followed by two linear transformations with nonlinear activation and dropout [115], and outputs the logits,

$$\mathbf{o} = \text{MLP}(\mathbf{h}) \in \mathbb{R}^{B \times C}, \quad (3.7)$$

where C denotes the number of target classes. These logits serve as the final output of the teacher network and are used both for supervised training with hard labels and for providing soft targets during the knowledge distillation process. As a result, the teacher network delivers a compact yet informative prediction that encapsulates multimodal physiological information in a form directly transferable to the student model.

Table 3.2: Structural summary of the multimodal teacher network.

Component	Input $\rightarrow$ Output	Key layers / settings
Modality encoders $\times 3$	$(B, 1, L) \rightarrow (B, D)$	Each modality (ECG/EDA/EMG) uses an independent 1DCNN encoder: Conv1D($1 \rightarrow 32, k=7, s=2, p=3$) + BN + ReLU $\rightarrow$ Conv1D($32 \rightarrow 64, k=5, s=2, p=2$) + BN + ReLU $\rightarrow$ Conv1D($64 \rightarrow 128, k=5, s=2, p=2$) + BN + ReLU $\rightarrow$ AdaptiveAvgPool1D(1) $\rightarrow$ squeeze $\rightarrow$ Linear($128 \rightarrow D$)
Tokenization	$3 \times (B, D) \rightarrow (B, 3, D)$	Stack modality embeddings as tokens; add learnable positional embedding $\text{pos} \in \mathbb{R}^{1 \times 3 \times D}$
Transformer fusion	$(B, 3, D) \rightarrow (B, 3, D)$	TransformerEncoder with 2 layers, $nhead=4$, FFN dim = $4D$, activation = GELU, dropout = 0.1, batch_first = True, norm_first = True
Pooling	$(B, 3, D) \rightarrow (B, D)$	Mean pooling over token dimension (average across the 3 modality tokens)
Classifier head	$(B, D) \rightarrow (B, C)$	LayerNorm $\rightarrow$ Linear($D \rightarrow D$) $\rightarrow$ GELU $\rightarrow$ Dropout(0.1) $\rightarrow$ Linear($D \rightarrow C$)

In summary, the multimodal teacher network is designed as a structured yet compact architecture that integrates modality specific feature extraction, token level representation, and sequence based fusion within a unified framework. As summarized in Table 3.2, the teacher consists of three independent 1DCNN encoders for ECG, EDA, and EMG signals,

followed by explicit tokenization, Transformer based fusion, parameter free mean pooling, and a lightweight classification head. Each component is carefully selected to balance representational capacity and architectural clarity, ensuring that multimodal interactions are explicitly modeled while avoiding unnecessary complexity.

By operating on standardized window level inputs and maintaining a fixed token length, the teacher network produces stable and interpretable logits that encapsulate cross modal physiological information. These logits not only support conventional supervised learning but also serve as reliable soft targets for guiding the training of a lightweight student model. Importantly, although the teacher uses multimodal input during training, its role is strictly confined to representation learning and supervision, while the final deployment objective remains centered on an ECG only inference model. This design establishes a clear separation between learning capacity and deployment constraints, which forms the basis for the knowledge distillation strategy introduced in the following section.

3.2.2.2 Student Network

The student network is designed as the final inference model for practical physiological stress recognition under constrained sensing and deployment conditions. While the training pipeline leverages multimodal supervision through the teacher network, the student model intentionally relies on electrocardiogram data only at inference time, which aligns with the objective of reducing sensor burden and improving robustness in wearable or real world settings [116]. Conceptually, the student is trained to reproduce the task discriminative behaviour of the multimodal teacher using a lightweight architecture, so that the learned decision boundary can be retained even when only a single modality is available [117]. In this study, the student receives window level inputs that are consistent with the unified preprocessing interface used across datasets, and it outputs task level logits for binary or multiclass classification depending on the experimental configuration. Compared with the teacher network, the student removes the multimodal fusion mechanism and reduces the representation dimensionality, thereby lowering computational cost and model complexity while preserving a compatible output space for supervised learning and knowledge distillation. The remainder of this subsection introduces the student architecture from an implementation oriented perspective, including the input definition and tensor flow, the compact one dimensional convolutional encoder, and the classification head that generates logits used for both standard training objectives and distillation based supervision.

The student network operates on the same window level input format as the teacher

during training, ensuring interface consistency across datasets, while enforcing a strict modality selection constraint at the model entry. Specifically, although each input sample is originally organized as a three channel tensor corresponding to electrocardiogram, electromyogram, and electrodermal activity signals, the student explicitly selects only the electrocardiogram channel for subsequent processing. This operation guarantees that the student network learns to rely solely on single modality information, regardless of whether multimodal signals are available during training. Formally, given an input batch tensor $\mathbf{X} \in \mathbb{R}^{B \times 3 \times L}$, the student input is defined as

$$\mathbf{x} = \mathbf{X}_{[:,0,:]} \in \mathbb{R}^{B \times 1 \times L}, \quad (3.8)$$

Where the student network enforces a deterministic channel selection operation that extracts only the electrocardiogram component from the original multimodal input tensor. This operation reduces the input from a three channel representation to a single channel signal while preserving the original temporal resolution and batch structure. Following channel selection, the tensor shape remains fixed throughout the encoder in terms of batch size, while the temporal dimension is progressively reduced by strided one dimensional convolutions. This deterministic shape flow ensures that all student representations are derived from temporally aligned and uniformly segmented electrocardiogram windows, which is essential for stable optimization and for maintaining compatibility with the teacher generated supervision signals during distillation based training.

The Tiny1DCNNEncoder is designed as a compact yet expressive temporal feature extractor tailored for electrocardiogram based inference under resource constrained conditions. Its architecture follows a progressive downsampling strategy using a small number of strided one dimensional convolutional blocks, enabling the model to capture discriminative temporal patterns while gradually reducing the input resolution. This design reflects the observation that stress related information in electrocardiogram signals is often distributed over local waveform structures as well as longer range temporal dynamics, both of which can be effectively modeled through hierarchical convolutional representations. By stacking convolution, batch normalization, and nonlinear activation layers, the encoder achieves stable optimization and improved generalization without relying on handcrafted filtering or feature extraction stages [118]. Furthermore, the use of global average pooling at the end of the convolutional stack converts variable length temporal features into a fixed dimensional representation, making the encoder robust to minor temporal misalignments and window boundary variations. Compared with the modality specific encoders used in the teacher network, the Tiny1DCNNEncoder reduces channel width and embedding di-

mensionality, which significantly lowers computational cost and parameter count. Despite this simplification, the encoder preserves a compatible representational interface, allowing its output embedding to be directly used for classification as well as for alignment with teacher supervision during knowledge distillation.

From an implementation perspective, the Tiny1DCNNEncoder is composed of three sequential one dimensional convolutional blocks, each consisting of a convolution layer, batch normalization, and a rectified linear unit activation. All convolutional layers employ a stride of two, resulting in progressive temporal downsampling along the encoder depth. Let $\phi(\cdot)$ denote the stacked convolutional transformation followed by global average pooling, then the encoder output can be expressed as

$$\mathbf{h} = \phi(\mathbf{x}) \in \mathbb{R}^{B \times 64}, \quad (3.9)$$

where $x \in \mathbb{R}^{B \times 1 \times L}$ is the selected electrocardiogram input window. Due to the use of three strided convolutional layers, the temporal resolution is approximately reduced by a factor of eight, producing an intermediate feature map of length $L/8$, which is subsequently collapsed to a single temporal unit through adaptive global average pooling. The resulting vector h encodes global temporal statistics of the input window and is projected into the final embedding space through a linear transformation,

$$\mathbf{z} = \mathbf{W}_p \mathbf{h} + \mathbf{b}_p \in \mathbb{R}^{B \times D}, \quad (3.10)$$

where W_p and b_p denote the projection parameters and $D = 64$ in the student configuration. This formulation highlights that the encoder enforces a deterministic shape flow from variable length temporal inputs to fixed dimensional embeddings, which simplifies optimization and ensures compatibility with both supervised classification and distillation objectives.

To further clarify why the student encoder can be regarded as a lightweight or tiny variant of the teacher design, Table 3.3(a) and Table 3.3(b) present a structured comparison between the convolutional encoders used in the teacher and student models, focusing on the electrocardiogram branch. As shown in the table, both encoders share an identical input interface and convolutional depth, consisting of three strided one dimensional convolutional blocks followed by global average pooling. This architectural alignment ensures that the difference between the two models does not arise from network depth or temporal resolution, but rather from channel width and representation capacity. In the student encoder, the number of channels at each convolutional stage is consistently reduced by

approximately one half compared with the teacher, leading to a quadratic reduction in convolutional parameters and multiply accumulate operations under standard convolutional formulations. In addition, the embedding dimension produced by the student encoder is reduced from 128 to 64, further shrinking the parameter count of the projection layer and the subsequent classification head. Importantly, the temporal downsampling strategy and pooling mechanism remain unchanged, which preserves the receptive field structure and the hierarchical temporal abstraction learned by the encoder. This design demonstrates that the student model achieves its compactness through controlled capacity reduction rather than structural simplification, allowing it to maintain a compatible feature extraction pathway while significantly improving deployability and computational efficiency.

Table 3.3(a): Comparison of CNN encoders in Teacher and Student models (ECG branch).

Item	Teacher (Conv weights)	Student (Conv weights)	Interpretation
Input shape	$(B, 1, L)$	$(B, 1, L)$	Same input interface enables a fair architectural comparison.
#Conv blocks	3	3	Same depth; the difference comes from channel width and embedding dimension.
Block 1	Conv1d $1 \rightarrow 32$, $k=7$, $s=2$, $p=3$ + BN + ReLU	Conv1d $1 \rightarrow 16$, $k=7$, $s=2$, $p=3$ + BN + ReLU	First-stage channel width is halved, reducing computation early.
Block 2	Conv1d $32 \rightarrow 64$, $k=5$, $s=2$, $p=2$ + BN + ReLU	Conv1d $16 \rightarrow 32$, $k=5$, $s=2$, $p=2$ + BN + ReLU	Middle-stage channels are consistently reduced.
Block 3	Conv1d $64 \rightarrow 128$, $k=5$, $s=2$, $p=2$ + BN + ReLU	Conv1d $32 \rightarrow 64$, $k=5$, $s=2$, $p=2$ + BN + ReLU	High-level feature channels are halved ($128 \rightarrow 64$).
Downsampling	Stride 2×3 ($\approx \times \frac{1}{8}$ in time length)	Stride 2×3 ($\approx \times \frac{1}{8}$ in time length)	Same temporal downsampling strategy; only capacity is reduced.
Pooling	Adaptive-AvgPool1d(1)	Adaptive-AvgPool1d(1)	Both produce a global vector representation.
Pre-projection feature dim	128	64	Feature dimension is halved before projection.
Projection (Linear)	$128 \rightarrow 128$ (emb_dim=128)	$64 \rightarrow 64$ (emb_dim=64)	Embedding dimension is reduced, further shrinking the head input.
Output embedding	$z \in \mathbb{R}^{B \times 128}$	$z \in \mathbb{R}^{B \times 64}$	Compact embedding improves deployability and reduces parameters.

Table 3.3(b): Comparison of CNN encoders in Teacher and Student models (ECG branch) continued.

Item	Teacher (Conv weights)	Student (Conv weights)	Interpretation
Usage scope	Three independent CNN encoders (ECG/EDA/EMG)	A single CNN encoder (ECG only)	Fewer modality branches leads to substantially smaller overall model.
Conv weight formula	$\sum_{\ell} C_{\text{in}}^{(\ell)} \cdot C_{\text{out}}^{(\ell)} \cdot k^{(\ell)}$	$\sum_{\ell} C_{\text{in}}^{(\ell)} \cdot C_{\text{out}}^{(\ell)} \cdot k^{(\ell)}$	For 1D Conv, weight count scales linearly with kernel size and quadratically with channel width.
Layer-wise counts (3 convs)	$1 \cdot 32 \cdot 7 + 32 \cdot 64 \cdot 5 + 64 \cdot 128 \cdot 5$	$1 \cdot 16 \cdot 7 + 16 \cdot 32 \cdot 5 + 32 \cdot 64 \cdot 5$	Student halves both C_{in} and C_{out} at each stage, yielding $\approx 4\times$ fewer conv weights per block.
Overall scale (rule-of-thumb)	Baseline	$\approx \frac{1}{4}$ (per conv block, rough)	Halving channel widths typically reduces conv parameters and MACs by about $4\times$ at comparable depth and stride.

After feature extraction by the Tiny1DCNNEncoder, the student network maps the resulting embedding to task level predictions through a lightweight classification head. Let $\mathbf{z} \in \mathbb{R}^{B \times D}$, denote the fixed dimensional embedding produced by the encoder, where $D = 64$ in the student configuration. The classification head is implemented as a compact multi-layer perceptron that consists of a normalization layer followed by linear transformations with nonlinear activation and dropout, and produces the output logits

$$\mathbf{o}_s = g(\mathbf{z}) \in \mathbb{R}^{B \times C}, \quad (3.11)$$

where $g(\cdot)$ denotes the head of student classification and C is the number of target classes. This design mirrors the output interface of the teacher network, ensuring that both models generate logits in the same label space.

Importantly, the student logits $\mathbf{o}_s$ are used for two complementary purposes during training. First, they are directly optimized using standard supervised classification objec-

Table 3.4: Compact structural comparison between Teacher and Student models.

Aspect	Teacher (Tri-modal)	Student (ECG-only)
Input modalities	ECG + EDA + EMG (3 modalities)	ECG only (single modality)
Encoder branches	3 independent CNN encoders	1 TinyCNN encoder
Fusion module	Transformer over 3 modality tokens	None (single-branch embedding)
Embedding dim	$D = 128$	$D = 64$
Inference requirement	Needs all 3 modalities	Needs ECG only
Lightweight source	Multi-branch + fusion + larger D	Branch removal + no fusion + smaller D

tives with hard labels. Second, they serve as input from the student side in the knowledge distillation framework, where they are aligned with the teacher logits o_t to transfer the relevant information of the task learned from multimodal supervision. By enforcing this logits level compatibility, the student network can effectively inherit the discriminative behavior of the teacher without replicating its multimodal architecture. As a result, the classification head provides a simple yet crucial bridge between the compact student encoder and the rich supervisory signals provided by the teacher model.

From a structural perspective, the student network differs from the teacher network primarily in model capacity and modality scope rather than in the fundamental feature extraction pathway. As summarized in Table 3.4, the teacher model operates on three physiological modalities using independent convolutional branches followed by a Transformer based fusion module, whereas the student model is restricted to a single electrocardiogram input and adopts a compact single branch encoder without explicit fusion. Importantly, both models preserve a consistent window level input interface, convolutional depth, and temporal downsampling strategy, ensuring that the student maintains the same hierarchical temporal abstraction and receptive field structure as the teacher. The reduction in complexity therefore arises from controlled capacity scaling, including narrower channel widths, a smaller embedding dimension, and the removal of cross modal fusion components, rather than from architectural simplification. This design allows the student to retain a compatible representation and output space while substantially reducing inference re-

quirements, making it suitable for real world deployment. In contrast, the teacher network serves as a high capacity reference model during training, a role separation that directly motivates the knowledge distillation framework introduced in the following section.

In summary, the student network is designed as a compact electrocardiogram based inference model that preserves the core temporal representation pathway of the multimodal teacher while substantially reducing model capacity and sensing requirements. Through controlled architectural scaling rather than structural simplification, the student maintains a compatible input representation, embedding space, and logits output interface, enabling it to serve as a practical deployment model. However, due to its reduced modality access and representational capacity, direct supervised training alone may not fully exploit the rich cross modal information available during training. This limitation motivates the introduction of a training strategy in which the high capacity teacher model provides additional supervisory signals to guide the optimization of the student. Consequently, the following section details the model training and evaluation protocol, with a particular focus on the knowledge distillation framework that bridges the representational gap between the multimodal teacher and the lightweight student.

3.3 Model Training and Evaluation Protocol

Building upon the architectural design of the multimodal teacher network and the lightweight electrocardiogram based student model presented in the preceding section, this section focuses on how these models are trained, deployed, and evaluated in a unified and reproducible manner. Although teacher and student networks differ substantially in capacity and modality access, their interaction during training is governed by a carefully designed protocol that enables effective knowledge transfer without compromising the inference constraints of the student. Consequently, this section first introduces the knowledge distillation framework that defines the supervisory relationship between the teacher-student models. It then describes the training procedure used to optimize the student model under combined hard label and soft target supervision, followed by the inference and testing protocol that reflects real world deployment conditions. Finally, the evaluation metrics and the assessment strategy are presented to ensure a fair and consistent performance comparison between the datasets and the experimental settings.

3.3.1 Knowledge Distillation Framework

Building on the teacher-student architectures introduced in Section 3.2, this section presents the knowledge distillation (KD) framework adopted to bridge the gap between multimodal training and ECG-only inference [119]. Although the teacher network is trained using synchronized ECG, EDA, and EMG signals to exploit complementary physiological information, the student model is strictly limited to operating on ECG signals alone during inference. This design reflects practical deployment requirements for wearable stress monitoring, where reducing sensor complexity and power consumption is essential. Knowledge distillation is therefore used as a training-time strategy to transfer discriminative knowledge learned from multimodal supervision into a compact student model with only ECG, without violating the single-modality constraint inference [120]. Rather than modifying the student architecture or introducing additional input, the proposed KD framework focuses on guiding the student’s decision boundaries through soft targets generated by the teacher, allowing the student to approximate multimodal decision behavior using only ECG representations. The following subsections detail the role assignment and data flow between teacher and student models, the formulation of distillation losses, and the training mechanisms designed to ensure stable and effective knowledge transfer.

As summarized in Table 3.5, the proposed knowledge distillation framework consists of four key components: a multimodal teacher network, an ECG-only student network, ground-truth labels, and a dedicated distillation loss module. During training, the teacher receives synchronized ECG, EDA, and EMG signals and produces class-level logits $\mathbf{z}_t$, which serve as soft targets for guiding the student. The teacher model remains frozen throughout the distillation process, ensuring that only the student parameters are updated. In contrast, the student network operates on ECG signals exclusively, receiving $\mathbf{X}_{ecg}$ as input and producing logits $\mathbf{z}_s$ with the same dimensionality as the teacher output. This output alignment enables direct distillation at the logit level without introducing additional projection layers. Supervision is provided through a combination of hard labels and teacher-generated soft targets, which are integrated by the KD loss module [121]. Importantly, multimodal signals are used only during training, while inference strictly relies on ECG inputs, preserving the deployment-oriented design of the student model.

Table 3.5: Roles and data flow in the proposed knowledge distillation framework.

Component	Input	Output	Description
Teacher (multimodal)	$\mathbf{X} \in \mathbb{R}^{B \times 3 \times L}$	$\mathbf{z}_t \in \mathbb{R}^{B \times C}$	Trained with ECG, EDA, and EMG during training; provides soft targets for distillation and is frozen during student training.
Student (ECG-only)	$\mathbf{X}_{ecg} = \mathbf{X}[:, 0:L, :] \in \mathbb{R}^{B \times 1 \times L}$	$\mathbf{z}_s \in \mathbb{R}^{B \times C}$	Lightweight inference model using ECG only; optimized using both hard labels and teacher-provided soft targets.
Ground-truth labels	$\mathbf{y} \in \{0, \dots, C-1\}^B$	–	Supervisory signal for standard cross-entropy loss applied to student outputs.
KD loss module	$\mathbf{z}_t, \mathbf{z}_s, \mathbf{y}$	$\mathcal{L}$	Combines classification loss and distillation loss with warm-up, ramping, and confidence masking mechanisms.

To enable effective knowledge transfer from the multimodal teacher to the ECG-only student, soft targets are generated by applying temperature scaling to the output logits of both models. Unlike hard labels, which provide only the index of the ground-truth class, soft targets encode inter-class similarity information captured by the teacher and therefore offer a richer supervisory signal during training [122]. Specifically, given the teacher logits $\mathbf{z}_t \in \mathbb{R}^{B \times C}$ and student logits $\mathbf{z}_s \in \mathbb{R}^{B \times C}$, temperature-scaled probability distributions are computed using the softmax function as follows:

$$p_{t,c}^{(T)} = \frac{\exp(z_{t,c}/T)}{\sum_{k=1}^C \exp(z_{t,k}/T)}, \quad p_{s,c}^{(T)} = \frac{\exp(z_{s,c}/T)}{\sum_{k=1}^C \exp(z_{s,k}/T)}, \quad (3.12)$$

where $T > 1$ denotes the temperature hyperparameter. A higher temperature produces a softer probability distribution by reducing the dominance of the maximum logit, thereby amplifying the relative probabilities of non-target classes. This softened distribution re-

flects the teacher’s uncertainty and class-level structure, which is particularly valuable when the student model operates under a reduced input modality constraint.

In this work, temperature scaling is applied consistently to both teacher and student outputs during distillation. The resulting soft probabilities are later used to compute the distillation loss, encouraging the student to match the teacher’s class distribution rather than only the hard decision boundary. By introducing temperature scaling, the student is guided to learn more informative decision surfaces that better approximate the multimodal teacher’s behavior, despite relying solely on ECG signals during inference.

Knowledge distillation in this work is formulated as a joint optimization objective that combines supervised learning from hard labels with distribution matching to the multimodal teacher [123]. Let $\mathbf{z}_s \in \mathbb{R}^{B \times C}$ denote the student logits and $\mathbf{z}_t \in \mathbb{R}^{B \times C}$ denote the teacher logits for a mini-batch of size B with C classes, and let $\mathbf{y} \in \{0, 1\}^{B \times C}$ be the one-hot encoded ground-truth labels. The supervised component is defined by the standard cross entropy loss:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{B} \sum_{b=1}^B \sum_{c=1}^C y_{b,c} \log p_{s,b,c}, \quad \mathbf{p}_s = \text{softmax}(\mathbf{z}_s), \quad (3.13)$$

where $p_{s,b,c}$ is the predicted probability of class c for sample b . The distillation component encourages the student to match the teacher softened class distribution under temperature scaling, which is implemented via the Kullback Leibler divergence:

$$\mathcal{L}_{\text{KD}} = T^2 \cdot \text{KL}(\mathbf{p}_t^{(T)} \parallel \mathbf{p}_s^{(T)}) = T^2 \cdot \frac{1}{B} \sum_{b=1}^B \sum_{c=1}^C p_{t,b,c}^{(T)} \log \left(\frac{p_{t,b,c}^{(T)}}{p_{s,b,c}^{(T)} + \varepsilon} \right), \quad (3.14)$$

where $\mathbf{p}_t^{(T)} = \text{softmax}(\mathbf{z}_t / T)$ and $\mathbf{p}_s^{(T)} = \text{softmax}(\mathbf{z}_s / T)$ are temperature softened probabilities, T is the temperature hyperparameter, and ε is a small constant used for numerical stability. The overall distillation objective is a weighted combination of the two terms:

$$\mathcal{L} = (1 - \alpha_{\text{eff}}) \mathcal{L}_{\text{CE}} + \alpha_{\text{eff}} \mathcal{L}_{\text{KD}}, \quad (3.15)$$

where $\alpha_{\text{eff}} \in [0, 1]$ controls the relative contribution of distillation in a given training stage. This formulation ensures that the student remains anchored to the ground truth labels while being guided to approximate the richer class structure learned by the multimodal teacher.

Table 3.6: Dynamic mechanisms for stabilising knowledge distillation during student training.

Mechanism	Definition	Purpose and effect
Warm-up training	For the first E_w epochs, distillation is disabled by setting $\alpha_{\text{eff}} = 0$, and the student is optimised using $\mathcal{L}_{\text{CE}}$ only.	Prevents early-stage instability and avoids forcing the student to mimic an immature or noisy teacher distribution.
Alpha ramp-up	After warm-up, the distillation weight α_{eff} is gradually increased over E_r epochs until it reaches the target α .	Introduces distillation smoothly, stabilising optimisation dynamics and improving convergence.
Teacher confidence masking	For each sample, distillation is applied only when the teacher confidence exceeds a threshold τ , otherwise only $\mathcal{L}_{\text{CE}}$ is used.	Reduces negative transfer from low-confidence teacher predictions and focuses distillation on reliable soft targets.

To improve the stability and reliability of student optimization, the proposed knowledge distillation procedure incorporates a set of dynamic mechanisms that control when and how strongly distillation is applied. These mechanisms are summarized in Table 3.6, including warm-up training, gradual ramp-up of the distillation weight, and teacher confidence masking. The underlying motivation is that, during early training or under uncertain teacher predictions, enforcing strict distribution matching may introduce noisy gradients and degrade student generalization. By dynamically adjusting the contribution of the distillation term, the student can first learn a robust decision boundary from hard labels and then progressively incorporate the richer supervisory signal provided by the teacher.

$$\alpha_{\text{eff}}(e, b) = \alpha \cdot \min\left(1, \max\left(0, \frac{e - E_w}{E_r}\right)\right) \cdot \mathbb{I}\left[\max_c p_{t,b,c}^{(T)} \geq \tau\right], \quad (3.16)$$

Formally, the combined effect of these mechanisms is implemented through an effective distillation weight α_{eff} , which modulates the contribution of the distillation loss term at epoch e and for sample b , as defined in Eq. 3.16. In this formulation, $\alpha \in [0, 1]$ denotes the target distillation weight, E_w is the number of warm-up epochs during which distillation is disabled, and E_r is the number of ramp-up epochs used to increase the distillation strength linearly from zero to α . The indicator function $\mathbb{I}[\cdot]$ implements teacher confidence masking,

where $p_{t,b,c}^{(T)}$ denotes the teacher soft probability for class c at temperature T for sample b . Distillation is applied only when the maximum teacher probability across classes exceeds a threshold τ , ensuring that unreliable teacher outputs do not dominate student updates. As a result, α_{eff} becomes sample-dependent and stage-dependent, providing fine-grained control over the distillation signal.

Practically, these dynamic mechanisms ensure that distillation complements rather than overrides standard supervised learning. Warm-up training provides an initial period in which the student learns from ground-truth labels only, establishing a stable representation and preventing premature overfitting to the teacher distribution. The ramp-up schedule then introduces the distillation loss progressively, allowing the student to incorporate multi-modal knowledge in a controlled manner as the optimization becomes more stable. Finally, confidence masking limits distillation to cases where the teacher offers reliable soft targets, improving robustness against uncertain predictions that may arise from cross-subject variability or transient noise in physiological signals. Together, the mechanisms in Table 3.6 and Eq. 3.16 form a practical and reproducible strategy to stabilize knowledge transfer from the multimodal teacher to the lightweight ECG-only student.

Model selection and inference are conducted under a protocol that reflects both fair evaluation and practical deployment constraints. During training, the student model is evaluated on a held-out validation set at the end of each epoch, and the model checkpoint that achieves the best validation F1-score is selected as the final student model [124]. This selection strategy ensures that the chosen model generalizes well beyond the training data and avoids bias from overfitting or late-stage optimization noise. After model selection, the teacher network is no longer involved, and all inference and testing procedures are performed using the selected student model-only. At inference time, the student operates strictly under the ECG-only constraint, receiving window-level ECG inputs and producing task-level predictions without access to additional physiological modalities or distillation signals. This protocol guarantees a clear separation between training-time supervision and deployment-time inference, ensuring that the reported results accurately reflect the performance of the lightweight ECG-based model in real-world usage scenarios.

The hyperparameters in the proposed framework are selected in a dataset-aware manner to account for differences in signal characteristics, class distributions, and task definitions across experimental settings. As a result, no single fixed configuration is enforced in all experiments. Instead, a consistent selection strategy is adopted, in which optimization-related parameters such as learning rate, batch size, and number of training epochs are tuned separately for each dataset using the corresponding training and validation splits.

For the knowledge distillation components, the temperature parameter, distillation weight, warm-up duration, and ramp-up schedule are chosen to ensure stable convergence and prevent dominance of the distillation signal in early training stages. All models are implemented using a unified training pipeline, ensuring that variations across experiments arise from dataset properties rather than differences in implementation. This approach balances reproducibility with flexibility, allowing the framework to adapt to heterogeneous datasets while maintaining a consistent methodological structure.

In summary, this section has presented a complete knowledge distillation framework designed to transfer multimodal supervisory knowledge into a lightweight ECG-only student model under realistic deployment constraints. By defining clear teacher–student roles, introducing temperature-scaled soft targets, formulating a joint distillation objective, and incorporating dynamic control mechanisms, the proposed framework enables stable and effective knowledge transfer during training without altering the inference-time structure of the student. Importantly, all distillation components are confined to the training phase, ensuring that the final model operates strictly under the ECG-only constraint at deployment. Building upon this framework, the next section details the concrete training procedure, describing how the distillation strategy is implemented in practice, including data batching, optimization steps, and evaluation checkpoints.

3.3.2 Training Procedure

This section describes the training procedure and evaluation protocol adopted in this study, with the objective of ensuring a fair and reproducible comparison between the multimodal teacher model, the ECG only student baseline, and the enhanced student model [125]. Building upon the model architectures introduced in the previous section, this part focuses on how different models are optimized, how supervision signals are constructed, and how inference performance is assessed under consistent experimental conditions.

Specifically, three training configurations are considered. The first configuration corresponds to the supervised training of the multimodal teacher network using electrocardiogram, electromyogram, and electrodermal activity signals. The second configuration involves direct training of a lightweight student model that operates solely on electrocardiogram signals. The third configuration introduces a knowledge distillation framework, in which the frozen teacher model provides soft supervisory signals to guide the training of the ECG only student model [126]. Through this design, the proposed training procedure explicitly decouples multimodal representation learning from deployment time inference,

allowing the student model to benefit from multimodal information during training while strictly adhering to an ECG only constraint at test time [127].

Table 3.7: Training settings and learning targets for different model configurations.

Training setting	Input modalities	Supervision signal	Optimization target
Teacher baseline training	ECG, EMG, EDA	Ground truth labels	Minimize cross entropy loss for multimodal classification
Student baseline training	ECG only	Ground truth labels	Minimize cross entropy loss under ECG only constraint
Knowledge distillation training	Teacher: ECG, EMG, EDA Student: ECG only	Ground truth labels and teacher soft targets	Joint optimization of hard label loss and distillation loss

To clearly distinguish the optimization objectives under different training configurations, the learning settings adopted in this study are summarized in Table 3.7. Although all models address the same classification task, their input modalities and supervision signals differ substantially, resulting in distinct learning targets and optimization behaviors.

In the teacher baseline setting, the multimodal fusion model is trained using electrocardiogram, electromyogram, and electrodermal activity signals, with ground truth labels serving as the sole supervision source. The objective of this stage is to learn a high capacity multimodal classifier that captures complementary physiological information across modalities. In contrast, the student baseline model is trained directly on electrocardiogram signals only, under identical label supervision. This configuration reflects a deployment oriented scenario in which inference is constrained to a single modality, but it does not benefit from additional multimodal guidance during training.

The knowledge distillation setting introduces a hybrid supervision strategy. While the student model continues to receive hard labels derived from the dataset annotations, it is additionally guided by soft targets produced by the pretrained multimodal teacher model. Importantly, the teacher model is frozen during this stage and is used exclusively to generate supervisory signals, whereas the student model is the only component being updated [128]. As shown in Table 3.7, this design enables the student model to leverage multimodal knowledge during training while preserving a strictly ECG only inference pathway.

The student model is optimized using a composite objective that combines hard label

supervision with soft target guidance from the teacher model. Specifically, the standard cross-entropy loss is first applied to enforce consistency between the student predictions and the ground truth labels [129], which is defined as

$$\mathcal{L}_{\text{CE}} = -\frac{1}{B} \sum_{i=1}^B \sum_{c=1}^C y_{i,c} \log(\sigma(z_s)_{i,c}) \quad (3.17)$$

where $\sigma(\cdot)$ denotes the softmax function, $z_s \in \mathbb{R}^{B \times C}$ denotes the logits produced by the student model for a batch of B samples, C is the number of classes, and y represents the corresponding ground truth labels. This term ensures that the student model preserves discriminative alignment with the labeled training data and serves as the primary optimization signal when no distillation is applied.

To enable knowledge transfer from the teacher, both teacher and student logits are transformed into softened probability distributions using temperature-scaled softmax,

$$p_t^{(T)} = \text{softmax}\left(\frac{z_t}{T}\right), \quad p_s^{(T)} = \text{softmax}\left(\frac{z_s}{T}\right) \quad (3.18)$$

where z_t denotes the logits generated by the frozen teacher model, and $T > 1$ is the temperature parameter controlling the smoothness of the output distributions. A higher temperature reduces the confidence of the predicted probabilities and amplifies relative differences among non-maximum classes, thereby exposing richer inter-class similarity information learned by the teacher. The student distribution $p_s^{(T)}$ is constructed in the same manner to ensure compatibility during distillation.

The discrepancy between the teacher and student soft distributions is quantified using the Kullback–Leibler divergence,

$$\mathcal{L}_{\text{KD}} = T^2 \cdot \frac{1}{B} \sum_{i=1}^B \sum_{c=1}^C p_{t,i,c}^{(T)} \log \frac{p_{t,i,c}^{(T)}}{p_{s,i,c}^{(T)}} \quad (3.19)$$

where the scaling factor T^2 compensates for the gradient magnitude attenuation introduced by the temperature scaling, following standard distillation practice. This formulation treats the teacher distribution as the target and explicitly encourages the student to approximate the softened class structure learned by the multimodal teacher. This distillation loss encourages the student model to approximate the teacher’s class-wise probability structure, effectively transferring multimodal knowledge into the ECG-only representation space.

Finally, the overall training objective for knowledge distillation is formulated as a weighted combination of the two loss terms,

$$\mathcal{L} = (1 - \alpha_{\text{eff}}) \mathcal{L}_{\text{CE}} + \alpha_{\text{eff}} \mathcal{L}_{\text{KD}} \quad (3.20)$$

where $\alpha_{\text{eff}} \in [0, 1]$ denotes the effective distillation weight. This parameter controls the relative contribution of hard label supervision and soft target guidance, allowing the training process to balance label fidelity and teacher knowledge transfer dynamically. Through this formulation, the student model is guided to achieve robust classification performance under ECG-only inference constraints while benefiting from the richer multimodal representations learned by the teacher during training.

To ensure stable optimization and reliable knowledge transfer during distillation, a set of dynamic mechanisms are introduced to regulate when and how the distillation signal is applied throughout training. These mechanisms are designed to mitigate early-stage student instability and to prevent the student from being forced to mimic teacher distributions before learning a reasonable ECG-only representation. As summarized in Table 3.8, the proposed framework incorporates three complementary strategies, including warm-up training, gradual increase in distillation weight, and masking of teacher confidence.

Table 3.8: Dynamic mechanisms used to stabilize knowledge distillation during student training.

Mechanism	Definition	Purpose
Warm-up training	Distillation is disabled for the first E_w epochs by setting $\alpha_{\text{eff}} = 0$.	Allows the student to learn a stable decision boundary from hard labels before distillation is introduced.
Distillation ramp-up	The effective distillation weight α_{eff} is increased linearly over E_r epochs.	Prevents abrupt optimization changes and improves training stability.
Teacher confidence masking	Distillation is applied only when $\max_c p_{t,c}^{(T)} \geq \tau$.	Suppresses unreliable teacher supervision and reduces negative transfer.

During the initial phase of training, a warm-up strategy is employed in which the student model is optimized solely using the hard-label cross-entropy loss, while the distillation loss is disabled. This phase allows the student to establish a stable decision boundary based on ground-truth supervision before being influenced by the teacher distribution. After the warm-up period, the contribution of the distillation loss is gradually introduced

through a linear ramp-up schedule. Formally, the effective distillation weight in the training epoch e is defined as

$$\alpha_{\text{eff}}(e) = \alpha \cdot \min\left(1, \max\left(0, \frac{e - E_w}{E_r}\right)\right), \quad (3.21)$$

where $\alpha \in [0, 1]$ denotes the target distillation weight, E_w is the number of warm-up epochs during which distillation is disabled, and E_r controls the duration of the ramp-up stage. This formulation ensures that the influence of the teacher model increases smoothly over time, preventing abrupt changes in the optimization landscape.

In addition to epoch-level scheduling, a confidence-based masking mechanism is applied at the sample level to further improve robustness. Note that the confidence threshold τ is defined on the temperature-scaled teacher distribution $p_t^{(T)}$ and is selected on the validation set to balance reliability and coverage of distillation signals. For each training sample, distillation is enabled only when the teacher prediction is sufficiently confident, as measured by the maximum probability of the temperature-scaled teacher distribution. The confidence mask is defined as

$$m_b = \mathbb{I}\left[\max_c p_{t,b,c}^{(T)} \geq \tau\right] \quad (3.22)$$

where $p_{t,b,c}^{(T)}$ denotes the temperature-scaled probability assigned by the teacher to class c for the sample b , τ is a predefined confidence threshold, and $\mathbb{I}[\cdot]$ is the indicator function. When the confidence condition is not satisfied, the distillation loss is suppressed and the student is updated using hard-label supervision only. This mechanism reduces the risk of negative transfer caused by uncertain or noisy teacher outputs, which may arise from subject variability or transient artifacts in physiological signals.

By jointly applying warm-up training, ramp-up scheduling, and confidence masking, the proposed dynamic distillation mechanisms provide fine-grained control over the strength and reliability of the distillation signal. As a result, the student model is guided to absorb multimodal knowledge progressively and selectively, leading to more stable training dynamics and improved generalization under ECG-only inference constraints.

The training process follows a structured epoch-based schedule to ensure stable optimization and fair comparison between different model configurations. For each training epoch, mini-batches are sequentially sampled from the training set and processed by the corresponding model according to the selected training setting. In the knowledge distillation configuration, each batch is first forwarded through the frozen teacher network to generate soft supervisory signals, followed by a forward pass through the student network

using ECG inputs only. The total loss is then computed based on the combined objective described in the previous section, after which the gradients are backpropagated exclusively through the student model. The model parameters are updated using a single optimizer with consistent optimization settings across experiments, while the teacher parameters remain fixed throughout training. For each dataset, all training configurations share the same validation protocol and hyperparameter tuning budget, ensuring that performance differences arise from learning strategies rather than optimisation advantages. This unified training schedule ensures that performance differences between models arise from learning strategies rather than discrepancies in optimization procedures.

Model selection is performed using a validation-driven strategy that reflects generalization performance rather than training accuracy. At the end of each training epoch, the student model is evaluated on a held-out validation set, and the checkpoint achieving the highest validation F1-score is retained as the final model [130]. This criterion is adopted to account for potential class imbalance and to provide a robust measure of classification quality. Once training is complete, the selected model is used for all subsequent testing and inference. During inference, only the student model is deployed, and predictions are generated strictly from ECG inputs without access to additional physiological modalities or distillation-related components. This protocol enforces a clear separation between training-time supervision and deployment-time inference, ensuring that reported results accurately reflect the performance of the ECG-only model under realistic application constraints.

All models are implemented within a unified training framework to ensure consistency and reproducibility across experiments. Identical data loading procedures, preprocessing pipelines, and evaluation scripts are used for all training configurations, minimizing variability introduced by implementation differences. To promote reproducibility, random seeds are fixed for model initialization, data shuffling, and optimization operations [131]. Training is carried out in the same computational environment using consistent software libraries and hardware resources. Although hyperparameters can vary between datasets to accommodate differences in signal characteristics and task definitions, the general training and evaluation protocol remains unchanged. This design ensures that experimental results can be reliably reproduced and that observed performance differences are attributable to methodological choices rather than uncontrolled implementation factors.

In summary, this section has described the complete training procedure adopted in this study, including the optimization schedule, the model selection strategy, and implementation considerations that ensure fair and reproducible experimentation. By maintaining a

unified training pipeline across all model configurations and strictly separating training-time supervision from deployment-time inference, the proposed procedure enables a reliable assessment of different learning strategies under consistent conditions. Having established how models are trained and selected, the next section focuses on the inference and testing protocol, detailing how trained models are evaluated under ECG-only constraints and how performance is measured in realistic deployment scenarios.

3.3.3 Inference and Testing Protocol

Building upon the training strategies and distillation mechanisms described in the previous section, this section presents a unified inference and testing protocol for the proposed teacher, student, and knowledge-distilled student models. While the training phase focuses on representation learning and supervision transfer, the inference stage imposes stricter constraints on input modalities, computational behavior, and evaluation consistency. To this end, the protocol explicitly defines the roles of different models during testing, the input–output conventions at window level, and the standardized procedures for model selection and performance evaluation.

The primary purpose of this section is to ensure that all reported results are obtained under a fair, reproducible, and deployment-consistent setting. In particular, although the teacher model leverages multimodal inputs during training and evaluation as an upper-bound reference, both the baseline student and the distilled student are strictly evaluated using ECG-only inputs, reflecting the intended real-world deployment scenario [132]. By clearly separating training-time design choices from inference-time constraints, this section establishes a common testing framework that enables reliable cross-model, cross-task, and cross-dataset comparisons in the experimental analysis.

During inference and testing, the teacher, baseline student, and knowledge-distilled student models serve distinct but complementary roles, which are explicitly defined to ensure evaluation consistency and deployment relevance. The teacher model acts as an upper-bound reference and diagnostic baseline, using all available physiological modalities provided by the dataset to generate prediction logits [122]. This multimodal inference is retained solely for benchmarking purposes and is not intended for deployment.

In contrast, both the baseline student and the knowledge-distilled student are evaluated under a strict ECG-only inference constraint. Although the input data are organized in a unified window-level tensor that may contain multiple modalities, only the ECG channel is selected and forwarded to the student encoder during inference. Importantly, the inference architecture and input pathway of the distilled student are identical to those of

the baseline student, ensuring that any observed performance differences arise exclusively from the training strategy rather than from additional test-time information or architectural modifications.

This role separation establishes a clear and fair evaluation protocol: the teacher provides a multimodal performance upper bound, while the student and distilled student represent deployable ECG-only models. As a result, comparisons between models directly reflect the effectiveness of knowledge distillation under realistic inference constraints, without confounding effects from modality availability or test-time fusion.

To ensure clarity and consistency across different datasets, tasks, and model variants, this subsection formally defines the input–output representations and tensor shape conventions adopted throughout the inference and testing stages. A unified notation is used for both multimodal teacher inference and ECG-only student inference, which facilitates fair comparison and reproducibility. At the window level, the input to the model is represented as a three-dimensional tensor

$$\mathbf{X} \in \mathbb{R}^{B \times M \times L}, \quad (3.23)$$

denote a batch of windowed input signals, where B is the batch size, M is the number of available physiological modalities, and L is the temporal length of each fixed-duration signal window. The modality dimension follows a fixed order determined during data preprocessing and remains consistent throughout training and evaluation. For inference based on only ECG, the student and distilled student models explicitly select the ECG channel from the multimodal input tensor, resulting in the output of the ECG channel.

$$X_{ecg} = X_{[:,0:1,:]} \in \mathbb{R}^{B \times 1 \times L}, \quad (3.24)$$

This operation enforces a strict single-modality inference constraint and prevents any form of test-time information leakage from non-ECG channels. For all models, the forward pass produces a vector of classification logits

$$\mathbf{z} \in \mathbb{R}^{B \times C}, \quad (3.25)$$

where C denotes the number of target classes, which can vary between tasks. The final class predictions are obtained by applying an argmax operation on the class dimension. This unified input–output specification enables consistent evaluation and fair comparison between multimodal teacher models and ECG-only student models under identical window-level testing conditions.

During inference and testing, the teacher and student models follow distinct architectural pathways that reflect their intended roles while maintaining a unified output interface. The teacher model adopts a multimodal inference architecture, in which window-level inputs containing all available physiological modalities are processed by independent modality-specific encoders. Each encoder maps its corresponding signal channel to a latent embedding, and the resulting modality embeddings are subsequently fused through a transformer-based fusion module. The fused representation is then aggregated and passed to a classification head to produce the final prediction logits. This inference pathway leverages the full multimodal context available in the dataset and serves as an upper-bound reference for representation quality.

In contrast, both the baseline student model and the knowledge-distilled student model employ an ECG-only inference architecture that is intentionally constrained for deployment consistency. During inference, the student model receives the same window-level input tensor format but explicitly operates only on the ECG channel. The selected ECG signal is encoded by a lightweight convolutional encoder, followed by a classification head that outputs logits in the same class space as the teacher. Importantly, the inference architecture of the student model remains identical regardless of whether it is trained with or without knowledge distillation. The distillation process influences only the training dynamics, while the inference-time structure and computational pathway of the student model are kept fixed. This design ensures a fair comparison between baseline and distilled student models and aligns the inference behavior with practical ECG-only deployment scenarios.

The inference and testing procedure follows a unified pipeline designed to ensure consistency, fairness, and reproducibility across all evaluated models. As summarized in Table 3.9, the pipeline begins with loading the best-performing model checkpoint selected during training and switching the model to evaluation mode to disable gradient computation. Test samples are then processed in a window-level manner using a standardized data loader, ensuring that all models are evaluated under identical preprocessing and batching conditions [133].

A key aspect of the unified pipeline is the explicit handling of input modalities during inference. While the teacher model receives all available physiological modalities as input, both the baseline student and the distilled student models strictly operate on ECG signals only, even when additional modalities are present in the stored input tensor. This modality selection step enforces deployment-consistent inference behavior and prevents any form of test-time information leakage. Following the forward pass through the corresponding inference architecture, class-level logits are generated for each input window, and predicted

labels are obtained using a deterministic decision rule.

Table 3.9: Unified inference and testing pipeline for teacher and student models.

Step	Operation	Description
1	Load trained model checkpoint	Load the best-performing model checkpoint selected during training based on validation performance. The teacher, baseline student, and distilled student are evaluated independently under the same protocol.
2	Set evaluation mode	Switch the model to evaluation mode and disable gradient computation to ensure deterministic behavior and reduce memory usage during inference.
3	Prepare test data loader	Iterate over the test dataset using a window-level data loader with fixed batch size and consistent preprocessing across all models.
4	Select input modalities	For the teacher model, all available physiological modalities are used as input. For the student and distilled student models, only the ECG channel is selected from the input tensor.
5	Forward pass	Perform a forward pass through the corresponding inference architecture to generate class-level logits for each input window.
6	Generate predictions	Obtain predicted class labels by applying an argmax operation over the logit outputs for each window.
7	Accumulate outputs	Collect predicted labels and ground-truth labels across all batches for subsequent performance evaluation.
8	Compute evaluation metrics	Compute evaluation metrics such as accuracy and F1-score on the aggregated predictions to quantify model performance.

Finally, predictions and ground-truth labels are accumulated throughout the test set to

enable a reliable computation of evaluation metrics. By standardizing each stage of the inference and testing process, the pipeline in Table 3.9 provides a common evaluation foundation for multimodal teacher models and ECG-only student models alike. This unified design ensures that performance comparisons reflect differences in learned representations and training strategies, rather than inconsistencies in test-time procedures.

During the testing stage, model predictions and performance evaluation are defined in a unified and task-independent manner to ensure fair comparison between the multimodal teacher, the ECG-only baseline student, and the distilled student model. For each input window, the model outputs a vector of class-level logits $z_i \in \mathbb{R}^C$, where C denotes the number of target classes. The predicted label is determined by selecting the class associated with the maximum logit value [134],

$$\hat{y}_i = \arg \max_c z_{i,c}. \quad (3.26)$$

This prediction rule is consistently applied to all models and datasets, regardless of the number of input modalities or task-specific settings, providing a unified decision criterion at inference time.

Model performance is evaluated primarily using classification accuracy and the F1 score. Accuracy is defined as the proportion of correctly classified samples over the test set,

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\hat{y}_i = y_i), \quad (3.27)$$

where N is the number of test samples and $\mathbb{I}(\cdot)$ denotes the indicator function. To account for potential class imbalance commonly observed in physiological stress datasets, the F1 score is adopted as the primary evaluation metric, as it balances precision and recall and provides a more informative assessment than accuracy alone. For binary classification tasks, the F1 score of the positive class is reported, while for multi-class tasks, the macro-averaged F1 score is used,

$$\text{F1}_{\text{macro}} = \frac{1}{C} \sum_{c=1}^C \text{F1}_c. \quad (3.28)$$

where $\text{F1}_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}$ denotes the F1 score of class c . In addition, the cross-entropy loss may optionally be computed on the test set as a reference measure of prediction confidence,

$$\mathcal{L}_{\text{test}} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C y_{i,c} \log(\sigma(z_i)_c) \quad (3.29)$$

although it is not used for model selection. Together, these metrics provide a comprehensive and consistent evaluation framework for both binary and multi-class stress recognition tasks.

During training, model selection is performed using a validation based strategy to ensure an unbiased assessment of generalization performance. For each training configuration, the model checkpoint that achieves the best performance in the validation set is retained as the final model for testing. The primary selection criterion is the F1 score, which provides a balanced evaluation of classification performance under both binary and multi class settings and is particularly suitable for datasets with potential class imbalance [135]. The validation set is used exclusively for model selection and early stopping decisions, while the test set remains strictly isolated and is accessed only once for final performance reporting. This separation prevents information leakage and ensures that the reported test results reflect the true generalization ability of the trained model rather than optimization to the evaluation data.

During inference and testing, a standardized set of output artifacts is recorded to support result verification, quantitative analysis, and reproducibility. For each evaluated model, the predicted labels and corresponding ground truth labels are saved at the window level, enabling consistent computation of performance metrics and detailed error analysis. In addition, confusion matrices and classification reports are generated to provide class-wise performance summaries. Aggregated result files are further produced to summarize validation and test performance across subjects or experimental settings. These artifacts ensure that all reported results can be independently reproduced and facilitate subsequent statistical analysis and visualization. The complete list of recorded outputs and their purposes is summarized in Table 3.10.

Table 3.10: Recorded output artifacts during inference and testing.

Artifact name	Content description	Purpose
Predicted labels	Window-level predicted class labels	Performance evaluation
Ground truth labels	Window-level reference labels	Metric computation
Confusion matrix	Class-wise prediction counts	Error pattern analysis
Classification report	Precision recall and F1 scores per class	Detailed performance summary
Result summary file	Aggregated validation and test metrics	Cross-subject comparison

To ensure that the proposed inference and testing protocol remains applicable across different datasets and task settings, a unified adaptation principle is adopted. All mod-

els operate on window-level signal segments with a consistent input representation, while dataset-specific characteristics such as sampling rate, window length, and the number of classes are handled during preprocessing and label definition. During inference, deployable models rely exclusively on electrocardiogram signals, whereas the teacher model may utilize all available modalities during training. Model outputs are consistently expressed as class logits, which enables uniform application of prediction rules and evaluation metrics across binary and multi-class tasks. Performance assessment follows a standardized metric selection strategy, using accuracy and F1-based measures aligned with the task definition. This design allows the proposed framework to generalize across datasets without modifying the core inference structure, thereby supporting fair comparison and reproducible evaluation.

In summary, this section establishes a unified and reproducible inference and testing protocol for both teacher and student models under different task settings. By clearly separating training and inference roles, enforcing consistent input and output definitions, and standardizing model selection and evaluation procedures, the proposed protocol ensures that performance comparisons are fair and methodologically sound. The exclusive use of electrocardiogram signals during student inference further aligns the evaluation process with realistic deployment constraints. Together, these design choices provide a transparent experimental foundation and prepare the ground for the quantitative results and analysis presented in the following section.

3.3.4 Evaluation Metrics

This subsection defines the evaluation metrics and the assessment protocol adopted in this study to ensure a consistent, fair, and reproducible comparison across all models. The evaluation framework is applied uniformly to the multimodal teacher model, the ECG-only student model, and the distilled student model, regardless of the training strategy or task configuration. All assessments are conducted at the window level, using identical input–output conventions and decision rules, thereby avoiding discrepancies arising from dataset structure or model modality. The selected metrics are designed to capture both overall predictive performance and class-wise discrimination ability, while remaining robust to class imbalance. In addition, the evaluation procedure clearly separates model selection in the validation set from final performance reporting in the test set, ensuring that the reported results reflect generalization performance rather than optimization bias.

During evaluation, all models operate on window-level input samples and follow a unified input–output convention to ensure comparability between datasets and training strate-

gies. Each evaluation batch consists of input windows X and the corresponding ground-truth labels y , where the labels take values from the discrete class set $\{0, \dots, C - 1\}$. Given an input batch, the model produces a matrix of unnormalised prediction scores, referred to as logits, defined as

$$z = f(x) \in \mathbb{R}^{B \times C}, \quad (3.30)$$

where B denotes the batch size, C is the number of target classes, and $f(\cdot)$ represents the forward mapping of the evaluated model. The logits encode the relative confidence of the model for each class before probability normalization. The final class predictions are obtained by selecting the class index with the maximum logit value for each sample [136], according to

$$\hat{y}_i = \operatorname{argmax}_c z_c. \quad (3.31)$$

where z_c denotes the logit corresponding to class c . This decision rule is consistently applied to the teacher model, the baseline student model, and the distilled student model. During evaluation, models are executed in inference mode with gradient computation disabled, and predictions are accumulated across all batches to compute dataset-level metrics. This unified pipeline ensures that performance differences arise solely from model design and training strategy, rather than from inconsistencies in evaluation logic or output interpretation.

To quantify predictive uncertainty and class separation during evaluation, cross-entropy loss is employed as an auxiliary assessment metric for all models. For a dataset containing N evaluation samples, the average cross-entropy loss is defined as

$$\mathcal{L}_{\text{CE}} = \frac{1}{N} \sum_{i=1}^N \left(-\log \frac{\exp(z_{i,y_i})}{\sum_{c=0}^{C-1} \exp(z_{i,c})} \right) \quad (3.32)$$

where $z_{i,c}$ denotes the logit predicted for class c by the model on the i -th sample, y_i represents the corresponding ground-truth class label, and C is the total number of classes. In practice, this loss is computed on a per-batch basis and accumulated across batches, yielding the equivalent averaged form

$$\mathcal{L} = \frac{\text{total_loss}}{N} \quad (3.33)$$

where *total_loss* denotes the sum of batch-wise cross-entropy losses weighted by batch size. Although cross-entropy loss is not used as the primary criterion for model selection in

this study, it provides a consistent scalar measure of prediction confidence and misclassification severity, complementing threshold-based metrics such as accuracy and F1-score in the evaluation protocol.

Accuracy is used as a fundamental evaluation metric to quantify the overall proportion of correctly classified samples across all classes. Let N denote the total number of evaluated windows, y_i the ground-truth label of the i -th sample, and $\hat{y}_i$ the corresponding predicted label obtained from the model logits. The accuracy is defined as

$$\text{Acc} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\hat{y}_i = y_i) \quad (3.34)$$

where $\mathbb{I}(\cdot)$ is the indicator function that equals one if the prediction is correct and zero otherwise. This metric measures the fraction of samples for which the predicted class matches the true class, providing an intuitive summary of classification performance. Accuracy is reported for all models under identical evaluation conditions and is particularly informative when class distributions are approximately balanced. However, because accuracy alone may obscure class-specific errors under imbalanced settings, it is interpreted alongside complementary metrics such as the F1-score in subsequent analyzes.

The F1-score family is adopted as the primary performance metric to evaluate classification effectiveness under both binary and multi-class settings, as it jointly reflects precision and recall and is less sensitive to class imbalance than accuracy alone [137]. For binary classification tasks, the F1-score is computed as

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3.35)$$

where precision is defined as $\text{Precision} = \frac{TP}{TP+FP}$ and recall is defined as $\text{Recall} = \frac{TP}{TP+FN}$. Here, TP , FP and FN denote the numbers of true positives, false positives, and false negatives, respectively. This formulation emphasises correct detection of the positive class while penalising both missed detections and false alarms. For multi-class classification tasks, the Macro-F1 score is employed to ensure equal contribution from each class, regardless of class frequency [138]. It is defined as

$$\text{F1}_{\text{macro}} = \frac{1}{C} \sum_{c=1}^C \text{F1}_c \quad (3.36)$$

where C denotes the total number of classes and F1_c represents the class-specific F1-score computed using a one-versus-rest strategy. This metric provides an unbiased estimate of overall classification quality in imbalanced multi-class scenarios. In addition, a Weighted-F1 score may be reported for completeness, defined as

$$\text{F1}_{\text{weighted}} = \sum_{c=1}^C \frac{N_c}{N} \text{F1}_c \quad (3.37)$$

where N_c is the number of samples in class c and N is the total number of samples. While Weighted-F1 reflects dataset class distribution, Macro-F1 is selected as the primary metric for model selection in multi-class tasks, ensuring that performance improvements are not dominated by majority classes.

The confusion matrix is used as a complementary diagnostic tool to provide a detailed class-wise analysis of the model predictions beyond the aggregate performance metrics. It records the distribution of predicted labels against the corresponding ground-truth labels, thereby revealing specific patterns of misclassification that may not be apparent from accuracy or F1-score alone. By examining the confusion matrix, it becomes possible to identify whether classification errors are concentrated in particular classes or arise from systematic confusion between specific class pairs. This information is particularly valuable for analyzing model behavior under class imbalance and for understanding failure modes in both binary and multi-class tasks. In this study, confusion matrices are generated during the final testing stage using the best validation-selected checkpoint and reported alongside summary metrics to support a transparent and interpretable performance evaluation.

The evaluation procedure is consistently applied across the training, validation, and testing stages in order to ensure fair model comparison and reliable performance reporting. As summarised in Table 3.11, the training stage focuses on monitoring optimisation behaviour by reporting loss, accuracy, and task appropriate F1 scores at the end of each epoch, which are used solely for convergence analysis. The validation stage serves a distinct role in model selection, where performance metrics are computed without parameter updates and the checkpoint yielding the best validation F1 score is selected as the final model. Importantly, the test stage is strictly reserved for final performance assessment and is executed only once using the selected checkpoint, thereby avoiding any information leakage into model development [139]. During testing, accuracy, F1 score, confusion matrix, and classification report are generated to provide both quantitative and diagnostic insights. This unified evaluation protocol ensures that optimisation, model selection, and performance reporting are clearly separated, improving reproducibility and methodological transparency.

Table 3.11: Evaluation procedure across training, validation, and testing stages.

Stage	Purpose	Reported outputs
Training	Model optimisation and convergence monitoring	Training loss, accuracy, F1 score computed at epoch end
Validation	Model selection and early stopping criterion	Validation loss, accuracy, F1 score used to select best checkpoint
Testing	Final performance evaluation	Test accuracy, F1 score, confusion matrix, classification report

The reported outputs and recorded artifacts are designed to support transparent evaluation and reproducibility of all experimental results. For each evaluated model, window-level predicted labels and corresponding ground truth labels are preserved to enable consistent metric computation and independent verification. In addition, summary performance metrics, including accuracy and F1-based measures, are recorded for each experimental setting. Diagnostic outputs such as confusion matrices and detailed classification reports are also generated to facilitate error analysis and interpretation of model behaviour. All results are stored in a structured and consistent format across datasets and tasks, ensuring that experimental findings can be reliably reproduced and that subsequent statistical analysis and visualization can be conducted without ambiguity.

In summary, this subsection has established a unified evaluation framework that defines how model performance is assessed across different tasks, datasets, and training configurations. By adopting window-level evaluation, clearly separating training, validation, and testing stages, and selecting metrics that are robust to class imbalance, the proposed approach ensures fair and reproducible performance comparison. The combination of aggregate metrics and diagnostic outputs further supports transparent interpretation of model behaviour. While these evaluation choices provide a consistent basis for quantitative analysis, they also rely on several methodological assumptions related to data segmentation, label definition, and deployment constraints. These assumptions, along with the inherent limitations of the proposed methodology, are discussed in the following section.

3.4 Methodological Assumptions and Constraints

This section outlines the key methodological assumptions and constraints underlying the proposed framework, with the aim of clarifying its scope of applicability and interpretation. These assumptions are introduced to ensure consistency in model training, evaluation, and cross-setting comparison. Rather than reflecting experimental limitations, they represent deliberate design choices that balance methodological rigor, reproducibility, and practical deployment considerations. By explicitly stating these assumptions, the evaluation results presented in subsequent sections can be interpreted within a clearly defined methodological context.

A primary assumption of this study concerns window-level analysis and label assignment. Each signal window is treated as an independent sample, and temporal dependencies across adjacent windows are not explicitly modeled [140]. In addition, labels are assumed to remain constant within each window, implying that transitional states occurring inside a window are not separately represented [141]. This assumption simplifies both training and evaluation by enabling a uniform window-based classification formulation across datasets. Furthermore, dataset-provided annotations are treated as ground truth without additional modeling of annotation uncertainty or subject-specific variability. These choices facilitate consistent evaluation and fair comparison across different experimental conditions.

Several constraints arise from the model design and deployment objectives. In particular, the student model is strictly constrained to operate using electrocardiogram signals only during inference, reflecting practical requirements for wearable and low-complexity monitoring systems. As a result, potentially informative signals such as electromyogram and electrodermal activity are excluded at deployment, even though they are available during teacher training. In addition, the student model adopts a compact architecture to reduce computational complexity, which inherently limits representational capacity. Knowledge distillation is therefore confined to the training stage and does not introduce additional components or overhead during inference. These constraints are intentional and align the methodological design with realistic application scenarios.

In summary, the assumptions and constraints described above define the conditions under which the proposed framework is developed and evaluated. They enable a unified and reproducible experimental pipeline while maintaining a clear focus on deployable ECG-based inference. At the same time, these choices may restrict performance in scenarios that require long-term temporal modeling or multi-sensor deployment at inference.

Understanding these boundaries is essential for contextualizing the reported results and for guiding future extensions of the proposed approach.

3.5 Chapter Summary

This chapter has presented the methodological framework adopted in this study, covering research design, signal processing, model architecture, training strategy, inference protocol, and evaluation criteria. A teacher–student learning paradigm with knowledge distillation was introduced to leverage multimodal supervision during training while enforcing ECG-only inference at deployment. Unified procedures were defined for training, evaluation, and testing to ensure fair comparison and reproducibility across datasets and task settings. Key methodological assumptions and constraints were also discussed to clarify the scope and applicability of the proposed approach. Building on this methodological foundation, the next chapter reports the experimental results and provides a quantitative evaluation of model performance under the defined protocols.

DATASETS AND EXPERIMENTAL DESIGN

This chapter introduces the datasets and the experimental design adopted in this study. Three datasets are used to support the evaluation of the proposed framework, including two publicly available benchmark datasets and one self collected multi-modal dataset. The selected datasets cover different experimental settings, data acquisition conditions, and labeling strategies, which together enable a comprehensive assessment of model performance and generalization. By presenting the data sources and the experimental context in a unified manner, this chapter establishes a consistent foundation for the methodological and evaluative analyzes reported in subsequent chapters.

The chapter is organized into three main parts. First, the public benchmark datasets are briefly described with a focus on their experimental design, signal modalities, and usage within this study. Second, the self collected dataset is introduced, including the experimental procedure, data acquisition process, and annotation strategy. Finally, the data organization and usage protocol is described to clarify how all datasets are processed, segmented, and incorporated into the training and evaluation pipeline. This structured presentation ensures transparency, reproducibility, and consistency between all experimental results.

4.1 Public Dataset

In this study, public benchmark datasets are adopted to provide a standardized and reproducible basis for evaluating the proposed framework. Compared with exclusively relying on self-collected data, public datasets enable direct comparison with existing studies and

help assess the generalizability of the method in well-established experimental settings. In addition, the use of multiple benchmark datasets reduces the risk of dataset-specific bias and allows the proposed approach to be examined in different experimental paradigms and subject populations. In this work, public datasets serve primarily as reference benchmarks for methodological validation, rather than as the focus of empirical investigation.

Two widely used public benchmark datasets, WESAD and MAUS, are used. Both datasets provide multimodal physiological recordings, with ECG serving as a core signal for stress related analysis. The WESAD dataset includes synchronized ECG, EMG, and EDA signals collected under controlled experimental conditions, together with clearly defined condition level annotations [142]. The MAUS dataset also provides multimodal recordings and is distributed as three separate signal files, namely ECG, EMG, and EDA. In the MAUS dataset used in this study, the EMG and EDA channels correspond to recordings based on PPG and GSR [143]. In the present work, ECG signals are consistently used as the primary input modality, while EMG and EDA related signals are incorporated only during the training stage for multimodal representation learning.

The selected public benchmark datasets are used to support systematic training, validation, and evaluation of the proposed framework. Although multimodal physiological signals are available in both datasets, the final inference setting is strictly constrained to only ECG signals, in order to reflect practical deployment scenarios. Additional modalities including EMG and EDA related signals are incorporated during the training stage to enhance representation learning, but are not required at inference time. This unified usage strategy is applied consistently across public datasets to ensure fair evaluation and comparability. In addition to the public benchmark datasets, a self collected multimodal dataset is constructed to further examine the proposed approach under a customized experimental setting.

4.2 Self-Collected Multimodal Stress Dataset

This study constructs a self-collected multimodal physiological dataset designed to investigate cognitive stress under controlled experimental conditions. The dataset was developed to examine how varying levels of mental workload influence physiological responses during task execution. Mental arithmetic tasks with graded difficulty were adopted as the primary stress induction paradigm [144]. These tasks were selected because of their well-established ability to elicit a stable and repeatable cognitive load. The collected data support both stress detection and difficulty-aware classification, with ECG signals serving as

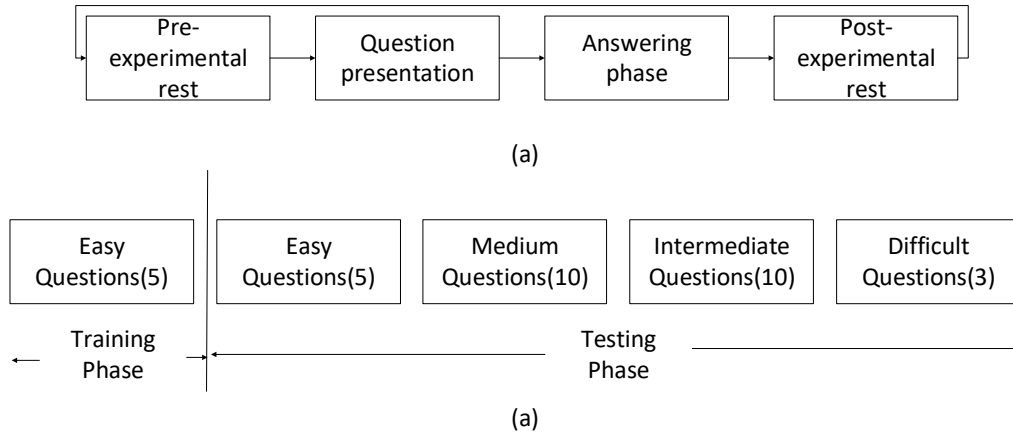


Figure 4.1: Overview of the experimental task design and trial structure. (a) Schematic illustration of the temporal structure within a single trial, including the pre-experimental rest, question presentation, answering phase, and post-experimental rest. (b) Organization of the training and testing phases, showing the distribution of mental arithmetic questions across different difficulty levels, including easy, medium, intermediate, and difficult conditions.

the central modality for subsequent analysis.

The experimental protocol was reviewed and approved by the University of Technology Sydney Human Research Ethics Committee under approval number ETH24-10203. A total of twenty one participants were recruited for this study, including 9 females and 12 males, with ages ranging from 18 to 25 years. All participants reported normal physical and mental health, with no known cardiovascular, neurological, or psychological conditions. The experiments were conducted in a quiet laboratory environment, where the participants were seated comfortably and instructed not to be moved or interrupted by unnecessary verbal interaction. Each experimental session lasted approximately 20 minutes and each participant completed five sessions on separate days with at least one day between sessions, to reduce fatigue effects and minimize accidental or transient influences.

During the experiment, participants performed a mental arithmetic task designed with multiple difficulty levels, including easy, normal, medium, and hard conditions. The general structure of the task and the sequence of the tests are illustrated in Figure 4.1. Each trial followed a fixed conceptual structure consisting of a rest period, a task related phase, and a post task rest period. During the task related phase, arithmetic questions were presented on the computer screen, and participants selected their answers using a mouse. Visual cues on the screen and brief auditory signals were used to indicate transitions between different phases, ensuring a consistent experimental rhythm across trials. A short training phase preceded the testing phase to familiarize participants with the task format and response

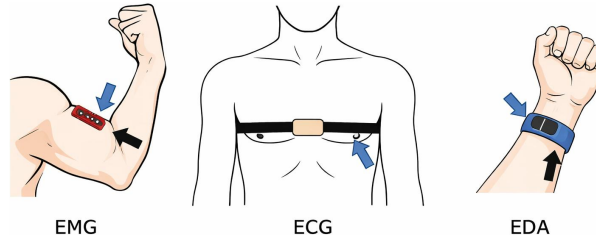


Figure 4.2: Illustration of multimodal physiological sensor placement.

procedure.

For each participant, the experimental session was organized into a training stage followed by a formal testing stage with a fixed trial structure. An initial set of trials was used for training and task familiarization, and these trials were excluded from all subsequent analyses. The formal experiment consisted of a predefined number of trials, each assigned to a single difficulty level. Specifically, the testing stage included easy, normal, medium and hard trials, with a fixed number of trials per difficulty condition. Each trial corresponded to one complete execution of the task sequence and was associated with exactly one difficulty label. It is important to note that trials represent task level units, whereas signal windows used in later analysis were derived from within trials. Therefore, task difficulty was treated as a trial level attribute rather than a window level label.

During the experiment, multimodal physiological signals were collected, including ECG, EDA, and EMG. Among these modalities, ECG signals were treated as the core signal for subsequent analysis, while EDA and EMG provided complementary information related to autonomic and muscular responses. The ECG sensor was worn on the chest near the nipple line to ensure stable cardiac signal acquisition. The EDA device was worn on the wrist, allowing continuous measurement of skin conductance changes. EMG sensors were placed on the forearm to capture muscle activity associated with task execution. A schematic illustration of sensor placement is shown in Figure 4.2. Throughout the experiment, participants were instructed to maintain a relaxed seated posture and minimize body movement in order to reduce motion related artifacts and ensure signal quality.

All physiological signals in the self-collected dataset were acquired in a synchronized manner under a unified experimental control program. The ECG signals were sampled at a rate of 512 Hz, while the EMG signals were sampled at 250 Hz. Despite differences in native sampling frequencies, all modalities were recorded within the same experimental timeline. A shared timestamp reference was used to align data streams across devices. In addition, experiment state markers were applied consistently to all modalities, enabling precise temporal correspondence between physiological signals and task phases.

The dataset adopts a state based annotation strategy to describe temporal structure within each trial. Labels do not represent emotional or stress categories. Instead, they indicate distinct experimental phases occurring during task execution. These phases include resting periods and task related periods that are consistently defined across trials. Each signal sample is associated with a phase label, a trial identifier, and a task difficulty level. This annotation design preserves the temporal dynamics of the experiment and supports window based analysis. It also ensures consistency between the data labeling and the modeling settings used in subsequent chapters.

After acquisition, the raw multimodal signals were processed through a structured pre-processing pipeline to ensure data consistency and usability for model training. Original recordings were first converted from device-specific data streams into continuous physiological waveforms. Signals from different modalities were then temporally aligned and re-sampled to a unified sampling frequency to support synchronized analysis. Training trials conducted at the beginning of each session were systematically removed to avoid learning effects and adaptation bias. All processed data were organized at the subject level, with a clear separation between sessions and trials. The final outputs were stored in standardized data formats suitable for downstream machine learning experiments and reproducible evaluation.

The proposed dataset is intended for controlled cognitive stress analysis under laboratory conditions and is particularly suitable for ECG-centered modeling approaches. By combining graded mental workload tasks with synchronized physiological recordings, the dataset provides a reliable basis for investigating stress-related patterns in a structured setting. However, the experimental design does not aim to capture stress responses in naturalistic daily environments, and the resulting data should be interpreted within this controlled context [145]. In the following chapters, this dataset is used as the primary experimental foundation for model development and performance evaluation.

In summary, the self-collected multimodal stress dataset provides a well-structured and ethically approved experimental foundation for studying cognitive stress under controlled conditions. The dataset integrates synchronized physiological signals with clearly defined task structure and annotation strategy, ensuring consistency across subjects and sessions. This design supports systematic analysis and reproducible modeling, and it establishes a clear link between experimental protocol, data organization, and subsequent methodological development.

4.3 Data Organization and Usage Protocol

All datasets used in this study follow a unified data organization and usage protocol to ensure fair comparison, reproducibility, and consistency across experiments. Regardless of whether the data originate from public benchmarks or the self-collected dataset, the same principles are applied for preprocessing, segmentation, and evaluation. This unified strategy minimizes dataset-specific bias and establishes a common experimental foundation for subsequent methodological analysis.

Data are organized at the subject level, with each subject containing one or more recording sessions composed of continuous physiological signals. Temporal alignment is maintained between signals and corresponding annotations along a shared time axis. Although trial information is preserved as a logical structure, the physiological recordings remain continuous, which supports flexible segmentation and avoids artificial boundary effects during analysis.

For model training and evaluation, continuous signals are segmented into fixed-length windows that serve as basic learning units. Each window is derived from within a single trial and inherits trial-level attributes such as task difficulty [146]. In this framework, trials represent experimental units that define task context, while windows represent learning units used for classification. This distinction ensures that task-level information is preserved without introducing label ambiguity at the signal level.

Different types of labels play distinct roles within the analysis pipeline. Phase labels describe the temporal state of the experiment and are used to guide segmentation and state-aware analysis. Task difficulty labels are treated as trial-level attributes and are used as classification targets or auxiliary supervisory signals, depending on the experimental setting. Stress is not assigned as a direct categorical label but is implicitly modeled through task-induced cognitive load, which aligns the annotation strategy with the experimental design.

Training, validation, and testing splits are performed in a subject-independent manner to prevent data leakage and ensure unbiased evaluation [147]. In all datasets, electrocardiogram signals are the only modality used during inference, reflecting practical deployment constraints. Additional modalities, including electromyography and electrodermal activity, are incorporated exclusively during training to support multimodal representation learning. This usage protocol is applied consistently to both public and self-collected datasets, enabling transparent comparison and reliable assessment in the following chapters.

RESULTS

This chapter presents the experimental results obtained to evaluate the proposed framework across multiple datasets and classification tasks. The evaluation focuses on subject dependent classification performance under different task complexities, including binary and multi class settings. Results are reported using standard performance metrics to provide a clear and consistent comparison between models. In addition to overall performance, detailed analyzes are conducted to examine the effect of knowledge distillation, class wise prediction behavior, and training dynamics. Together, these results offer a comprehensive assessment of the proposed approach under the considered experimental conditions.

5.1 Overview of Evaluation Tasks and Metrics

This section provides an overview of the evaluation tasks and performance metrics adopted in the experimental analysis. To comprehensively assess the proposed framework under different classification complexities and data conditions, experiments were conducted on three datasets, including two public benchmark datasets and one self collected dataset. For each dataset, both binary and multi class classification tasks were considered when applicable. This design enables a systematic evaluation of model behavior across varying numbers of target classes and ensures that performance trends can be examined under both simpler and more challenging classification settings. A summary of all evaluation tasks is presented in Table 5.1, which outlines the dataset sources, task definitions, and corre-

Table 5.1: Evaluation tasks overview.

Dataset	Task	#Classes	Metric
WESAD	2-class	2	Accuracy, Macro F1
WESAD	4-class	4	Accuracy, Macro F1
MAUS	2-class	2	Accuracy, Macro F1
MAUS	3-class	3	Accuracy, Macro F1
MyDB	2-class	2	Accuracy, Macro F1
MyDB	4-class	4	Accuracy, Macro F1

sponding class numbers.

As shown in Table 5.1, the WESAD and MAUS datasets were evaluated under two-classification configurations each, while the self collected dataset was assessed using both two-class and four class settings. This unified task configuration allows results from different datasets to be compared in a consistent manner. Model performance was evaluated using Accuracy and Macro F1 score across all tasks. Accuracy reflects the overall classification correctness, while Macro F1 provides a balanced assessment by equally weighting each class, which is particularly important for multi class problems. By adopting the same evaluation metrics across all datasets and task settings, the reported results remain directly comparable and facilitate a clear interpretation of performance differences in subsequent sections.

All evaluations are conducted under a subject dependent setting, where training and testing samples are drawn from the same subject. In this study, Cross subject evaluation is not considered. Performance results are reported at the subject level and are summarized using descriptive statistics to facilitate comparison across datasets and task types. This evaluation protocol establishes a consistent basis for the result analyzes presented in the following sections.

5.2 Overall Classification Performance

Table 5.2 presents the overall classification performance obtained across all datasets and task settings. The reported results summarize the average model behavior over all subjects for each evaluation task, with performance expressed as mean values with standard deviation. Across the three datasets, the model achieved stable classification accuracy under both binary and multi class conditions, indicating that the proposed framework maintains reliable performance across different levels of task complexity. The inclusion of multiple datasets in this evaluation further supports the general applicability of the reported results.

Table 5.2: Overall classification performance of the distilled model across datasets and tasks.

Dataset	Task	Accuracy (%)	Macro F1	n
WESAD	2-class	88.85 ± 4.40	0.869 ± 0.045	15
WESAD	4-class	80.71 ± 4.28	0.782 ± 0.044	15
MAUS	2-class	88.07 ± 3.94	0.864 ± 0.039	22
MAUS	3-class	78.82 ± 5.05	0.755 ± 0.046	22
MyDB	2-class	74.86 ± 5.19	0.723 ± 0.052	21
MyDB	4-class	64.87 ± 5.17	0.622 ± 0.051	21

As shown in Table 5.2, higher performance is consistently observed for binary classification tasks compared with their multi class counterparts. On the WESAD and MAUS datasets, two-class tasks achieved accuracy values close to or above eighty percent, while performance decreased when the number of target classes increased. A similar trend is observed on the self collected dataset, where the four class task exhibits lower accuracy and Macro F1 compared with the corresponding two-class setting. These results reflect the increased difficulty associated with finer class distinctions and demonstrate consistent performance trends across different data sources.

In addition to accuracy, Macro F1 provides further insight into classification performance under varying class distributions. As reported in Table 5.2, Macro F1 values follow trends similar to those observed for accuracy, with lower scores in multi class tasks. Despite this expected reduction, the model maintains balanced performance across classes, as reflected by the relatively stable Macro F1 values across datasets. Overall, the results in Table 5.2 indicate that the proposed approach achieves robust and consistent classification performance across diverse datasets and task configurations.

5.3 Effect of Knowledge Distillation

This section examines the effect of knowledge distillation on classification performance across different datasets and task configurations. The comparison focuses on the performance difference between the distilled student model and the corresponding student model trained without distillation. Across all evaluated datasets and classification tasks, the distilled student consistently achieves higher accuracy than the non distilled counterpart. This improvement is observed under both binary and multi class settings, indicating that the benefit of knowledge distillation is not restricted to a specific dataset or task complexity. The teacher model is included as a reference to illustrate the performance upper bound,

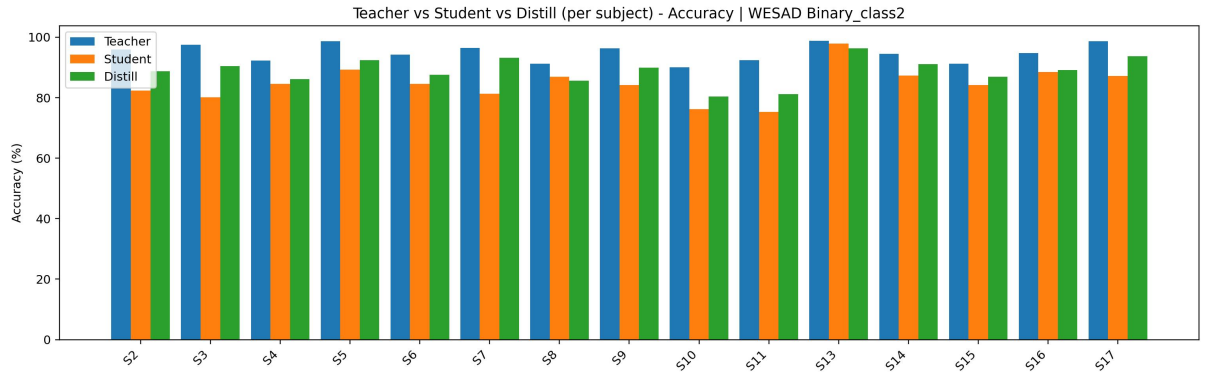


Figure 5.1: Per subject classification accuracy comparison between the teacher model, the baseline student model, and the distilled student model on the WESADB two-class task.

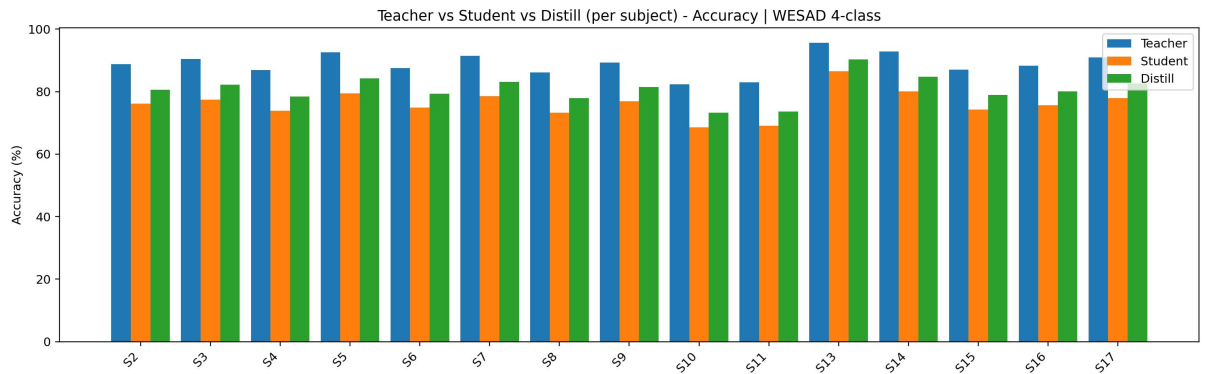


Figure 5.2: Per subject classification accuracy comparison between the teacher model, the baseline student model, and the distilled student model on the WESADB four class task.

while the primary emphasis remains on the relative gain achieved by the distilled student model.

The per subject accuracy results on the WESAD dataset further illustrate the effect of knowledge distillation under both binary and multi class settings. As shown in the Figure 5.1, and Figure 5.2, the distilled student model consistently outperforms the baseline student model for the majority of subjects. This improvement is observed across nearly all evaluated subjects, indicating that the performance gain is not driven by a small subset of individuals. In addition, the relative improvement of the distilled model is more pronounced in the four class task, where the classification problem is more challenging. These observations suggest that knowledge distillation provides a stable and subject independent performance benefit on the WESAD dataset across different task complexities.

Similar performance trends are observed on the MAUS dataset for both two-class and three-class tasks. Although the overall accuracy levels are lower than those reported for the WESAD dataset, the distilled student model maintains a consistent advantage over the

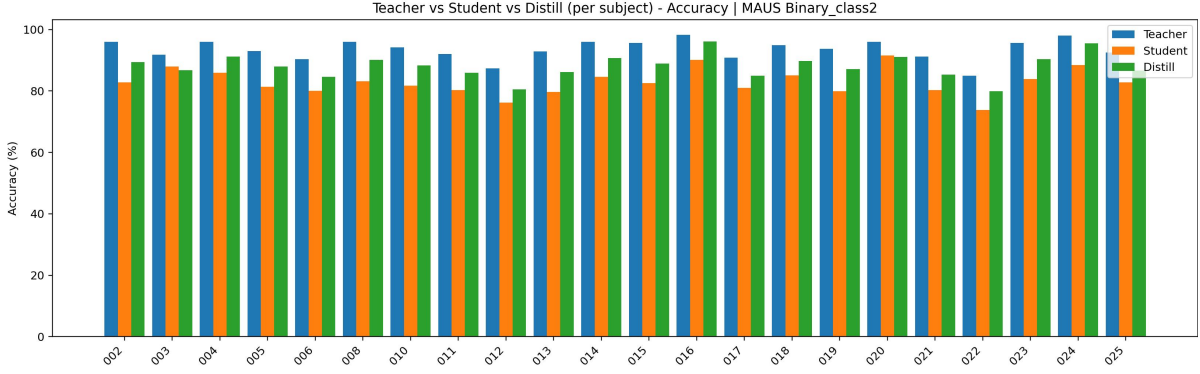


Figure 5.3: Per subject classification accuracy comparison between the teacher model, the baseline student model, and the distilled student model on the maus two-class task.

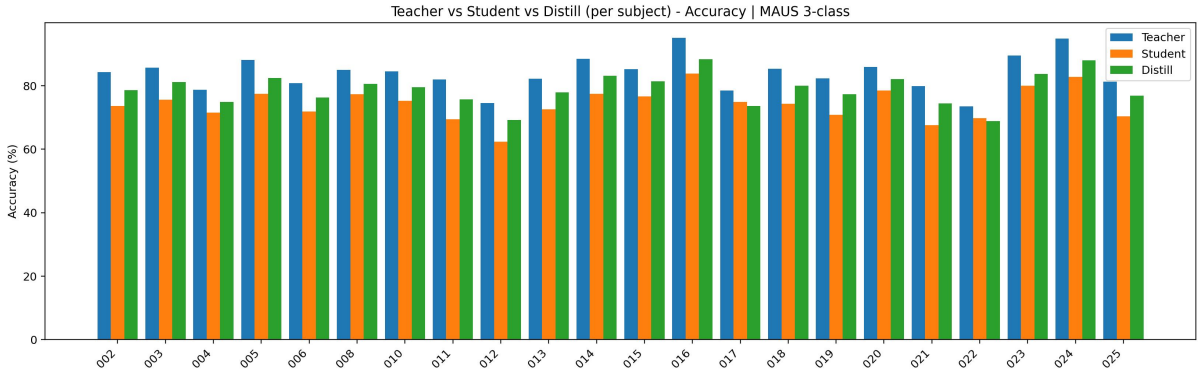


Figure 5.4: Per subject classification accuracy comparison between the teacher model, the baseline student model, and the distilled student model on the maus three-class task.

baseline student model across most subjects, indicating that the benefit of knowledge distillation remains stable under different data characteristics.

The self collected MyDB dataset exhibits a similar performance pattern to the public benchmarks, while overall accuracy levels are further reduced compared with the MAUS dataset. Despite this decrease, the distilled student model continues to outperform the baseline student model for most subjects in both two-class and four class tasks, indicating that the positive effect of knowledge distillation persists under more challenging and noisy data conditions.

When comparing results across datasets and task configurations, a consistent pattern emerges regarding the effect of knowledge distillation. While the absolute classification accuracy varies across WESAD, MAUS, and MyDB, the relative performance gain of the distilled student model over the baseline student model remains stable. In particular, the improvement introduced by knowledge distillation is more evident in multi class tasks, where the classification problem is inherently more challenging. This observation holds across

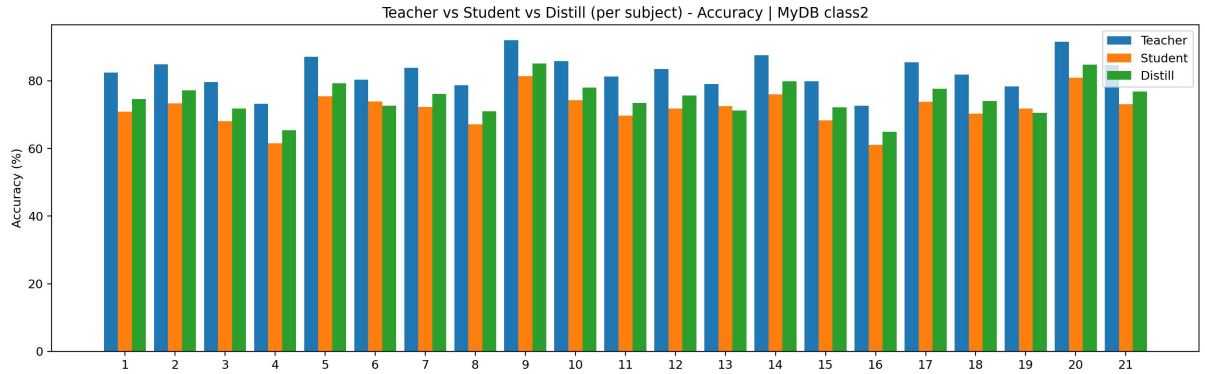


Figure 5.5: Per subject classification accuracy comparison between the teacher model, the baseline student model, and the distilled student model on the maus two-class task.

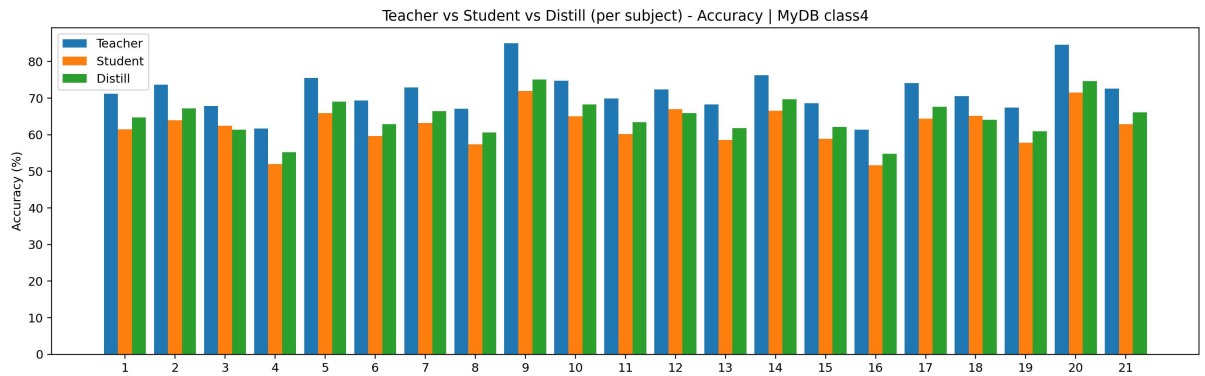


Figure 5.6: Per subject classification accuracy comparison between the teacher model, the baseline student model, and the distilled student model on the maus three-class task.

datasets with different data sources, signal characteristics, and experimental conditions, indicating that the effectiveness of knowledge distillation is not limited to a specific dataset or task setting.

Overall, the results presented in this section demonstrate that knowledge distillation provides a reliable and consistent performance improvement for the student model across diverse datasets and classification tasks. The observed gains are not driven by isolated subjects or specific experimental settings, but instead reflect a general trend that persists under varying levels of task complexity and data quality. These findings establish a solid empirical basis for the use of knowledge distillation in ECG based stress classification and motivate further discussion on its practical implications and limitations in the following section.

Although direct comparisons with existing ECG-only methods could provide additional reference points, such comparisons were not the primary focus of this study. The main objective of this work is to investigate whether multimodal physiological information can be used effectively during training to enhance ECG-only inference, rather than to establish a

new state-of-the-art ECG-only model in specific benchmark settings. Existing ECG-only approaches often differ substantially in dataset characteristics, preprocessing pipelines, task definitions, and evaluation protocols, which makes direct and fair comparison challenging and potentially misleading. Instead, this study adopts a controlled teacher–student and ablation-based evaluation framework, where the contribution of multimodal training is isolated by comparing models trained with and without access to complementary physiological signals under identical conditions. This design allows the performance gains introduced by multimodal-informed training to be examined in a transparent and internally consistent manner, without confounding effects introduced by heterogeneous experimental settings.

5.4 Confusion Matrix and Class-wise Performance

This section presents a class wise performance analysis based on confusion matrices in order to complement the overall metrics reported in previous sections. While Accuracy and Macro F1 provide a global assessment of classification performance, confusion matrices offer more detailed insight into error distribution across different classes. To avoid redundancy and maintain clarity, only representative tasks are selected for visualization and analysis. All confusion matrices are presented in a row normalized form, which enables a clearer comparison of class specific prediction behavior by reducing the influence of class sample size. The analysis in this section focuses on two aspects, including model level comparison within the same dataset and pattern comparison across different datasets under multi class settings.

The class wise comparison on the WESAD multi class task further highlights the effect of knowledge distillation on error distribution. As illustrated in Figure 5.7, the baseline student model exhibits notable inter class confusion, as reflected by the increased off diagonal values in the row normalized confusion matrix shown in Figure 5.7(c). In contrast, the distilled student model demonstrates a clearer separation between classes, with stronger diagonal dominance and reduced misclassification rates across multiple classes, as shown in Figure 5.7(b). When compared with the teacher model in Figure 5.7(d), the distilled student presents a prediction pattern that more closely aligns with the teacher, indicating a consistent shift toward improved class wise discrimination. These observations suggest that knowledge distillation not only improves overall accuracy, but also contributes to a more balanced and structured distribution of classification errors at the class level.

The cross dataset comparison of class wise performance further demonstrates the con-

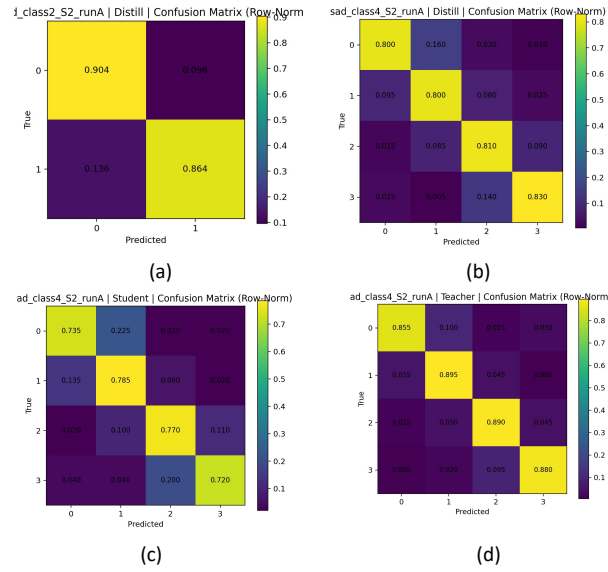


Figure 5.7: Row normalized confusion matrices illustrating class wise prediction performance under different classification settings on the WESAD dataset. (a) Confusion matrix of the distilled student model for the binary classification task, showing clear class separation with high diagonal dominance. (b) Confusion matrix of the distilled student model for the four class task, where increased off diagonal entries indicate higher classification difficulty under finer class granularity. (c) Confusion matrix of the baseline student model for the four class task, exhibiting more pronounced inter class confusion and reduced diagonal values compared with the distilled model. (d) Confusion matrix of the teacher model for the four class task, serving as a performance reference and demonstrating stronger class separation. All matrices are row normalized to highlight class specific error distributions and enable direct comparison of prediction behavior across models and task settings.

sistency of prediction patterns produced by the distilled student model. As shown in Figure 5.8, all three datasets exhibit dominant diagonal elements, indicating that the model maintains a clear primary classification tendency across different data sources. However, the degree of inter class confusion varies across datasets. The WESAD dataset in Figure 5.8(a) shows relatively stronger class separation, while the MyDB dataset in Figure 5.8(b) presents increased confusion between certain classes, reflecting higher task difficulty. The MAUS dataset in Figure 5.8(c) exhibits intermediate behavior, with moderate diagonal dominance and localized confusion patterns. Despite these differences in absolute performance, the overall structure of the confusion matrices remains similar across datasets, suggesting that the distilled model preserves consistent class wise discrimination behavior under varying data conditions.

Overall, the confusion matrix analysis provides complementary evidence to the quantitative metrics reported in earlier sections. The observed error distributions indicate that

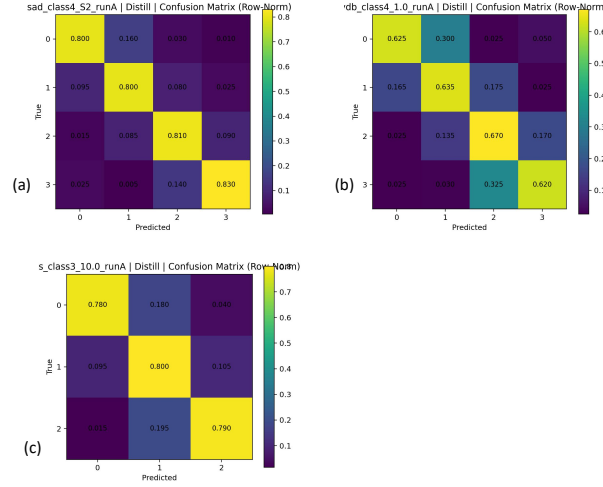


Figure 5.8: Row normalized confusion matrices of the distilled student model for multi class classification tasks across different datasets. (a) Confusion matrix for the four class task on the WESAD dataset, showing relatively strong diagonal dominance with moderate inter class confusion. (b) Confusion matrix for the four class task on the MyDB dataset, where increased off diagonal values indicate higher classification difficulty under more complex and noisy data conditions. (c) Confusion matrix for the three-class task on the MAUS dataset, illustrating intermediate class separation performance between WESAD and MyDB. All confusion matrices are row normalized to emphasize class wise error distribution and enable a direct comparison of prediction patterns across datasets with different characteristics and task definitions.

knowledge distillation improves class wise discrimination by reducing inter class confusion and strengthening diagonal dominance across multiple tasks. These improvements are consistently observed within the same dataset and across different datasets under multi class settings. The alignment between the confusion matrix patterns and the previously reported performance metrics further supports the effectiveness of the proposed approach. In the following section, the training dynamics and convergence behavior of the models are examined to provide additional insight into the optimization process.

5.5 Training Dynamics and Convergence Behavior

This section analyzes the training dynamics and convergence behavior of the evaluated models in order to further validate the reliability of the reported performance gains. While the previous sections focus on final classification results and class wise error patterns, examining the training process provides additional insight into whether these improvements are achieved through stable and consistent optimization. To this end, representative train-

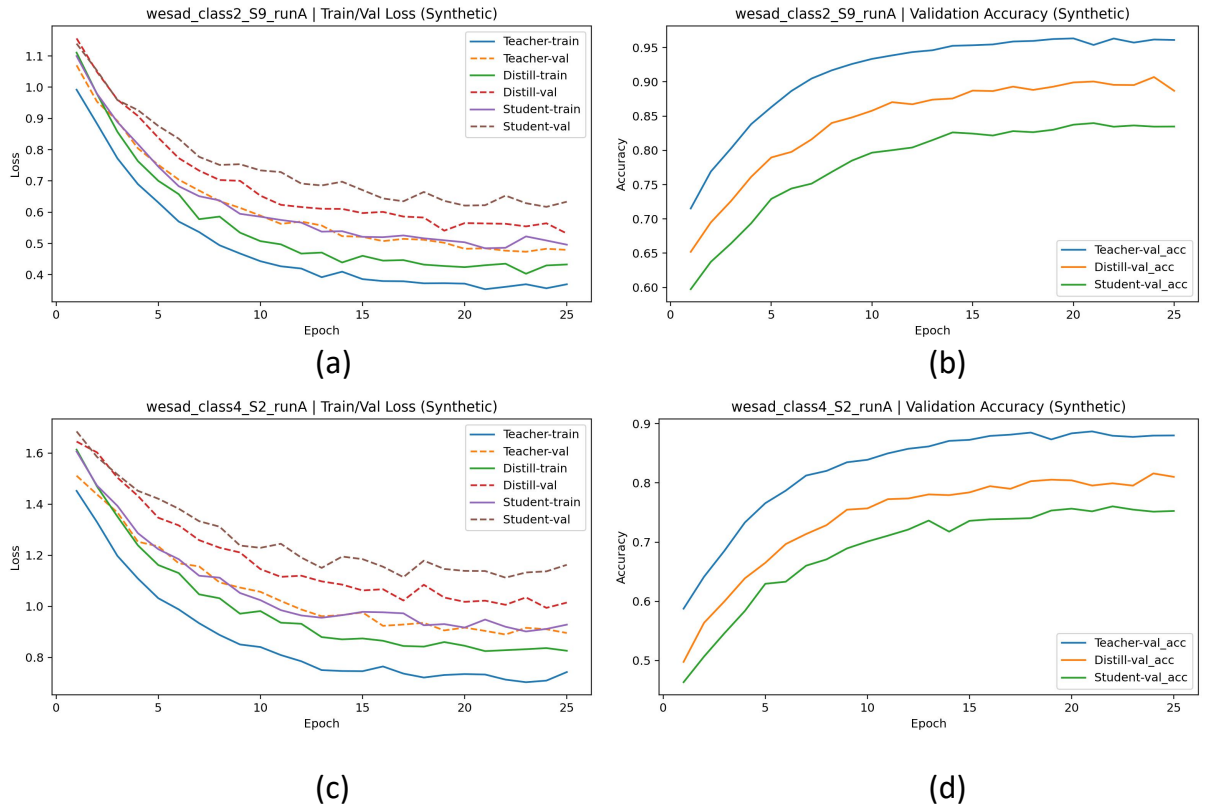


Figure 5.9: Training dynamics and convergence behavior of the evaluated models on the WESAD dataset under different classification settings. (a) Training loss curves for the binary classification task, illustrating the optimization process of the teacher, student, and distilled student models. (b) Validation accuracy curves for the binary classification task, showing the evolution of predictive performance across training epochs. (c) Training loss curves for the four class task, where increased task complexity leads to higher loss values and slower convergence. (d) Validation accuracy curves for the four class task, highlighting differences in convergence speed and final performance among the three models. In all cases, the distilled student model demonstrates smoother convergence and improved validation performance compared with the baseline student model, while the teacher model serves as a reference upper bound.

ing and validation curves are presented to illustrate convergence characteristics, followed by a statistical analysis that quantifies the significance of performance differences between models. Together, these analyses aim to ensure that the observed improvements are not attributable to random variation or unstable training behavior.

The training and validation curves shown in Figure 5.9 provide further insight into the convergence behavior of the models under different task complexities on the WESAD dataset. For the binary classification task, the distilled student model converges more rapidly than the baseline student model and achieves higher validation accuracy, as illustrated in Fig-

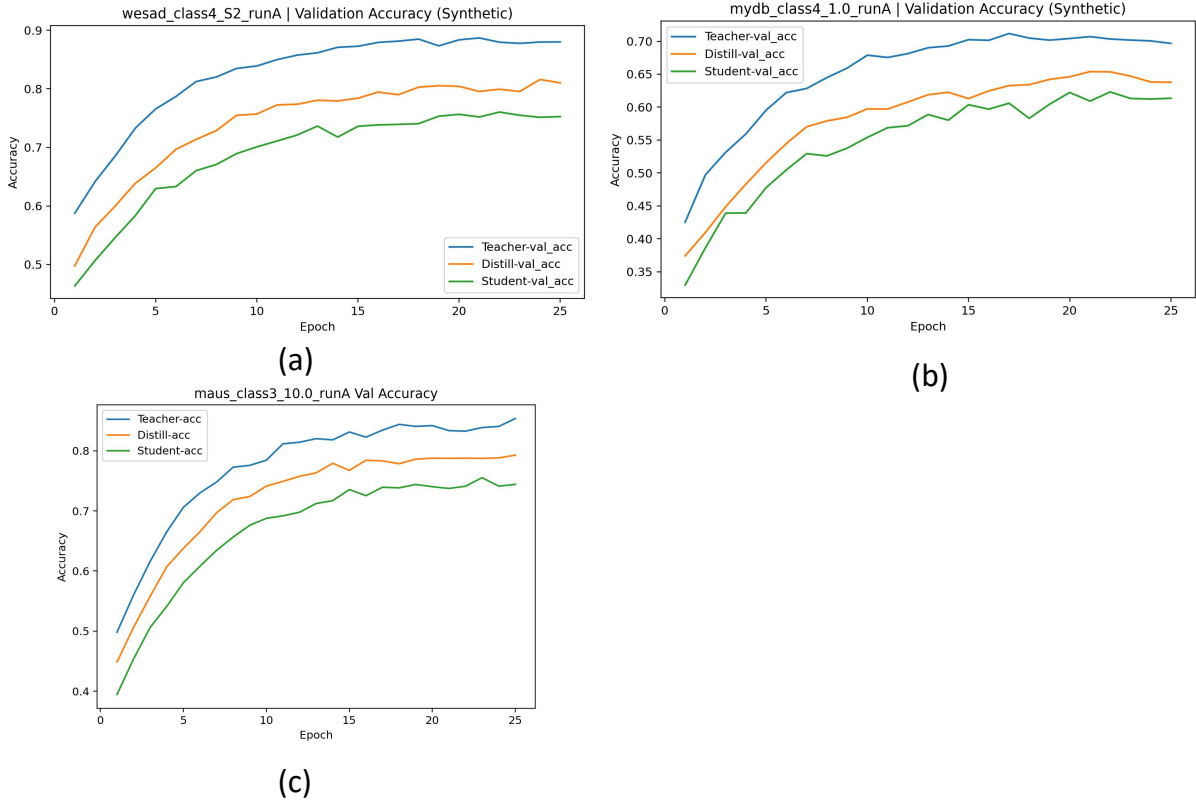


Figure 5.10: Validation accuracy curves for multi class classification tasks across different datasets. (a) Validation accuracy on the WESAD four class task, illustrating the convergence behavior of the teacher, student, and distilled student models under a public benchmark setting. (b) Validation accuracy on the MyDB four class task, where overall accuracy levels are reduced due to increased data complexity and noise. (c) Validation accuracy on the MAUS three-class task, showing intermediate convergence performance between WESAD and MyDB. Across all datasets, the distilled student model consistently achieves higher validation accuracy than the baseline student model throughout training, while the teacher model serves as a reference upper bound.

ure 5.9(b). A similar pattern is observed for the four class task, where increased task difficulty results in slower convergence and lower overall accuracy for all models, as shown in Figure 5.9(d). Despite this increase in complexity, the distilled student maintains a consistent advantage over the student model throughout the training. In addition, the corresponding loss curves in Figure 5.9(a) and Figure 5.9(c) exhibit smooth and stable decreases without signs of divergence, indicating that the observed performance improvements are achieved through stable optimization rather than unstable training behavior.

The cross dataset comparison of training dynamics further confirms the consistency of the distilled student model under different data conditions. As shown in Figure 5.10, all three datasets exhibit similar convergence trends in validation accuracy, despite noticeable

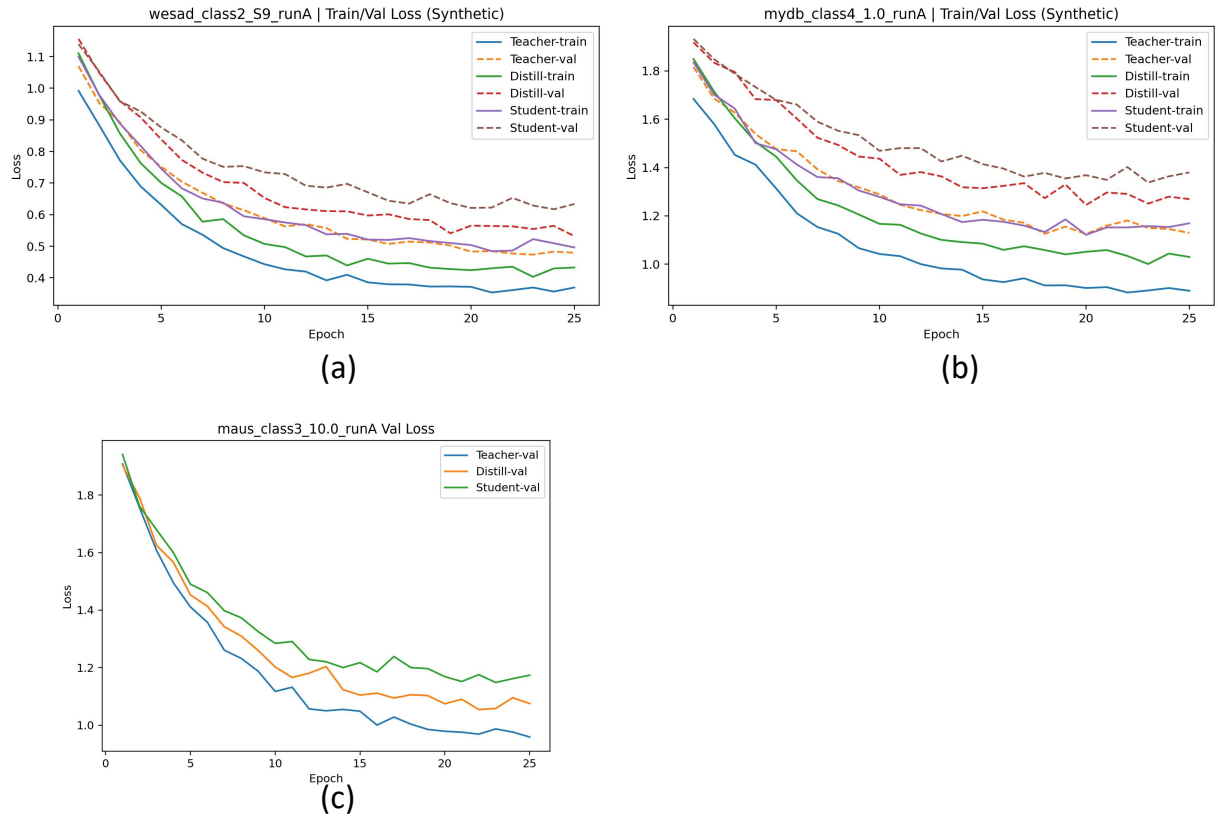


Figure 5.11: Training and validation loss curves for multi class classification tasks across different datasets. (a) Loss curves for the WESAD four class task, demonstrating smooth and stable optimization behavior for all evaluated models. (b) Loss curves for the MyDB four class task, where higher loss values and slower convergence reflect increased task difficulty. (c) Loss curves for the MAUS three-class task, exhibiting moderate convergence behavior between the WESAD and MyDB datasets. In all cases, the distilled student model shows consistently lower loss values compared with the baseline student model, indicating improved optimization stability during training.

differences in absolute performance levels. The WESAD dataset in Figure 5.10(a) achieves the highest validation accuracy, while the MyDB dataset in Figure 5.10(b) presents lower accuracy due to increased data complexity. The MAUS dataset in Figure 5.10X(c) demonstrates intermediate behavior. Across all datasets, the distilled student model consistently outperforms the baseline student model throughout training. Correspondingly, the loss curves in Figure 5.11 indicate stable and monotonic decreases without divergence, further supporting that the observed performance gains are achieved through consistent and reliable optimization rather than dataset specific effects.

Table 5.3: Statistical significance analysis of performance improvements achieved by knowledge distillation using paired t-tests.

Dataset	Task	Metric	n	Student	Distill	Mean Diff	CI _{95%} Low	CI _{95%} High	p-value
WESAD	2-class	Accuracy (%)	15	84.67	88.85	4.18	2.09	6.28	0.00075
WESAD	2-class	Macro-F1	15	0.823	0.869	0.046	0.027	0.065	0.00013
WESAD	4-class	Accuracy (%)	15	76.17	80.71	4.55	4.42	4.68	$< 10^{-18}$
WESAD	4-class	Macro-F1	15	0.733	0.782	0.049	0.045	0.053	$< 10^{-13}$
MAUS	2-class	Accuracy (%)	22	82.89	88.07	5.18	4.20	6.16	$< 10^{-9}$
MAUS	2-class	Macro-F1	22	0.828	0.864	0.036	0.027	0.044	$< 10^{-7}$
MAUS	3-class	Accuracy (%)	22	74.26	78.82	4.57	3.62	5.51	$< 10^{-8}$
MAUS	3-class	Macro-F1	22	0.720	0.755	0.035	0.027	0.043	$< 10^{-7}$

To further quantify the reliability of the observed performance improvements, paired t-tests were conducted across subjects for all evaluated datasets and task configurations. The variable n denotes the number of subjects included in each evaluation. The statistical results are summarized in Table 5.3. Across all tasks and metrics, the distilled student model consistently outperforms the baseline student model, as indicated by positive mean differences in both Accuracy and Macro F1. The corresponding confidence intervals do not include zero, confirming that the observed improvements are statistically robust. In addition, all reported p-values indicate statistically significant differences, supporting that the performance gains achieved through knowledge distillation are unlikely to result from random variation. These findings provide strong statistical evidence that complements the training dynamics and performance analyzes presented earlier in this section.

In summary, the training dynamics and statistical analyzes presented in this section demonstrate that the performance improvements achieved by knowledge distillation are supported by stable optimization behavior and consistent convergence patterns. Across different datasets and task configurations, the distilled student model exhibits smoother training trajectories, higher validation performance, and statistically significant gains compared with the baseline student model. These observations indicate that the reported improvements are not incidental outcomes of specific training runs but instead reflect reliable and reproducible learning behavior. In the next chapter, the implications of these findings are discussed in relation to practical deployment considerations and methodological limitations.

This chapter presents a comprehensive evaluation of the proposed framework in multi-

ple datasets and classification tasks. The experimental results demonstrate consistent performance improvements achieved through knowledge distillation, as evidenced by overall classification metrics, class wise error analysis, training dynamics, and statistical significance testing. Together, these results confirm that the observed gains are robust across subjects, task complexities, and data sources, and are supported by stable and reproducible training behavior. Although the Results chapter focuses on empirical observations and quantitative evidence, the following Discussion chapter interprets these findings in a broader context, examines their implications for practical deployment, and addresses the limitations and assumptions underlying the proposed approach.

DISCUSSION

This chapter discusses the experimental results presented in the previous chapter and provides an interpretive analysis of the observed performance trends. The discussion focuses on how model behavior varies across different datasets and classification tasks, with particular attention to the impact of task complexity and training strategy on classification performance. Rather than restating quantitative results, this chapter aims to explain the underlying factors that contribute to the observed differences in performance and stability. The discussion is further contextualized by comparing the findings with existing studies in ECG based stress recognition and by reflecting on their methodological and practical implications. Finally, the limitations of the current study are acknowledged to clarify the scope and applicability of the reported results.

Across all evaluated settings, binary classification tasks consistently exhibited stronger and more stable performance than multi class tasks [148]. This trend can be attributed to the reduced decision complexity associated with binary formulations, where the model is required to separate two broadly defined states rather than multiple overlapping categories. In emotional stress recognition, physiological responses often share similar patterns across related stress levels, which increases class ambiguity in multi class settings [149]. As a result, feature representations learned under binary objectives tend to be more discriminative and robust, while multi class learning is more sensitive to inter class similarity and label uncertainty. This observation aligns with the intrinsic characteristics of ECG based signals, where stress related variations are often gradual rather than sharply separated.

Performance differences across datasets further reflect variations in experimental de-

sign, data quality, and task structure. Public benchmark datasets benefit from standardized protocols and clearer labeling schemes, which support more consistent model learning [150]. In contrast, self collected data introduce greater variability due to individual differences, recording conditions, and subjective annotation, leading to less stable performance trends. Within this context, knowledge distillation provides an effective mechanism to enhance student model learning by transferring structured representations from a stronger teacher model [151]. Through this process, the student model benefits from additional supervisory information beyond hard labels, which improves generalization and mitigates performance degradation under constrained input conditions. The observed improvements indicate that informed training strategies can partially compensate for reduced model capacity and limited modality availability.

Existing studies on ECG based stress recognition have primarily focused on improving classification performance through increasingly complex model architectures or multimodal signal integration. While these approaches have demonstrated strong results in controlled settings, they often rely on rich input modalities or computationally intensive models, which limits their practicality for deployment in wearable and resource constrained environments. In contrast, the current study adopts a different perspective by prioritizing ECG only inference while enhancing representation learning through informed training strategies. By leveraging knowledge distillation during training, the proposed approach seeks to balance performance and deployability, positioning this work as a complementary contribution rather than a direct performance driven alternative to existing methods.

The findings of this study offer several practical implications for the design of ECG based stress recognition systems. First, they suggest that acceptable and stable performance can be achieved without relying on multimodal inputs during inference, which supports simpler and more wearable friendly system designs. Second, the results highlight the value of training strategies that emphasize representation quality rather than model scale, indicating that performance gains can be obtained through informed supervision even under constrained input conditions. Finally, the observed effectiveness of knowledge distillation indicates that lightweight models can benefit from guidance provided by stronger models during training, offering a viable pathway for balancing accuracy, efficiency, and deployability in real world applications.

Several limitations of the present study should be acknowledged. First, the performance gains introduced by knowledge distillation were not consistently observed across all settings, and in some cases the student model achieved comparable or superior performance without distillation. This outcome may be related to factors such as task simplicity, limited

data scale, or suboptimal teacher student alignment, where the additional supervisory information did not provide further benefit. Second, the evaluation protocol did not adopt a leave one subject out strategy, which limits the assessment of cross subject generalization and may overestimate performance stability. Third, model performance was found to be sensitive to signal quality, particularly in recordings affected by noise or motion artifacts, which remains a practical challenge for ECG based systems. In addition, the scale and diversity of the available data constrain the generalizability of the findings, and the subjective nature of stress labeling may introduce uncertainty into the learning process.

In summary, this chapter has interpreted the experimental findings by examining performance trends across tasks, datasets, and training strategies. The discussion has situated the results within the context of existing ECG based stress recognition studies and highlighted their methodological and practical implications. At the same time, key limitations have been acknowledged to clarify the scope of the current work. Together, these discussions provide a coherent understanding of the reported results and establish a foundation for the conclusions presented in the following chapter.

CONCLUSION

This study has investigated ECG based emotional stress recognition under different datasets and classification settings, with an emphasis on performance robustness and deployment oriented design. By evaluating multiple task formulations and training strategies, the results demonstrate consistent performance trends related to task complexity, data characteristics, and model capacity. The findings indicate that effective stress recognition can be achieved under constrained inference conditions when appropriate training strategies are adopted. In general, the study provides a comprehensive assessment of ECG centered stress classification and establishes a clear basis for the contributions and conclusions discussed in this chapter.

This study contributes to the field of ECG based stress recognition from both methodological and experimental perspectives. Methodologically, it demonstrates that ECG only inference can be effectively supported through informed training strategies, offering a practical alternative to fully multimodal deployment. Experimentally, the study provides a systematic evaluation across multiple datasets and task settings, highlighting consistent trends related to task complexity and data characteristics. In addition, the analysis of knowledge distillation within this context offers empirical insight into its benefits and limitations when applied to lightweight models. Together, these contributions help clarify design trade offs between performance, model complexity, and practical applicability in ECG centered stress recognition systems.

Several directions for future research can be identified based on the findings and limitations of this study. Further work could adopt more rigorous cross subject evaluation pro-

protocols to better assess generalization performance in realistic settings. Improvements in data quality and labeling consistency may also help reduce performance variability, particularly under noisy or unconstrained recording conditions. In addition, future studies may explore more adaptive or task aware training strategies to better align teacher and student representations when knowledge distillation is applied. These directions may support the development of more robust and generalizable ECG based stress recognition systems.

In conclusion, this thesis presents a deployment oriented investigation of ECG based emotional stress recognition that balances methodological rigor with practical considerations. By examining performance trends across datasets, tasks, and training strategies, the study clarifies both the potential and the limitations of ECG centered approaches under constrained inference conditions. It is hoped that the insights provided in this work will support future research and system design in wearable stress recognition and related applications.

BIBLIOGRAPHY

- [1] Katrin Starcke and Matthias Brand. "Decision making under stress: a selective review." In: *Neuroscience & Biobehavioral Reviews* 36.4 (2012), pp. 1228–1248.
- [2] Bruce S McEwen. "Protection and damage from acute and chronic stress: allostasis and allostatic overload and relevance to the pathophysiology of psychiatric disorders." In: *Annals of the new York Academy of Sciences* 1032.1 (2004), pp. 1–7.
- [3] Tashfia Ahmed, Meha Qassem, and Panicos A Kyriacou. "Physiological monitoring of stress and major depression: A review of the current monitoring techniques and considerations for the future." In: *Biomedical Signal Processing and Control* 75 (2022), p. 103591.
- [4] Marvin R Golfried and Ronald N Kent. "Traditional versus behavioral personality assessment: a comparison of methodological and theoretical assumptions." In: *Psychological Bulletin* 77.6 (1972), p. 409.
- [5] Giorgos Giannakakis et al. "Review on psychological stress detection using biosignals." In: *IEEE transactions on affective computing* 13.1 (2019), pp. 440–460.
- [6] Sylvia D Kreibig. "Autonomic nervous system activity in emotion: A review." In: *Biological psychology* 84.3 (2010), pp. 394–421.
- [7] Giorgos Giannakakis et al. "Review on Psychological Stress Detection Using Biosignals." In: *IEEE Transactions on Affective Computing* 13.1 (2022), pp. 440–460. DOI: 10.1109/TAFFC.2019.2927337.
- [8] Maria Egger, Matthias Ley, and Sten Hanke. "Emotion Recognition from Physiological Signal Analysis: A Review." In: *Electronic Notes in Theoretical Computer Science* 343 (2019), pp. 35–55. ISSN: 1571-0661. DOI: <https://doi.org/10.1016/j.entcs.2019.04.009>. URL: <https://www.sciencedirect.com/science/article/pii/S157106611930009X>.

- [9] Sanjeev Kumar Saini and Rashmi Gupta. "A Review on ECG Signal Analysis for Mental Stress Assessment." In: *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)*. 2019, pp. 915–918.
- [10] PC De Looft et al. "Associations of sympathetic and parasympathetic activity in job stress and burnout: A systematic review." In: *PLoS One* 13.10 (2018), e0205741.
- [11] Hye-Geum Kim et al. "Stress and heart rate variability: a meta-analysis and review of the literature." In: *Psychiatry investigation* 15.3 (2018), p. 235.
- [12] Bob JA Schijvenaars, Gerard van Herpen, and Jan A Kors. "Intraindividual variability in electrocardiograms." In: *Journal of Electrocardiology* 41.3 (2008), pp. 190–196.
- [13] Muhammad Zeeshan Baig and Manolya Kavakli. "A Survey on Psycho-Physiological Analysis and Measurement Methods in Multimodal Systems." In: *Multimodal Technologies and Interaction* 3.2 (2019). ISSN: 2414-4088. DOI: 10.3390/mti3020037. URL: <https://www.mdpi.com/2414-4088/3/2/37>.
- [14] Manju Priya Arthanarisamy Ramaswamy and Suja Palaniswamy. "Multimodal emotion recognition: A comprehensive review, trends, and challenges." In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 14.6 (2024), e1563.
- [15] Priya Karthikeyan, Murugappan Murugappan, and Sazali Yaacob. "Detection of human stress using short-term ECG and HRV signals." In: *Journal of Mechanics in Medicine and Biology* 13.02 (2013), p. 1350038.
- [16] Fang Li and Dan Zhang. "Multimodal physiological signals from wearable sensors for affective computing: A systematic review." In: *Intelligent Sports and Health* 1.4 (2025), pp. 210–222.
- [17] Maria Kyrou, Ioannis Kompatsiaris, and Panagiotis C Petrantonakis. "Deep learning approaches for stress detection: A survey." In: *IEEE Transactions on Affective Computing* 16.2 (2024), pp. 499–517.
- [18] P. C. de Looft et al. "Associations of sympathetic and parasympathetic activity in job stress and burnout: A systematic review." In: *PLOS ONE* 13.10 (Oct. 2018),

- pp. 1–24. DOI: 10.1371/journal.pone.0205741. URL: <https://doi.org/10.1371/journal.pone.0205741>.
- [19] Ilan S. Wittstein et al. “Neurohumoral Features of Myocardial Stunning Due to Sudden Emotional Stress.” In: *New England Journal of Medicine* 352.6 (2005), pp. 539–548. DOI: 10.1056/NEJMoa043046. URL: <https://www.nejm.org/doi/full/10.1056/NEJMoa043046>.
- [20] AS Anusha et al. “Physiological signal based work stress detection using unobtrusive sensors.” In: *Biomedical Physics & Engineering Express* 4.6 (2018), p. 065001.
- [21] Sara Pourmohammadi and Ali Maleki. “Stress detection using ECG and EMG signals: A comprehensive study.” In: *Computer methods and programs in biomedicine* 193 (2020), p. 105482.
- [22] Jun-Zhi Xiang et al. “A multi-modal deep learning approach for stress detection using physiological signals: integrating time and frequency domain features.” In: *Frontiers in Physiology* 16 (2025), p. 1584299.
- [23] Dingwen Li et al. “Feasibility study of monitoring deterioration of outpatients using multimodal data collected by wearables.” In: *ACM Transactions on Computing for Healthcare* 1.1 (2020), pp. 1–22.
- [24] Yu Liu et al. “Machine Learning, Physiological Signals, and Emotional Stress/Anxiety: Pitfalls and Challenges.” In: *Applied Sciences* 15.21 (2025), p. 11777.
- [25] Dimitrios Andrikopoulos et al. “Machine learning-enabled detection of attention-deficit/hyperactivity disorder with multimodal physiological data: a case-control study.” In: *BMC psychiatry* 24.1 (2024), p. 547.
- [26] Oliver J Chrzanowski-Smith et al. “Variability in exercise physiology: can capturing intra-individual variation help better understand true inter-individual responses?” In: *European Journal of Sport Science* 20.4 (2020), pp. 452–460.
- [27] Dimitris Giakoumis, Dimitrios Tzovaras, and George Hassapis. “Subject-dependent biosignal features for increased accuracy in psychological stress detection.” In: *International Journal of Human-Computer Studies* 71.4 (2013), pp. 425–439.

- [28] Shruti Gedam and Sanchita Paul. "A review on mental stress detection using wearable sensors and machine learning techniques." In: *IEEE Access* 9 (2021), pp. 84045–84066.
- [29] Nilava Mukherjee, Sumitra Mukhopadhyay, and Rajarshi Gupta. "Real-time mental stress detection technique using neural networks towards a wearable health monitor." In: *Measurement Science and Technology* 33.4 (2022), p. 044003. DOI: 10.1088/1361-6501/ac3aae. URL: <https://doi.org/10.1088/1361-6501/ac3aae>.
- [30] Feng-Tso Sun et al. "Activity-Aware Mental Stress Detection Using Physiological Sensors." In: *Mobile Computing, Applications, and Services*. Ed. by Martin Gris and Guang Yang. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 282–301. ISBN: 978-3-642-29336-8.
- [31] Eman Abdelfattah, Shreehar Joshi, and Shreekar Tiwari. "Machine and Deep Learning Models for Stress Detection Using Multimodal Physiological Data." In: *IEEE Access* 13 (2025), pp. 4597–4608. DOI: 10.1109/ACCESS.2024.3525459.
- [32] Chariklia Chatzaki and Manolis Tsiknakis. "An Overview of Stress Analysis Based on Physiological Signals: Systematic Review of Open Datasets and Current Trends." In: *Sensors* 25.23 (2025). ISSN: 1424-8220. DOI: 10.3390/s25237108. URL: <https://www.mdpi.com/1424-8220/25/23/7108>.
- [33] Louise V. Coutts et al. "Deep learning with wearable based heart rate variability for prediction of mental and general health." In: *Journal of Biomedical Informatics* 112 (2020), p. 103610. ISSN: 1532-0464. DOI: <https://doi.org/10.1016/j.jbi.2020.103610>. URL: <https://www.sciencedirect.com/science/article/pii/S1532046420302380>.
- [34] Juan Antonio Castro-García et al. "Towards Human Stress and Activity Recognition: A Review and a First Approach Based on Low-Cost Wearables." In: *Electronics* 11.1 (2022). ISSN: 2079-9292. DOI: 10.3390/electronics11010155. URL: <https://www.mdpi.com/2079-9292/11/1/155>.
- [35] Deepak Berwal et al. "Motion Artifact Removal in Ambulatory ECG Signal for Heart Rate Variability Analysis." In: *IEEE Sensors Journal* 19.24 (2019), pp. 12432–12442. DOI: 10.1109/JSEN.2019.2939391.

- [36] Julian F. Thayer et al. "A meta-analysis of heart rate variability and neuroimaging studies: Implications for heart rate variability as a marker of stress and health." In: *Neuroscience & Biobehavioral Reviews* 36.2 (2012), pp. 747–756. ISSN: 0149-7634. DOI: <https://doi.org/10.1016/j.neubiorev.2011.11.009>. URL: <https://www.sciencedirect.com/science/article/pii/S0149763411002077>.
- [37] Jing Zhang et al. "Real-time mental stress detection using multimodality expressions with a deep learning framework." In: *Frontiers in Neuroscience* Volume 16 - 2022 (2022). ISSN: 1662-453X. DOI: 10.3389/fnins.2022.947168. URL: <https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2022.947168>.
- [38] Aya Khalaf et al. "Analysis of multimodal physiological signals within and between individuals to predict psychological challenge vs. threat." In: *Expert Systems with Applications* 140 (2020), p. 112890. ISSN: 0957-4174. DOI: <https://doi.org/10.1016/j.eswa.2019.112890>. URL: <https://www.sciencedirect.com/science/article/pii/S0957417419306001>.
- [39] Philip Schmidt et al. "Wearable-Based Affect Recognition—A Review." In: *Sensors* 19.19 (2019). ISSN: 1424-8220. DOI: 10.3390/s19194079. URL: <https://www.mdpi.com/1424-8220/19/19/4079>.
- [40] Hadi Banaee, Mobyen Uddin Ahmed, and Amy Loutfi. "Data Mining for Wearable Sensors in Health Monitoring Systems: A Review of Recent Trends and Challenges." In: *Sensors* 13.12 (2013), pp. 17472–17500. ISSN: 1424-8220. DOI: 10.3390/s131217472. URL: <https://www.mdpi.com/1424-8220/13/12/17472>.
- [41] Savannah L. Buchner et al. "Assessing Physiological Signal Utility and Sensor Burden in Estimating Trust, Situation Awareness, and Mental Workload." In: *Journal of Cognitive Engineering and Decision Making* 19.2 (2025), pp. 154–173. DOI: 10.1177/15553434241310084. URL: <https://doi.org/10.1177/15553434241310084>.
- [42] Guobin Chen et al. "Learning Efficient Object Detection Models with Knowledge Distillation." In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/e1e32e235eee1f970470a3a6658dfdd5-Paper.pdf.

- [43] A. Sathiya et al. "IoT Enabled Healthcare Framework Using Edge AI and Advanced Wearable Sensors for Real Time Health Monitoring." In: *2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)*. 2025, pp. 384–392. DOI: 10.1109/ICMSCI62561.2025.10894492.
- [44] Xing Wang et al. "Deep Convolutional Networks With Tunable Speed–Accuracy Tradeoff for Human Activity Recognition Using Wearables." In: *IEEE Transactions on Instrumentation and Measurement* 71 (2022), pp. 1–12. DOI: 10.1109/TIM.2021.3132088.
- [45] Jacqueline Wijsman et al. "Towards mental stress detection using wearable physiological sensors." In: *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2011, pp. 1798–1801. DOI: 10.1109/IEMBS.2011.6090512.
- [46] Khalid Masood et al. "Consistency and Validity of Self-reporting Scores in Stress Measurement Surveys." In: *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2012, pp. 4895–4898. DOI: 10.1109/EMBC.2012.6347091.
- [47] Jiayuan He et al. "Real-Time Detection of Acute Cognitive Stress Using a Convolutional Neural Network From Electrocardiographic Signal." In: *IEEE Access* 7 (2019), pp. 42710–42717. DOI: 10.1109/ACCESS.2019.2907076.
- [48] Amandeep Cheema et al. "Combined empirical mode decomposition and phase space reconstruction based psychologically stressed and non-stressed state classification from cardiac sound signals." In: *Biomedical Signal Processing and Control* 82 (2023), p. 104585. ISSN: 1746-8094. DOI: <https://doi.org/10.1016/j.bspc.2023.104585>. URL: <https://www.sciencedirect.com/science/article/pii/S1746809423000186>.
- [49] L. Vanitha and G. R. Suresh. "Hybrid SVM classification technique to detect mental stress in human beings using ECG signals." In: *2013 International Conference on Advanced Computing and Communication Systems*. 2013, pp. 1–6. DOI: 10.1109/ICACCS.2013.6938735.
- [50] George Tanev et al. "Classification of acute stress using linear and non-linear heart rate variability analysis derived from sternal ECG." In: *2014 36th Annual*

- International Conference of the IEEE Engineering in Medicine and Biology Society*. 2014, pp. 3386–3389. DOI: 10.1109/EMBC.2014.6944349.
- [51] Maria Kyrou, Ioannis Kompatsiaris, and Panagiotis C. Petrantonakis. “Deep Learning Approaches for Stress Detection: A Survey.” In: *IEEE Transactions on Affective Computing* 16.2 (2025), pp. 499–517. DOI: 10.1109/TAFFC.2024.3455371.
- [52] Maciej Dzieżyc et al. “Can We Ditch Feature Engineering? End-to-End Deep Learning for Affect Recognition from Physiological Sensor Data.” In: *Sensors* 20.22 (2020). ISSN: 1424-8220. DOI: 10.3390/s20226535. URL: <https://www.mdpi.com/1424-8220/20/22/6535>.
- [53] Xiaowei Zhang et al. “Dynamic Alignment and Fusion of Multimodal Physiological Patterns for Stress Recognition.” In: *IEEE Transactions on Affective Computing* 15.2 (2024), pp. 685–696. DOI: 10.1109/TAFFC.2023.3290177.
- [54] Kuldeep Mahato et al. “Hybrid multimodal wearable sensors for comprehensive health monitoring.” In: *Nature Electronics* 7.9 (2024), pp. 735–750.
- [55] Sara Aristizabal et al. “The Feasibility of Wearable and Self-Report Stress Detection Measures in a Semi-Controlled Lab Environment.” In: *IEEE Access* 9 (2021), pp. 102053–102068. DOI: 10.1109/ACCESS.2021.3097038.
- [56] Jajack Heikenfeld et al. “Wearable sensors: modalities, challenges, and prospects.” In: *Lab on a Chip* 18.2 (2018), pp. 217–248.
- [57] S.Sheeba Rani et al. “Advanced IoT-enabled embedded wearable system for continuous cardiovascular health monitoring using edge intelligence.” In: *Biomedical Signal Processing and Control* 111 (2026), p. 108362. ISSN: 1746-8094. DOI: <https://doi.org/10.1016/j.bspc.2025.108362>. URL: <https://www.sciencedirect.com/science/article/pii/S1746809425008730>.
- [58] Giorgos Giannakakis et al. “Review on psychological stress detection using biosignals.” In: *IEEE Transactions on Affective Computing* 13.1 (2022), pp. 440–460. DOI: 10.1109/TAFFC.2019.2927337.
- [59] Mohamed Hammad et al. “Detection of abnormal heart conditions based on characteristics of ECG signals.” In: *Measurement* 125 (2018), pp. 634–644.

- [60] George E. Billman. "The LF/HF ratio does not accurately measure cardiac sympatho-vagal balance." In: *Frontiers in Physiology* 4 (2013), p. 26. ISSN: 1664-042X. DOI: 10.3389/fphys.2013.00026. URL: <https://www.frontiersin.org/articles/10.3389/fphys.2013.00026>.
- [61] R. Castaldo et al. "Acute mental stress assessment via short-term HRV analysis in healthy adults: A systematic review with meta-analysis." In: *Biomedical Signal Processing and Control* 18 (2015), pp. 370–377. ISSN: 1746-8094. DOI: 10.1016/j.bspc.2015.02.012. URL: <https://www.sciencedirect.com/science/article/pii/S1746809415000245>.
- [62] Mariska Kret et al. "Perception of Face and Body Expressions Using Electromyography, Pupillometry and Gaze Measures." In: *Frontiers in Psychology* Volume 4 - 2013 (2013). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2013.00028. URL: <https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2013.00028>.
- [63] TDRR Moritani, D Stegeman, and R Merletti. "Basic physiology and biophysics of EMG signal generation." In: *Electromyography: physiology, engineering, and noninvasive applications* (2004), pp. 1–25.
- [64] Michael Pluess, Ansgar Conrad, and Frank H. Wilhelm. "Muscle tension in generalized anxiety disorder: A critical review of the literature." In: *Journal of Anxiety Disorders* 23.1 (2009), pp. 1–11. ISSN: 0887-6185. DOI: <https://doi.org/10.1016/j.janxdis.2008.03.016>. URL: <https://www.sciencedirect.com/science/article/pii/S088761850800090X>.
- [65] Mohammad Ahmed et al. "A Comprehensive Analysis of Trapezius Muscle EMG Activity in Relation to Stress and Meditation." In: *BioMedInformatics* 4.2 (2024), pp. 1047–1058.
- [66] Jason J Braithwaite et al. "A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments." In: *Psychophysiology* 49.1 (2013), pp. 1017–1034.
- [67] Osmalina Nur Rahma et al. "Electrodermal activity for measuring cognitive and emotional stress level." In: *Journal of Medical Signals & Sensors* 12.2 (2022), pp. 155–162.

- [68] Dindar S Bari. "Gender differences in tonic and phasic electrodermal activity components." In: *Science Journal of University of Zakho* 8.1 (2020), pp. 29–33.
- [69] Matin Khalili et al. "Motion artifacts in capacitive ECG monitoring systems: a review of existing models and reduction techniques." In: *Medical & Biological Engineering & Computing* 62.12 (2024), pp. 3599–3622.
- [70] Marianne Boyer et al. "Reducing Noise, Artifacts and Interference in Single-Channel EMG Signals: A Review." In: *Sensors* 23.6 (2023). ISSN: 1424-8220. DOI: 10.3390/s23062927. URL: <https://www.mdpi.com/1424-8220/23/6/2927>.
- [71] Keith B. Burt and Jelena Obradović. "The construct of psychophysiological reactivity: Statistical and psychometric issues." In: *Developmental Review* 33.1 (2013), pp. 29–57. ISSN: 0273-2297. DOI: <https://doi.org/10.1016/j.dr.2012.10.002>. URL: <https://www.sciencedirect.com/science/article/pii/S0273229712000524>.
- [72] Giorgos Giannakakis et al. "Review on Psychological Stress Detection Using Biosignals." In: *IEEE Transactions on Affective Computing* 13.1 (2022), pp. 440–460. DOI: 10.1109/TAFFC.2019.2927337.
- [73] Georgios N. Yannakakis, Roddy Cowie, and Carlos Busso. "The Ordinal Nature of Emotions: An Emerging Approach." In: *IEEE Transactions on Affective Computing* 12.1 (2021), pp. 16–35. DOI: 10.1109/TAFFC.2018.2879512.
- [74] Kurt Plarre et al. "Continuous inference of psychological stress from sensory measurements collected in the natural environment." In: *Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks*. 2011, pp. 97–108.
- [75] Jing Zhai and Armando Barreto. "Stress Detection in Computer Users Based on Digital Signal Processing of Noninvasive Physiological Variables." In: *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*. 2006, pp. 1355–1358. DOI: 10.1109/IEMBS.2006.259421.
- [76] Beatrice De Marchi, Endi Agovi, and Andrea Aliverti. "Stress level assessment by a multi-parametric wearable platform: Relevance of different physiological signals." In: *Information Systems Frontiers* 27.1 (2025), pp. 113–137.

- [77] Pramod Bobade and M. Vani. "Stress Detection with Machine Learning and Deep Learning using Multimodal Physiological Data." In: *2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA)*. 2020, pp. 51–57. DOI: 10.1109/ICIRCA48905.2020.9183244.
- [78] James A Meurs et al. "Physiological reactions to and recovery from acute stressors: the roles of chronic anxiety and stable resources." In: *Health psychology report* 11.3 (2022), p. 223.
- [79] Alexandra D Crosswell and Kimberly G Lockwood. "Best practices for stress measurement: How to measure psychological stress in health research." In: *Health psychology open* 7.2 (2020), p. 2055102920933072.
- [80] Han Yu and Akane Sano. "Semi-Supervised Learning for Wearable-based Momentary Stress Detection in the Wild." In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7.2 (2023). DOI: 10.1145/3596246. URL: <https://doi.org/10.1145/3596246>.
- [81] Gal Richter-Levin and Carmen Sandi. "Labels Matter: Is it stress or is it Trauma?" In: *Translational Psychiatry* 11.1 (2021), p. 385.
- [82] Reda Yacouby and Dustin Axman. "Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models." In: *Proceedings of the first workshop on evaluation and comparison of NLP systems*. 2020, pp. 79–91.
- [83] Aryan Rastogi, Qian Liu, and Erik Cambria. "Stress Detection from Social Media Articles: New Dataset Benchmark and Analytical Study." In: *2022 International Joint Conference on Neural Networks (IJCNN)*. 2022, pp. 1–8. DOI: 10.1109/IJCNN55064.2022.9892889.
- [84] Ruishi Zhou et al. "ECG-based biometric under different psychological stress states." In: *Computer Methods and Programs in Biomedicine* 202 (2021), p. 106005. ISSN: 0169-2607. DOI: <https://doi.org/10.1016/j.cmpb.2021.106005>. URL: <https://www.sciencedirect.com/science/article/pii/S0169260721000808>.

- [85] Ramyashri B Ramteke, Gaurav O Gajbhiye, and Vijaya R Thool. “Discriminating psychological stress levels: multi-level attentive LSTM approach.” In: *Neural Computing and Applications* 37.30 (2025), pp. 25579–25599.
- [86] Shiqi Yang et al. “A deep learning approach to stress recognition through multimodal physiological signal image transformation.” In: *Scientific Reports* 15.1 (2025), p. 22258.
- [87] Zankhana Bhatt and Ashwin Dobariya. “Stress Level Classification Using Multimodal Deep Learning and Physiological Signal Analysis.” In: *SN Computer Science* 6.5 (2025), p. 499.
- [88] Maxine He et al. “Objective anxiety level classification using unsupervised learning and multimodal physiological signals.” In: *Smart Health* (2025), p. 100572.
- [89] Zhifeng Wang et al. “PhysioFormer: Integrating multimodal physiological signals and symbolic regression for explainable affective state prediction.” In: *PLoS One* 20.10 (2025), e0335221.
- [90] Hongxiang Gao et al. “Multimodal Fusion of Behavioral and Physiological Signals for Enhanced Emotion Recognition Via Feature Decoupling and Knowledge Transfer.” In: *IEEE Journal of Biomedical and Health Informatics* (2025), pp. 1–11. DOI: 10.1109/JBHI.2025.3597398.
- [91] SS Aravinth et al. “Dynamic cross-domain transfer learning for driver fatigue monitoring: multi-modal sensor fusion with adaptive real-time personalizations.” In: *Scientific Reports* 15.1 (2025), p. 15840.
- [92] Stevica Cvetkovic, Sandra Stankovic, and Sasa V. Nikolic. “BioEdgeNet: A compact deep residual network for stress recognition on edge devices.” In: *Biomedical Signal Processing and Control* 102 (2025), p. 107361. ISSN: 1746-8094. DOI: <https://doi.org/10.1016/j.bspc.2024.107361>. URL: <https://www.sciencedirect.com/science/article/pii/S1746809424014198>.
- [93] Wenhao Li et al. “BrainFusion: a Low-Code, Reproducible, and Deployable Software Framework for Multimodal Brain–Computer Interface and Brain–Body Interaction Research.” In: *Advanced Science* 12.32 (2025), e17408. DOI:

- <https://doi.org/10.1002/adv.202417408>. URL: <https://advanced.onlinelibrary.wiley.com/doi/abs/10.1002/adv.202417408>.
- [94] Evgenia Lazarou and Themis P Exarchos. “Predicting stress levels using physiological data: Real-time stress prediction models utilizing wearable devices.” In: *AIMS neuroscience* 11.2 (2024), p. 76.
- [95] Karisma Trinanda Putra et al. “A Review on the Application of Internet of Medical Things in Wearable Personal Health Monitoring: A Cloud-Edge Artificial Intelligence Approach.” In: *IEEE Access* 12 (2024), pp. 21437–21452. DOI: 10.1109/ACCESS.2024.3358827.
- [96] Ekta Singh Dahiya et al. “Wearable technology for monitoring electrocardiograms (ECGs) in adults: a scoping review.” In: *Sensors* 24.4 (2024), p. 1318.
- [97] Aboul Hassane Cisse and Shoya Ishimaru. “From Signals to Cognition: Self-Supervised Multimodal Learning for Cognitive State Inference from Wearable Sensor Streams.” In: *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. UbiComp Companion '25. Finland: Association for Computing Machinery, 2026, pp. 105–109. ISBN: 9798400714771. DOI: 10.1145/3714394.3754403. URL: <https://doi.org/10.1145/3714394.3754403>.
- [98] Mohamed Naeem et al. “Wearable ECG systems for accurate mental stress detection: a scoping review.” In: *Journal of Public Health* 33.6 (2025), pp. 1181–1197.
- [99] Meena Anandan and Rohini Palanisamy. “Wearable ECG Sensor-Based Cognitive Stress Detection During Physical Activity Using Multifractal Detrended Fluctuation Analysis.” In: *IEEE Sensors Letters* 9.11 (2025), pp. 1–4. DOI: 10.1109/LSENS.2025.3605112.
- [100] Yixuan Zhang et al. “An Adaptive Knowledge Distillation Framework for Short-Form Remote Physiological Measurement.” In: *IEEE Transactions on Instrumentation and Measurement* 74 (2025), pp. 1–15. DOI: 10.1109/TIM.2025.3580874.

- [101] Julien Rimok and Kelly A Larkin-Kaiser. “From lab to life: Wearables, real-world data, and the future of autonomic research.” In: *Autonomic Neuroscience* (2025), p. 103377.
- [102] U. Pluntke et al. “Evaluation and Classification of Physical and Psychological Stress in Firefighters using Heart Rate Variability.” In: *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2019, pp. 2207–2212. DOI: 10.1109/EMBC.2019.8856596.
- [103] Wenqi Zheng and Yutaka Arakawa. “A Review of Wearables-Based Activities of Daily Living Recognition With LLMs: Overview, Progress, and Trends.” In: *IEEE Access* 13 (2025), pp. 143813–143830. DOI: 10.1109/ACCESS.2025.3598088.
- [104] Kang Yin et al. “EEG-Based Multimodal Representation Learning for Emotion Recognition.” In: *2025 13th International Conference on Brain-Computer Interface (BCI)*. 2025, pp. 1–4. DOI: 10.1109/BCI65088.2025.10931743.
- [105] Jiyao Wang et al. “PhysMLE: Generalizable and Priors-Inclusive Multi-Task Remote Physiological Measurement.” In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47.6 (2025), pp. 4908–4925. DOI: 10.1109/TPAMI.2025.3545598.
- [106] Akhilesh Kumar and Awadhesh Kumar. “Human emotion recognition using machine learning techniques based on the physiological signal.” In: *Biomedical Signal Processing and Control* 100 (2025), p. 107039.
- [107] Shikang Liu, Xiren Zhou, and Huanhuan Chen. “Multiscale Temporal Dynamic Learning for Time Series Classification.” In: *IEEE Transactions on Knowledge and Data Engineering* 37.6 (2025), pp. 3543–3555. DOI: 10.1109/TKDE.2025.3542799.
- [108] Durgesh Nandini et al. “An ensemble deep learning framework for emotion recognition through wearable devices multi-modal physiological signals.” In: *Scientific Reports* 15.1 (2025), p. 17263.
- [109] Tian Xie and Ning Ma. “Tracking vigilance fluctuations in real-time: a sliding-window heart rate variability-based machine-learning approach.” In: *Sleep* 48.2 (2025), zsae199.

- [110] Frigyes Samuel Racz and Gabor Csukly. “Information Leakage and Performance Overestimation in EEG-Based Schizophrenia Detection: Evidence from Literature and Empirical Analyses.” In: *medRxiv* (2025), pp. 2025–12.
- [111] Megha Khandelwal and Arun Sharma. “Machine learning and deep learning techniques to detect mental stress using various physiological signals: A critical insight.” In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 15.3 (2025), e70035.
- [112] Serkan Kiranyaz et al. “1-D Convolutional Neural Networks for Signal Processing Applications.” In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2019, pp. 8360–8364. DOI: 10.1109/ICASSP.2019.8682194.
- [113] Ting-Yun Hsiao et al. “Filter-based deep-compression with global average pooling for convolutional networks.” In: *Journal of Systems Architecture* 95 (2019), pp. 9–18.
- [114] Yujin Wu, Mohamed Daoudi, and Ali Amad. “Transformer-Based Self-Supervised Multimodal Representation Learning for Wearable Emotion Recognition.” In: *IEEE Transactions on Affective Computing* 15.1 (2024), pp. 157–172. DOI: 10.1109/TAFFC.2023.3263907.
- [115] Kan Hong. “Classification of emotional stress and physical stress using a multispectral based deep feature extraction model.” In: *Scientific Reports* 13.1 (2023), p. 2693.
- [116] Han Yu and Akane Sano. “Semi-supervised learning for wearable-based momentary stress detection in the wild.” In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7.2 (2023), pp. 1–23.
- [117] Pengguang Chen et al. “Distilling knowledge via knowledge review.” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 5008–5017.
- [118] Shibani Santurkar et al. “How does batch normalization help optimization?” In: *Advances in neural information processing systems* 31 (2018).

- [119] Hui Tang et al. “Blood pressure Estimation based on PPG and ECG signals using knowledge distillation.” In: *Cardiovascular Engineering and Technology* 15.1 (2024), pp. 39–51.
- [120] Mohammad Mahdi Azizi and Bagher BabaAli. “Biosignals Based Automated Driver Cognitive Load Assessment Using a Pre-Trained Transformer.” In: *IEEE Transactions on Intelligent Vehicles* 10.11 (2025), pp. 4829–4840. DOI: 10.1109/TIV.2024.3501016.
- [121] Shangquan Sun et al. “Logit standardization in knowledge distillation.” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2024, pp. 15731–15740.
- [122] Tao Huang et al. “Knowledge distillation from a stronger teacher.” In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 33716–33727.
- [123] Kyungyul Kim et al. “Self-knowledge distillation with progressive refinement of targets.” In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 6567–6576.
- [124] Ramatoulaye Diallo, Codjo Edalo, and O Olawale Awe. “Machine learning evaluation of imbalanced health data: a comparative analysis of balanced accuracy, MCC, and F1 score.” In: *Practical Statistical Learning and Data Science Methods: Case Studies from LISA 2020 Global Network, USA*. Springer, 2024, pp. 283–312.
- [125] Sorelle A Friedler et al. “A comparative study of fairness-enhancing interventions in machine learning.” In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 329–338.
- [126] Yucheng Liu, Ziyu Jia, and Haichao Wang. “EmotionKD: A Cross-Modal Knowledge Distillation Framework for Emotion Recognition Based on Physiological Signals.” In: *Proceedings of the 31st ACM International Conference on Multimedia*. Ottawa ON, Canada: Association for Computing Machinery, 2023, pp. 6122–6131. ISBN: 9798400701085. DOI: 10.1145/3581783.3612277. URL: <https://doi.org/10.1145/3581783.3612277>.

- [127] Jun Kong et al. “Multimodality Self-distillation for Fast Inference of Vision and Language Pretrained Models.” In: *IEEE Transactions on Multimedia* 26 (2024), pp. 8928–8940. DOI: 10.1109/TMM.2024.3384060.
- [128] Qihan Hu et al. “Unleashing the Power of Pretrained Transformer for Dense Prediction in Physiological Signals.” In: *IEEE Journal of Biomedical and Health Informatics* (2025), pp. 1–11. DOI: 10.1109/JBHI.2025.3592687.
- [129] Happy Nkanta Monday et al. “Enhancing ECG Classification in Cardiac Diagnostics: A Novel Approach Using Adaptive Focal Cross-Entropy Loss Function.” In: *IEEE Journal of Biomedical and Health Informatics* 29.10 (2025), pp. 7161–7174. DOI: 10.1109/JBHI.2025.3566531.
- [130] Firuz Kamalov, Fadi Thabtah, and Ho Hon Leung. “Feature selection in imbalanced data.” In: *Annals of Data Science* 10.6 (2023), pp. 1527–1541.
- [131] Pritam Sarkar et al. “Detection of maternal and fetal stress from the electrocardiogram with self-supervised representation learning.” In: *Scientific reports* 11.1 (2021), p. 24146.
- [132] Han Yu and Akane Sano. “Semi-Supervised Learning for Wearable-based Momentary Stress Detection in the Wild.” In: *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7.2 (June 2023). DOI: 10.1145/3596246. URL: <https://doi.org/10.1145/3596246>.
- [133] S Jerritta et al. “Physiological signals based human emotion Recognition: a review.” In: *2011 IEEE 7th International Colloquium on Signal Processing and its Applications*. 2011, pp. 410–415. DOI: 10.1109/CSPA.2011.5759912.
- [134] Jim Davis et al. “Hierarchical Classification with Confidence using Generalized Logits.” In: *2020 25th International Conference on Pattern Recognition (ICPR)*. 2021, pp. 1874–1881. DOI: 10.1109/ICPR48806.2021.9412867.
- [135] Ni Wayan Surya Wardhani et al. “Cross-validation Metrics for Evaluating Classification Performance on Imbalanced Data.” In: *2019 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*. 2019, pp. 14–18. DOI: 10.1109/IC3INA48034.2019.8949568.

- [136] G. Rajender Naik and K. Ashoka Reddy. “Comparative Analysis of ECG Classification Using Neuro-Fuzzy Algorithm and Multimodal Decision Learning Algorithm: ECG Classification Algorithm.” In: *2016 3rd International Conference on Soft Computing & Machine Intelligence (ISCMI)*. 2016, pp. 138–142. DOI: 10.1109/ISCMI.2016.35.
- [137] Sanghoon Choi et al. “Performance Improvement of Deep Learning Based Multi-Class ECG Classification Model Using Limited Medical Dataset.” In: *IEEE Access* 11 (2023), pp. 53185–53194. DOI: 10.1109/ACCESS.2023.3280565.
- [138] Vanessa Borst, Robert Leppich, and Samuel Kounev. “MACRO: A Multi-Head Attentional Convolutional Recurrent Network for the Classification of Co-Occurring Diseases in 12-Lead ECGs.” In: *2024 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*. 2024, pp. 1–8. DOI: 10.1109/BHI62660.2024.10913961.
- [139] Rikuto Yoshizawa, Kohei Yamamoto, and Tomoaki Ohtsuki. “Investigation of Data Leakage in Deep-Learning-Based Blood Pressure Estimation Using Photoplethysmogram/Electrocardiogram.” In: *IEEE Sensors Journal* 23.12 (2023), pp. 13311–13318. DOI: 10.1109/JSEN.2023.3272921.
- [140] Bharati Sharma and Jenkin Suji. “Analysis of various window techniques used for denoising ECG signal.” In: *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*. 2016, pp. 1–5. DOI: 10.1109/CDAN.2016.7570885.
- [141] Pritam Sarkar and Ali Etemad. “Self-Supervised ECG Representation Learning for Emotion Recognition.” In: *IEEE Transactions on Affective Computing* 13.3 (2022), pp. 1541–1554. DOI: 10.1109/TAFFC.2020.3014842.
- [142] Philip Schmidt et al. “Introducing wesad, a multimodal dataset for wearable stress and affect detection.” In: *Proceedings of the 20th ACM international conference on multimodal interaction*. 2018, pp. 400–408.
- [143] Win-Ken Beh, Yi-Hsuan Wu, et al. “MAUS: A dataset for mental workload assessment on N-back task using wearable sensor.” In: *arXiv preprint arXiv:2111.02561* (2021).

- [144] Edith Galy and Claudine Mélan. “Effects of cognitive appraisal and mental workload factors on performance in an arithmetic task.” In: *Applied psychophysiology and biofeedback* 40.4 (2015), pp. 313–325.
- [145] Simão Ferreira et al. “From Controlled to Chaotic: Disparities in Laboratory vs Real-World Stress Detection.” In: *2024 International Conference on Content-Based Multimedia Indexing (CBMI)*. 2024, pp. 1–7. DOI: 10.1109/CBMI62980.2024.10859254.
- [146] Gede Angga Pradipta et al. “Enhanced Fetal Arrhythmia Classification by Non-Invasive ECG Using Cross Domain Feature and Spatial Differences Windows Information.” In: *IEEE Access* 13 (2025), pp. 6729–6749. DOI: 10.1109/ACCESS.2025.3526586.
- [147] Wenqian Lin and Chao Li. “Review of studies on emotion recognition and judgment based on physiological signals.” In: *Applied Sciences* 13.4 (2023), p. 2573.
- [148] Jon Andreas Mortensen et al. “Multi-Class Stress Detection Through Heart Rate Variability: A Deep Neural Network Based Study.” In: *IEEE Access* 11 (2023), pp. 57470–57480. DOI: 10.1109/ACCESS.2023.3274478.
- [149] Youhui Deng et al. “A multispectral feature framework for predicting soybean high temperature resistance grades based on masked autoencoding and supervised contrastive learning with dual-branch pretraining.” In: *Engineering Applications of Artificial Intelligence* 165 (2026), p. 113487.
- [150] Hongzhuo Kang et al. “A composite classification algorithm of motor imagery EEG signals based on PLV-GCN+ LSTM model.” In: *Biomedical Signal Processing and Control* 112 (2026), p. 108419.
- [151] Tianlong Feng et al. “PathFusion-Net: A Rough Path Theory-Based Deep Learning Model for ECG Arrhythmia Classification.” In: *IEEE Journal of Biomedical and Health Informatics* (2026).