A Deep Learning Framework for Quantitative Stress Estimation using Wearables

by

Israa Ashraf Moustafa

A thesis submitted to the
Graduate Program in Computing
in conformity with the requirements for
the degree of Master of Science

Queen's University
Kingston, Ontario, Canada
July 2025

Abstract

Repeated exposure to everyday stressors without sufficient recovery can progressively lead to chronic stress and depression, highlighting the importance of early stress monitoring and prevention. Although advances in wearable devices have enabled stress detection, most methods adopt discrete classification, introducing several critical limitations. Discretization masks individual differences and reduces statistical power by ignoring within-group variance when studying physiological stress associations. It also weakens reliability by averaging out fluctuations and reduces temporal sensitivity for real-time monitoring.

This thesis presents a framework for estimating a continuous Quantitative Stress Index (QSI) by integrating psychometric modeling with physiological features from wearables. As a first contribution, a continuous questionnaire-based score, $\text{QSI}_{\text{Quest}}$, is constructed using Exploratory Factor Analysis (EFA) on standardized self-report measures, capturing four factors representing stress dimensions. These are aggregated using variance weighting to yield a robust, personalized stress score that outperforms the discriminative power and reliability of conventional questionnaire-derived scores. Building on $\text{QSI}_{\text{Quest}}$ as a reference, two deep learning architectures, QSI-MLP and QSI-CNN, are proposed and evaluated for continuous stress estimation from physiological signals, including electrocardiogram (ECG), blood volume pulse (BVP), and electrodermal activity (EDA). The models are benchmarked against baseline machine learning methods and established quantitative indices, demonstrating improved session-level discrimination and smoother temporal trends. In classification tasks, the ECG-EDA and BVP-EDA configurations achieved high mean accuracies of 98.81% and 98.58%, respectively, in distinguishing stress from non-stress sessions. In regression on a 0–100 stress scale, both models yielded mean absolute errors below 10, with QSI-CNN performing slightly better (ECG-EDA: 9.58; BVP-EDA: 9.82). The comparable performance of BVP-EDA to ECG-EDA underscores the practical potential of wrist-worn systems for real-world stress monitoring. By modeling stress as a continuous, multidimensional construct and integrating psychometric and physiological modalities, the proposed framework enables more realistic, personalized, and temporally sensitive stress estimation. Physiological signals capture intra-session variability, reflecting natural fluctuations in stress responses, while psychometric inputs provide subject-specific baselines. This integrated approach supports high-resolution monitoring and has broad applicability in clinical decision-making, adaptive human-computer interaction, and personalized health management, particularly in affective computing and telehealth environments.

Acknowledgments

First, I would like to thank my husband, Mohamed, for being my greatest support throughout this journey. His patience, love, and constant encouragement kept me going, especially during the toughest times. I truly couldn't have done this without him.

I want to express my heartfelt gratitude to my family: my mother, Rasha; my father, Ashraf; my sister, Salma; and my brothers, Abdulla and Mohamed. Thank you for always being there for me, believing in me, and cheering me on every step of the way.

I am deeply grateful to my supervisor, Professor Hossam Hassanein, for his guidance and support throughout my research. His expertise and insight were invaluable in shaping the direction of this work. I also extend my sincere thanks to my co-supervisor, Professor Sharief Oteafy, for his mentorship, thoughtful guidance, and consistent encouragement during my master's journey.

Thanks to everyone in the TRL Lab for providing a welcoming and collaborative research environment. I'm especially thankful to Hebatalla Ouda for her help, advice, and continuous support throughout my thesis.

Finally, I would like to thank everyone who supported me along the way. I'm genuinely grateful for this experience and for all those who were part of it.

Contents

List of Tables

List of Figures

Glossary of Acronyms

ACTH adrenocorticotropic hormone.

ANOVA one-way analysis of variance.

ANS autonomic nervous system.

BPF Band-Pass Filter.

BVP blood volume pulse.

CNN convolutional neural network.

CRH corticotropin-releasing hormone.

ECG electrocardiogram.

EDA electrodermal activity.

EDASymp sympathetic EDA power.

EEG electroencephalogram.

EFA exploratory factor analysis.

EMD empirical mode decomposition.

Extra Trees extremely randomized trees.

FIR finite impulse response.

GB gradient boosting.

HF high-frequency spectral power.

HPA hypothalamic-pituitary-adrenal axis.

HR heart rate.

HRV heart rate variability.

KMO kaiser-meyer-olkin.

KNN K-Nearest Neighbors.

LF low-frequency spectral power.

LF/HF LF to HF ratio.

LR logistic regression.

MAE mean absolute error.

MEDIAN NN NN intervals median.

ML maximum likelihood.

MLP multi-layer perceptron.

MSI mental stress index.

PANAS positive and negative affect schedule.

PCA principal component analysis.

pNN20 percentage of interval differences greater than 20 ms.

pNN50 percentage of interval differences greater than 50 ms.

PPG photoplethysmography.

PSM physiological stress model.

PSS perceived stress scale.

QSI quantitative stress index.

$\mathbf{QSI_{Quest}}$ questionnaire-based quantitative stress index.

QSI-CNN convolutional quantitative stress index model.

QSI-MLP fully connected quantitative stress index model.

RBF radial basis function.

RF random forest.

RMSEA root mean square error of approximation.

RMSSD root mean square of successive differences.

RR respiratory rate.

sAA salivary alpha-amylase.

SAM self-assessment manikin.

SCL skin conductance level.

SCR skin conductance response.

SD standard deviation.

SDNN standard deviation of NN intervals.

SNS sympathetic nervous system.

STAI state-trait anxiety inventory.

SVM support vector machine.

SVR support vector regression.

TEMP skin temperature.

TSST trier social stress test.

VAS visual analogue scale.

XGBoost extreme gradient boosting.

Chapter 1

Introduction

1.1 Motivation

Acute stress is a short-term physiological and psychological reaction to immediate perceived threats or challenges [1]. It activates the autonomic nervous system (ANS) and hypothalamic-pituitary-adrenal axis (HPA) axis, resulting in rapid bodily responses such as increased heart rate, heightened skin conductance, and cortisol release, collectively known as the "fight-or-flight" response [1, 2].

While acute stress plays a vital role in adaptive responses necessary for survival and performance, its repeated or poorly regulated activation contributes to significant long-term health risks. It is strongly associated with mental health conditions like anxiety and depression and a range of physical health conditions, including cardiovascular disease, diabetes, and cancer [3].

Globally, stress is increasingly recognized as a critical public health and economic issue. According to the world health organization (WHO), stress-related mental disorders are among the most prevalent worldwide. The American Psychological Association reported that 34% of young adults experienced extremely high stress in 2023, up from 26% in 2019 [4]. Economically, stress-related conditions, including depression and anxiety, result in the loss of 12 billion workdays annually, contributing to a global economic burden of approximately US$1 trillion due to reduced productivity, as reported by the WHO [5].

Despite growing interest in stress detection, most existing research considers stress as discrete, binary, or limited to categorical levels, oversimplifying the inherently continuous and dynamic nature of stress. This dichotomization introduces several limitations. It masks inter-individual variability in stress responses. It also reduces the statistical power of analytical methods, such as correlation and regression, by discarding within-group variance, which limits the ability to detect subtle associations between physiological indicators and stress levels. Furthermore, it overlooks potential non-linear relationships between physiological signals and stress, compromises measurement reliability by averaging out meaningful fluctuations, and decreases temporal sensitivity, hindering the detection of dynamic stress transitions in real-time applications [6].

Quantitative high-resolution stress monitoring has the potential to enhance applications such as clinical decision-making, adaptive human-computer interaction, and personal health management, especially in Affective interactions across tele-health networks [7]. In parallel, the rising adoption of wearable devices, particularly wristbands, has enabled real-time monitoring of physiological signals such as blood volume pulse (BVP) and electrodermal activity (EDA). These are directly influenced by ANS activity and are consequently highly responsive to stress-related changes, making them a promising foundation for developing quantitative, wearable-based stress scales for continuous monitoring.

1.2 Research Objective

This research aims to develop and validate a framework for estimating a continuous, individualized quantitative stress index (QSI) based on physiological signals from wearable devices. The approach addresses the limitations of discrete stress labeling by integrating psychometric and physiological data to enable high-resolution and interpretable stress quantification. The specific objectives are as follows:

1. Construct a continuous questionnaire-based quantitative stress index ($\text{QSI}_{\text{Quest}}$) using exploratory factor analysis (EFA) applied to self-reported psychological questionnaires, capturing the latent structure of acute stress responses.

2. Develop and optimize supervised deep learning regression models, fully connected quantitative stress index model (QSI-MLP) and convolutional quantitative stress index model (QSI-CNN), and compare their performance against baseline traditional machine learning models for estimating stress score from physiological features.

3. Assess the effectiveness of chest-worn and wrist-worn physiological modalities for stress estimation, analyze the performance of individual and multimodal sensor combinations, and investigate the impact of auxiliary dataset integration on model generalizability.

1.3 Contributions

This thesis presents the following key contributions:

- Development of a questionnaire-based continuous stress index, $\text{QSI}_{\text{Quest}}$, using

EFA, capturing four latent stress dimensions and applying variance-weighted aggregation to produce a robust stress score.

- Design and validation of a BVP preprocessing pipeline integrating detrending, Band-Pass Filter, and Hilbert demodulation to enhance HRV parameter estimation accuracy, using ECG-derived intervals as a reference.
- Proposal of two deep learning architectures, QSI-MLP and QSI-CNN, for continuous stress estimation from wearable physiological signals (ECG, BVP, EDA), demonstrating superior performance over traditional machine learning models. This includes empirical evaluation of wrist-worn and chest-worn configurations, showing comparable performance and supporting the feasibility of wrist-based stress monitoring.

1.4 Organization of Thesis

This thesis is organized into five chapters. Chapter 2 provides an overview of the physiological and psychological foundations of acute stress, reviewing prior work on stress detection using wearable technologies with a particular emphasis on the gap in the literature regarding quantitative stress estimation. Chapter 3 presents the development of the $\mathrm{QSI}_{\mathrm{Quest}}$, detailing the application of exploratory factor analysis to self-report questionnaire data, the validation of the resulting stress score, and comparisons with other questionnaire-based models. Chapter 4 outlines the physiological stress modeling framework, including the preprocessing of electrocardiogram (ECG), BVP, and EDA signals, the extraction and selection of relevant features, and the implementation of machine learning and deep learning models (QSI-MLP and QSI-CNN) for continuous stress estimation. This chapter also includes a comprehensive evaluation of individual and multimodal sensor configurations, examines the comparative performance of chest-worn versus wrist-worn modalities, and investigates the impact of integrating auxiliary datasets on model generalizability. Chapter 5 concludes this thesis by summarizing the key findings and contributions and outlining potential directions for future research.

Chapter 2

Background and Literature Review

The definition of stress has evolved over time, with recent literature defining it as psychological and physical strain triggered by physical, emotional, social, and cognitive stimuli that disrupt the homeostasis balance of the individual, activating a stress response to cope with the stressor [8]. Depending on its duration and intensity, it is typically classified as either acute or chronic.

2.1 Overview of the Body's Stress Response

Acute stress, a short-term reaction to an immediate stimulus, activates the sympathetic branch of the autonomic nervous system (ANS) along with the hypothalamic-pituitary-adrenal axis (HPA) axis. The stress response is initiated in the amygdala, a key brain structure involved in emotional processing. Upon perceiving a threat, the amygdala signals the hypothalamus to activate a coordinated physiological cascade. This initiates the HPA axis, wherein the hypothalamus secretes corticotropin-releasing hormone (CRH), which in turn stimulates the pituitary gland to release adrenocorticotropic hormone (ACTH). ACTH circulates through the bloodstream and activates the adrenal glands, leading to the secretion of the primary stress hormones cortisol and adrenaline [2].

Concurrently, the ANS, specifically its sympathetic branch, is engaged to support the immediate "fight-or-flight" response. Adrenaline acts rapidly on various organs, increasing heart rate and blood pressure to optimize cardiovascular output, and dilating the airways in the lungs to enhance oxygen uptake. Additionally, sweat gland activity is increased, facilitating thermoregulation and physical readiness. Blood flow is simultaneously redirected away from the digestive system, resulting in decreased gastrointestinal activity, which conserves energy for vital muscular and neural functions. Meanwhile, cortisol promotes gluconeogenesis in the liver, increasing blood glucose levels to ensure sufficient energy availability. This integrated physiological cascade is summarized in Figure 2.1. While the response enables the organism to cope effectively with acute threats, persistent activation in response to non-life-threatening stressors, such as occupational or social pressures, can result in autonomic dysregulation and adverse health outcomes [9].

2.2 Key Indicators of the Body's Stress Response

As stress activates multiple physiological systems and body organs, both biochemical and physiological measures have been extensively employed as objective indicators of stress.

2.2.1 Biochemical Markers of Stress

Biochemical markers provide valuable insight into the endocrine and neurochemical processes underlying the stress response. Among the most widely studied are cortisol

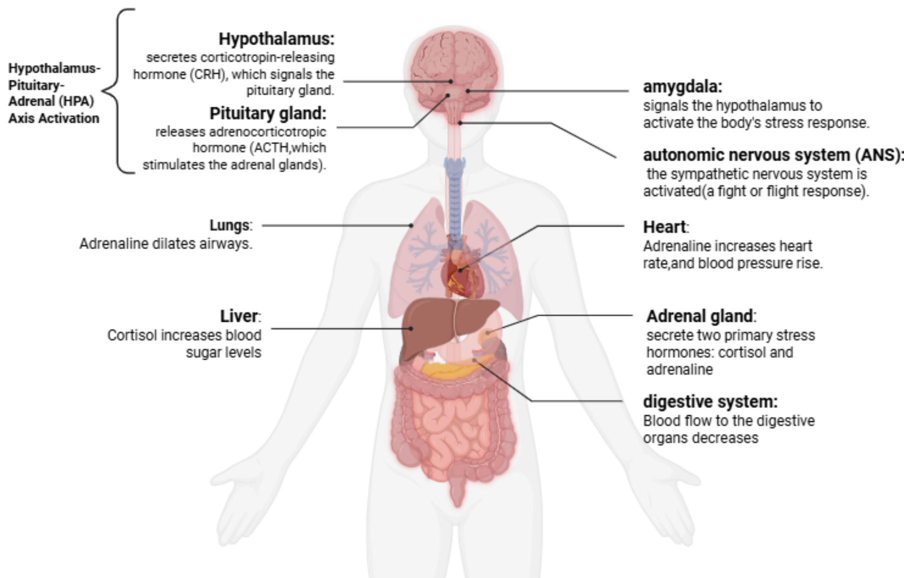


Figure 2.1: Overview of the physiological body stress response. Illustration created using BioRender.

and salivary alpha-amylase (sAA), which serve as indicators of HPA axis activity and sympathetic nervous system (SNS) activation, respectively.

Cortisol

Cortisol is considered a well-established biomarker for assessing physiological stress, as it is the principal hormone released by the adrenal cortex following HPA axis stimulation. However, single-point cortisol measurements are not considered reliable for comparing stress level between individuals due to significant fluctuations throughout the day and sensitivity to factors such as genetics, environment, physical activity, dietary intake, lifestyle and demographic factors such as age, gender. As a result, relying on a single measurement may lead to inaccurate or misleading conclusions [10]. To improve reliability, it is recommended to collect multiple samples per individual, typically between 10 and 15, to adequately capture cortisol secretion patterns and minimize measurement error [11].

Salivary Alpha-Amylase

SAA has emerged as a promising marker for acute stress due to its rapid responsiveness to sympathetic nervous activation. Unlike cortisol, which reflects slower endocrine responses, sAA levels increase rapidly in response to psychological or physical stressors, making it suitable for assessing immediate stress reactivity. While sAA can be collected relatively easily through saliva sampling in both laboratory settings, it is not suitable for real-time monitoring, limiting its application in continuous or dynamic stress assessment. Moreover, sAA levels can be influenced by factors such as sampling technique, diurnal rhythms, and individual variability, which require careful methodological control. Despite these limitations, sAA provides valuable insight into sympathetic arousal and is increasingly used alongside other biomarkers to capture a more comprehensive picture of the stress response [12].

2.2.2 Physiological Markers of Stress

Heart Activity Features

Under stressful conditions, SNS activation plays a crucial role in regulating cardiovascular function, influencing both the rate of heartbeats and the variability between successive beats, thereby supporting the body's adaptive "fight-or-flight" response. These cardiac changes enable the body to quickly mobilize resources to cope with psychological or physical threats.

Heart rate (HR), defined as the number of heartbeats per minute, is directly influenced by the autonomic nervous system. During stress, HR typically increases due to sympathetic activation and the release of catecholamines such as adrenaline. Elevated HR is considered a reliable indicator of physiological arousal and is commonly observed in response to emotional, cognitive, or physical stressors [13].

Heart rate variability (HRV) refers to the fluctuations in the time intervals between consecutive heartbeats. Under normal conditions, a healthy heart exhibits natural variability with irregular beat-to-beat intervals, reflecting the cardiovascular system's ability to adapt to changing physiological demands. HRV serves as a noninvasive marker of autonomic balance, providing insight into both sympathetic and parasympathetic nervous system activity. In stressful situations, HRV typically decreases and becomes more uniform, indicating reduced parasympathetic tone and heightened sympathetic activation [14]. HRV can be analyzed through time-domain and frequency-domain metrics.

Time-domain analysis quantifies the statistical variability of NN intervals. Commonly used features include the standard deviation of NN intervals (SDNN), the root mean square of successive differences (RMSSD), the percentage of interval differences greater than 50 ms (pNN50), and the percentage of interval differences greater than 20 ms (pNN20). These metrics are sensitive to autonomic changes and are consistently found to decrease during acute stress or cognitive load [13, 15, 16, 17, 18].

Frequency-domain analysis applies power spectral density to decompose HRV into frequency bands that reflect different aspects of autonomic function. The high-frequency spectral power (HF), typically ranging from 0.15 to 0.4 Hz, reflects parasympathetic activity, while the low-frequency spectral power (LF), ranging from 0.04 to 0.15 Hz, is modulated by both sympathetic and parasympathetic influences. The LF to HF ratio (LF/HF) is frequently used to estimate sympathovagal balance. Psychological stress is often associated with a reduction in HF power and an increase in LF power, resulting in an elevated LF/HF ratio and indicating sympathetic dominance [13, 16, 17].

Electrodermal Activity Features

Electrodermal activity (EDA), also referred to as galvanic skin response, measures the electrical conductance of the skin, which fluctuates in response to sweat gland activity. Physiologically, EDA is a direct reflection of sympathetic nervous system activation, particularly of the eccrine sweat glands. As there is minimal parasympathetic influence on these glands, EDA is considered a specific and sensitive marker of sympathetic arousal.

The EDA signal can be decomposed into two primary components, tonic and phasic, each characterized by distinct temporal dynamics and psychophysiological relevance. The tonic component, also known as the skin conductance level (SCL), represents slow-varying baseline conductance that reflects an individual's general psychophysiological state and autonomic regulation. In contrast, the phasic component captures rapid, stimulus-evoked fluctuations known as skin conductance responses (SCRs). A typical SCR is characterized by a rapid increase in conductance followed by a slower, exponential decay, corresponding to transient sympathetic activations [19].

During periods of stress or psychological arousal, the tonic SCL typically increases due to generalized sweat secretion. At the same time, stress triggers more frequent and pronounced SCRs. Under psychological stress or emotional stimulation, individuals tend to exhibit a higher frequency of SCR peaks and greater SCR amplitudes compared to baseline conditions, indicating increased sympathetic arousal[17].

To quantify EDA signals, various statistical features are extracted from both tonic and phasic components. These include the median (medianTonic, medianPhasic), mean (mTonic, mPhasic), standard deviation (stdTonic, stdPhasic), maximum amplitude (maxTonic, maxPhasic), and area under the curve (aucTonic, aucPhasic). Additionally, the sympathetic EDA power (EDASymp) index is computed to quantify the power spectral density within the 0.045–0.25 Hz band, providing a frequency-domain measure of sympathetic nervous system activity. All of these features have been shown to correlate significantly with stress and are widely utilized in affective computing stress estimation research [20].

Skin Temperature

skin temperature (TEMP) reflects changes in peripheral blood flow regulated by sympathetic nervous system activity. During psychological stress, sympathetic arousal induces vasoconstriction, leading to a decrease in TEMP, most notably in extremities such as the hands and fingers. In contrast, areas such as the face may exhibit increased temperature in response to stress. TEMP is recognized as a sensitive indicator of autonomic nervous system activity and is commonly used to assess stress responses. However, TEMP is also influenced by a variety of external and individual factors, including ambient temperature, physical activity, and the specific body site where measurements are taken. These influences can introduce variability across individuals and settings, reducing the reliability of TEMP as a standalone stress biomarker in uncontrolled environments. For this reason, environmental factors such as room temperature and lighting must be carefully controlled to obtain reliable temperature measurements in stress research [21].

Respiration Rate

Respiration is a robust physiological marker of stress, with stress typically resulting in increased respiratory rate (RR), reduced tidal volume, and irregular breathing patterns. Under acute stress, respiration often shifts from abdominal to thoracic breathing, becoming faster and shallower in response. These changes reflect increased sympathetic activity and are associated with emotional arousal. Although RR is widely used as a noninvasive measure of stress, practical challenges arise when monitoring respiration in real-world conditions. Respiratory signals are highly susceptible to artifacts from body movement, speech, and changes in posture, which can compromise measurement accuracy and reliability. Moreover, many existing methods require chest-worn intrusive sensors, limiting the feasibility of continuous monitoring in daily life [21].

2.3 Stress Level Estimation and Detection Techniques

Stress assessment can be generally categorized into subjective self-report measures and objective measures based on physiological and biochemical body responses to stress.

2.3.1 self-report Questionnaire-Based Assessments

In clinical and research settings, subjective self-report questionnaires are widely employed to quantify perceived emotions and label stress levels. Several established instruments have been validated as effective tools for assessing perceived stress [14].

These self-reports include:

- The state-trait anxiety inventory (STAI): distinguishes between temporary state anxiety and the general personality trait of trait anxiety.
- The positive and negative affect schedule (PANAS): independently measures positive and negative emotional states.
- The self-assessment manikin (SAM): a non-verbal pictorial tool that assesses emotional responses across the dimensions of valence (pleasure level) and arousal (emotional intensity), with an optional dominance dimension to reflect perceived control over emotions.
- The perceived stress scale (PSS): evaluates the extent to which individuals perceive their life circumstances as stressful, reflecting general stress appraisal.
- The visual analogue scale (VAS): a simple single-item instrument in which individuals rate their perceived stress by marking a point on a scale.

Several publicly available stress estimation datasets have employed self-report questionnaires for ground-truth labelling. The WESAD dataset [22] includes the STAI, PANAS, and SAM to label stress states under various experimental conditions. The MuSE dataset [23] incorporates the SAM and the PSS to capture participants' perceived stress level. Furthermore, Arza et al. [24] combined PSS, STAI, VAS, and Sympathetic Stress Scale questionnaires to construct a multidimensional stress label.

2.3.2 Biochemical Stress Measurement Approaches

Cortisol and sAA are widely recognized biochemical markers of stress. Petrakova et al. [25] validated sAA as a reliable indicator of sympathetic nervous system activation, while Wu et al. [26] used cortisol as a physiological reference in a quantitative stress model. Recent findings also report significant differences in the levels of sAA, cortisol, prolactin, and copeptin between stress and relaxation sessions, further supporting their association with stress responses [27].

Despite their validity, biochemical markers face several limitations for real-time or continuous stress monitoring. They are often invasive, require laboratory processing, and are affected by factors such as sampling time, environmental conditions, and individual variability. Additionally, the timing and magnitude of their physiological responses may not consistently reflect the subjective experience of stress [24]. Therefore, while useful in controlled research or clinical settings, biochemical markers are less practical for daily, continuous stress monitoring.

2.3.3 Physiological Signal-Based Stress Estimation

On the other hand, physiological stress measures have gained increasing attention due to recent advancements in wearable sensor technologies, which enable noninvasive, real-time monitoring of the body's stress response. Unlike biochemical markers, physiological signals can be continuously recorded using wearable devices in everyday settings.

Chest-worn devices are commonly used in clinical and research environments due to their proximity to core physiological sources. These devices provide high-fidelity measurements such as electrocardiogram (ECG), which records the electrical activity of the heart and is considered the gold standard for assessing cardiac activity, including HR and HRV. Respiration rate can also be captured using inductive sensors that measure chest expansion and contraction during breathing. Despite their accuracy, chest-worn sensors are generally less practical for continuous, daily use due to their discomfort and intrusiveness.

In contrast, wrist-worn devices offer greater practicality and user compliance in real-world settings. These devices can simultaneously monitor multiple physiological parameters. HRand HRV are derived from photoplethysmography (PPG), which uses low-intensity infrared light to detect changes in peripheral blood volume, generating a signal known as the blood volume pulse (BVP). EDA is measured via skin conductance using electrodes placed on the fingers or palm, reflecting SNS. TEMP is measured using infrared thermopile sensors that detect changes in wrist temperature. Each of these signals is known as a key indicator of stress, making wrist-worn devices a promising solution for continuous, real-life stress monitoring.

Among these, the Empatica E4 wristband has emerged as one of the most validated and widely adopted devices in research. It captures multiple physiological signals, including BVP, EDA, and TEMP. Schuurmans et al. [28] confirmed the validity of the Empatica E4 in estimating physiological signals, and its effectiveness in stress and affect detection has been demonstrated across various research studies, incorporating it into multiple datasets, which reinforces its trustworthiness among researchers [29].

Numerous studies have investigated stress detection using physiological signals acquired from wearable devices, with the Empatica E4 wristband being one of the most extensively utilized. Zhu et al. [30] integrated BVP, ECG, and EDA signals collected from both chest-worn and wrist-worn devices, including the Empatica E4, and trained six machine learning models: support vector machine (SVM), K-Nearest Neighbors (KNN), random forest (RF), naive bayes, logistic regression (LR), and stacking ensemble learning for stress detection. Fauzi et al. [31] applied the Empatica

E4 in a study involving continuous stress detection for hospital staff. They extracted features from EDA, TEMP, acceleration, and BVP from the WESAD dataset. , and used a wide range of classifiers including Naive Bayes, SVM, neural networks, KNN, LR, RF, Decision Trees, and ensemble techniques for stress binary classification.

Benita et al. [32] also utilized the WESAD dataset, focusing on HRV features extracted from PPG signals collected using the Empatica E4. A convolutional neural network (CNN) was implemented for stress classification, with Random Forest employed for feature selection. The selected features, HR, LF, NN intervals median (MEDIAN NN), SDNN, and SD1 (also referred to as RMSSD), were used to differentiate between 'stressed' and 'not stressed' states.

Alshareef et al. [33] proposed two deep learning architectures, a transformer and a CNN model, using multimodal signals (ECG, EDA, TEMP, and respiration) from both chest-worn devices and the Empatica E4, leveraging the WESAD dataset to enhance stress classification performance.

Chalabianloo et al. [34] conducted a comparative study using seven wearable devices (wrist- and chest-worn), including the Empatica E4, BITalino, Firstbeat Bodyguard2, Polar H10, Zephyr HxM, Samsung Gear S2, and CoreSense. They extracted multimodal features and applied classification algorithms such as SVM, RF, extremely randomized trees (Extra Trees), and light gradient boosting machine to evaluate device performance in stress detection tasks.

Zanetti et al. [35] investigated multilevel stress classification using a combination of physiological signals, including one-lead ECG, RR, BVP, and electroencephalogram (EEG). The ECG and respiration were recorded using a chest-worn Smartex device, while BVP signals were obtained from the wrist-worn Empatica wristband. The study evaluated three classification algorithms, SVM, RF, and LR, to classify stress levels based on multimodal signal input.

Outside of Empatica-based studies, Betti et al. [36] used a three-sensor setup recording ECG, EDA, and electroencephalogram (EEG) to detect stress during the Maastricht acute stress test. Physiological features were selected based on their correlation with salivary cortisol levels, and a SVM classifier was employed for stress classification.

Booth et al. [37] employed the Garmin Vivosmart 3, a wrist-worn commercial wearable, to collect HR and physical activity data. Features such as RMSSD, SDNN, and SDANN were extracted and used in models including Elastic Net, RF, multi-layer perceptron (MLP), and time-aware deep learning architectures like gated recurrent units (GRU) and long short-term memory (LSTM) networks to capture the dynamics of stress over time for long-term monitoring.

These studies demonstrate the growing integration of multimodal physiological signals with advanced machine learning and deep learning models for wearable-based stress detection. While chest-worn sensors provide high signal fidelity, wrist-worn devices, such as the Empatica E4, have gained significant research interest as a promising, more practical alternative for real-time, continuous stress monitoring.

2.4 Limitations in Existing Stress Estimation Approaches

2.4.1 Constraints of Categorical Stress Modeling

Despite the increasing interest in stress detection using physiological signals, most existing approaches treat stress as a categorical classification problem, typically binary or a few ordinal levels. While this simplification facilitates model training and evaluation, it often fails to capture the complex, dynamic, and individualized nature of stress responses.

Session-Level Labeling Assumptions

A common limitation across these studies is the use of task-driven, session-level labeling. Participants are uniformly assigned stress labels based on the type of task performed (e.g., mental arithmetic, Stroop, or public speaking), regardless of how each individual experiences or responds to the stimulus. This approach assumes homogeneity in stress perception and overlooks individual variability in stress responses that arise from differences in intellectual and social abilities, as well as emotional capacity, which influence how individuals perceive and react to identical stimuli.

Examples from Existing Studies

Several studies illustrate this reliance on categorical, session-based labeling. Most approaches adopt a binary classification framework, assigning discrete stress or non-stress labels based on task type. For instance, Alshareef et al. [33] formulated stress detection as a binary classification task by approximating a mapping function from input samples to discrete output labels representing either stress or non-stress states.

Similarly, Zhu et al. [30] and Benita et al. [32] excluded amusement-related sessions from the WESAD dataset and retained only sessions associated with stress (e.g., TSST tasks [38] such as public speaking and mental arithmetic) and non-stress (e.g., baseline and meditation). The labels were then converted into a binary format to align with the study's classification goals.

Fauzi et al. [31] followed a comparable strategy, merging amusement and relaxation sessions into a single "non-stress" class, thereby simplifying the classification task to two categories.

While some researchers have attempted to capture more nuanced stress states using multi-level labeling, these approaches still rely on task definitions rather than individualized responses. For example, Betti et al. [36] presented two classification cases: one where stress was treated as a binary problem, distinguishing between stress and non-stress sessions, and another where stress was classified into three levels: Relaxed, Stressed, and an intermediate Recovery phase that followed the stress exposure, reflecting a transitional recovery state. Zanetti et al. [35] proposed a three-level framework based on sessions: a rest condition, a moderate stress level induced by a serious game requiring motor coordination, and a high stress level caused by a mental arithmetic task.

Although multi-level labeling introduces additional granularity, it remains rooted in fixed session assumptions and continues to overlook inter-individual differences in stress reactivity. Whether binary or multi-class, such categorical frameworks risk oversimplifying the stress continuum and limit the development of personalized or continuous stress modeling approaches.

2.4.2 Challenges in Continuous Stress Quantification

While categorizing stress into discrete classes simplifies model development and evaluation, it fails to capture the inherently continuous and latent nature of psychological stress. This represents a key limitation in the existing literature. Although some efforts have been made to derive continuous stress scores, efforts remain limited in the current literature..

One common strategy involves constructing stress scales using self-report questionnaires. However, most existing methods are relatively simple, making it difficult to capture the dynamic and complex nature of stress. Common approaches include taking the direct value of a single-item scale, such as the VAS, or computing the mean or sum of items from multi-item scales, such as the Negative Affect subscale of PANAS [39]. Single-item scales are particularly vulnerable to random error and have been shown to possess lower validity compared to multi-item scales with unknown reliability [40]. Moreover, using mean or sum scores is not a valid approach either. As McNeish and Wolf [41] argue, sum scores represent a constrained form of factor analysis, imposing constraints of equal contributions of all items and equal error variances, which are assumptions that are typically untested. Relying on such constraints can lead to misleading conclusions about the underlying psychological construct.

A more structured approach to deriving a continuous stress score is exemplified by the mental stress index (MSI) proposed by Wu et al. [26], which utilized salivary cortisol as a reference to develop a regression model based on HRV features. While cortisol is widely used as a biochemical stress biomarker, its reliability as a standalone quantitative reference is limited. As previously discussed in Section 2.2.1, Cortisol levels are influenced by various confounding factors, including genetic, environmental, behavioural, and demographic factors, and exhibit significant diurnal variation, rendering single-point measurements unreliable for between-subject comparisons. As noted earlier, collecting 10–15 samples per individual is recommended to improve reliability and reduce measurement error.

Arza et al. [24, 27] proposed a different approach, combining questionnaire scores and biochemical markers into a composite stress index using principal component analysis (PCA), which was then used to train physiological stress model (PSM) using features derived from ECG, BVP, EDA, and TEMP. However, PCA is a data-driven dimensionality reduction method that focuses on maximizing variance without a conceptual basis for modeling latent psychological constructs, which limits the interpretability and theoretical relevance of its components in modeling latent constructs such as stress.

In contrast, exploratory factor analysis (EFA) offers a theory-driven statistical method specifically designed to identify latent variables underlying observed correlations. Unlike PCA,EFA assumes that the measured variables reflect common latent factors, making it more suitable for modeling psychological latent constructs [42].

2.5 Latent Variable Modeling in Psychology

2.5.1 Overview of Exploratory Factor Analysis

EFA is a statistical technique used to identify the latent structure underlying a set of observed variables. The method assumes that the covariation among observed variables can be explained by a smaller number of unobserved latent factors. Each observed variable is modelled as a linear combination of these latent factors plus a unique error term representing variance specific to that variable [42].

The mathematical representation of the EFA model is given by:

$$\mathbf{X} = \mathbf{\Lambda f} + \boldsymbol{\epsilon} \tag{2.1}$$

Where:

- $\mathbf{X} \in \mathbb{R}^{p\times 1}$ is the vector of p observed variables,

- $\mathbf{\Lambda} \in \mathbb{R}^{p \times m}$ is the matrix of factor loadings, indicating the relationship between observed variables and m latent factors (with $m < p$),
- $\mathbf{f} \in \mathbb{R}^{m \times 1}$ is the vector of latent variables (factors),
- $\boldsymbol{\epsilon} \in \mathbb{R}^{p \times 1}$ is the vector of unique errors or specific variances.

This model assumes that the observed data can be decomposed into shared variance due to common factors ($\mathbf{\Lambda f}$) and residual variance specific to each item ($\boldsymbol{\epsilon}$). The primary goal of EFA is to estimate the factor loading matrix $\mathbf{\Lambda}$, which reveals the latent dimensions that account for the observed relationships among variables.

The factor model assumes that the correlations among observed variables arise from their shared dependence on a smaller number of latent factors. The factor loadings quantify the strength and direction of association between each observed variable and each latent construct.

Under the assumption that the common factors are mutually uncorrelated and also uncorrelated with the unique error components, the total variance-covariance matrix of the observed variables can be expressed as:

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi} \tag{2.2}$$

Where:

- $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$ is the covarience matrix of the p observed variables,
- $\mathbf{\Lambda}\mathbf{\Lambda}^\top \in \mathbb{R}^{p \times p}$ represents the shared variance explained by the latent factors,
- $\mathbf{\Psi} \in \mathbb{R}^{p \times p}$ is a diagonal matrix containing the unique variances (specific errors) of each observed variable.

The diagonal structure of $\mathbf{\Psi}$ reflects the assumption that measurement errors are uncorrelated across items. A well-fitting factor model aims to ensure that the model-implied covariance structure, $\mathbf{\Sigma}$, closely approximates the observed covariance matrix, thereby capturing the latent relationships among variables.

2.5.2 Maximum Likelihood Extraction Method

The maximum likelihood (ML) method [43] is a widely used model-based approach for estimating the parameters of the EFA model. It estimates the factor loading matrix $\mathbf{\Lambda}$ and the unique variance matrix $\mathbf{\Psi}$ by maximizing the likelihood of the observed data, assuming that the observed variables follow a multivariate normal distribution.

The goal of ML estimation is to identify the model parameters that make the model-implied covariance matrix $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}$ as close as possible to the observed sample covariance matrix $\mathbf{S}$. The likelihood function to be maximized is:

$$\log \mathcal{L} = -\frac{n}{2}\left[p \log(2\pi) + \log|\mathbf{\Sigma}| + \mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{S})\right] \tag{2.3}$$

Where:

- $\mathbf{S} \in \mathbb{R}^{p\times p}$ is the sample covariance matrix of the observed variables,
- $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}$ is the model-implied covariance matrix,
- n is the sample size,
- $\mathrm{tr}(\cdot)$ denotes the trace of a matrix (sum of diagonal elements).

The estimation is performed iteratively through likelihood optimization. The ML approach yields estimates of the factor loadings and unique variances that are most likely to reproduce the observed correlations under the assumed model.

Chapter 3

Development of Questionnaire-based Quantitative Stress index

Stress is fundamentally a continuous construct that reflects varying degrees of mental and physical strain in response to different stimuli. Despite this, most studies simplify stress assessment by reducing it to binary or limited categorical levels, which obscures inter-individual variability, reduces statistical power, compromises reliability, and limits temporal sensitivity [6]. Modeling stress as a continuous variable enables more precise analysis, allowing for the detection of subtle associations with physiological indicators and the exploration of nonlinear relationships. This approach supports the development of personalized and context-aware stress recognition systems.

This chapter presents a methodology for constructing a continuous questionnaire-based stress score as a reference for the physiological stress model. Exploratory factor analysis (EFA) is applied to the items of validated questionnaires to extract latent factors representing core psychological dimensions of stress. These factor scores are aggregated using a variance-weighted approach to generate a continuous stress index, as illustrated in Figure 4.1.

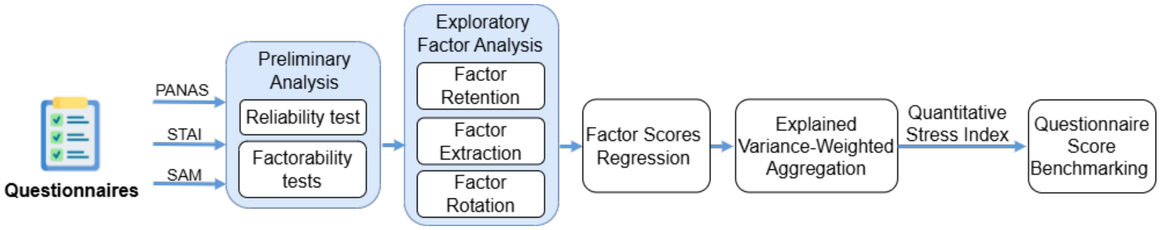


Figure 3.1: Framework for deriving the questionnaire-based quantitative stress index ($\mathrm{QSI}_{\mathrm{Quest}}$) through exploratory factor analysis followed by variance-weighted aggregation of the factor scores.

3.1 Dataset Overview

To develop a robust and interpretable quantitative stress index (QSI), we surveyed multiple publicly available physiological datasets containing synchronized self-report measures. Table 3.1 summarizes six widely used datasets in stress research and the associated subjective annotations.

Table 3.1: Comparison of stress-related datasets and their self-report annotations

Dataset	Number of Subjects	Self-report Annotations
WESAD	15	PANAS, STAI, SAM
CLAS	62	SAM
MuSE	28	PSS, SAM
StressID	65	SAM, Stress and Relax assesment
UBFC-Phys	56	Cognitive/Somatic Anxiety, Self-confidence
FatigueSet	12	Mental Fatigue VAS

3.1.1 Justification for Dataset Selection

The WESAD dataset [22] is selected for developing a continuous stress scale due to its inclusion of three validated self-report questionnaires (positive and negative affect schedule (PANAS), state-trait anxiety inventory (STAI), self-assessment manikin (SAM)) with item-level responses per session. This structure supports robust EFA and flexible modeling of latent stress dimensions, enabling the development of a reliable ground truth for the physiological stress model. Additionally, the StressID dataset [44] is used as an auxiliary dataset for model training in Chapter 4. While it lacks item-level, multi-dimensional questionnaires, it combines SAM with structured stress and relaxation assessments, offering greater inter-subject variability (65 subjects versus WESAD's 15), which enhances model generalizability.

Other datasets were excluded for the following reasons:

- **CLAS [45]:** Contains only the SAM scale, which, being a two-item measure, lacks the dimensional depth necessary for reliable factor extraction.

- **MuSE [23]:** Includes both SAM and the perceived stress scale (PSS). However, the PSS assesses stress over the past month and does not reflect session-specific variations. The limited scope of SAM alone is insufficient for meaningful factor analysis.

- **UBFC-Phys [46]:** Provides pre- and post-session scores on cognitive/somatic anxiety and self-confidence, but these are not standardized instruments and lack internal consistency and discriminative validity across stress conditions.

- **FatigueSet [47]:** Relies on a single-item visual analogue scale (VAS) for mental fatigue, which is prone to random error and lacks the validity of multi-item scales, with unknown reliability.

Accordingly, WESAD offers the most suitable structure and quality for constructing a reliable and interpretable stress index.

3.1.2 Experimental Protocol and Self-reports in WESAD

WESAD is a publicly available multimodal dataset designed for stress analysis, utilizing self-report measures and physiological signals collected via wearable devices. It comprises recordings from 15 participants who completed a structured laboratory protocol involving five distinct sessions.

The experimental protocol was designed to elicit three affective states: stress, amusement, and neutral. The stress condition involved the trier social stress test (TSST) [38], which included public speaking and mental arithmetic tasks to induce acute psychological stress. The amusement session featured exposure to humorous video clips intended to evoke positive emotional responses. Following both the stress and amusement sessions, guided meditation was used to return participants to a relaxed baseline before the subsequent session. The neutral condition, serving as a baseline, involved low-arousal activities such as quietly reading magazines.

Subjective emotional states were assessed using three validated self-report measures. The Positive and Negative Affect Schedule PANAS includes 24 items rated on a 5-point Likert scale, assessing both positive and negative affect. The STAI consists of 6 items rated on a 4-point scale, measuring anxiety by distinguishing between temporary state anxiety and more general trait anxiety. The SAM assesses emotional experience along dimensions of valence and arousal using a 9-point scale.

3.2 Preliminary Analysis for Factor Analysis Suitability

Prior to conducting EFA, the questionnaire items were evaluated to ensure their reliability for statistical analysis. Additionally, the dataset was evaluated for its suitability for factor analysis, including assessment of sampling adequacy and factorability. IBM

SPSS Statistics software, Version 29.0.2.0 (20), was used to perform all analyses, including the preliminary tests and EFA procedures.

3.2.1 Reliability Assessment

Internal consistency is assessed using Cronbach's alpha [48], a reliability coefficient that evaluates the extent to which items measure a common underlying construct. The formula for Cronbach's alpha is:

$$\alpha = \frac{n\bar{r}}{1 + (n-1)\bar{r}} \tag{3.1}$$

where n is the number of items, and $\bar{r}$ is the average inter-item correlation. The value of Cronbach's alpha ranges from 0 to 1, with values above 0.7 generally considered acceptable. A higher alpha indicates greater internal consistency among items.

3.2.2 Sampling Adequacy and Factorability

To assess whether the dataset met the assumptions for EFA, the kaiser-meyer-olkin (KMO) measure and Bartlett's Test of Sphericity are applied.

Kaiser-Meyer-Olkin Measure

Sampling adequacy refers to the extent to which observed variables share common variance, determining the suitability of a dataset for factor analysis. The KMO statistic [49] is a widely used measure for assessing sampling adequacy, evaluating whether the correlations among variables are sufficiently strong and systematic to justify the extraction of latent constructs.

It quantifies the proportion of variance that may be attributed to underlying common factors rather than unique or error variance, and is defined as:

$$\text{KMO} = \frac{\sum_{i \neq j} R_{ij}^2}{\sum_{i \neq j} R_{ij}^2 + \sum_{i \neq j} U_{ij}^2} \tag{3.2}$$

where R_{ij} is the correlation coefficient between variables i and j, and U_{ij} is the corresponding partial correlation coefficient. KMO values range from 0 to 1. Values between 0.8 and 1.0 indicate excellent adequacy, values between 0.7 and 0.79 are considered middling, and values below 0.6 suggest the sampling is inadequate for factor analysis.

Bartlett's Test of Sphericity

Bartlett's Test of Sphericity [49] assesses whether the observed correlation matrix significantly deviates from an identity matrix, thereby indicating the presence of sufficient inter-item correlations. A significant test result supports the suitability of the data for factor analysis by confirming that the variables are not orthogonal.

The test statistic χ^2_{observed}, which quantifies the deviation from the identity matrix under the null hypothesis of no inter-variable correlation, is calculated as:

$$\chi^2_{\text{observed}} = -\left(n - 1 - \frac{2p + 5}{6}\right) \ln |R| \tag{3.3}$$

where n is the sample size, p is the number of observed variables, and $|R|$ denotes the determinant of the correlation matrix.

The degrees of freedom, representing the number of unique off-diagonal elements in the correlation matrix, define the shape of the chi-square distribution used to evaluate the statistical significance of the test, and are calculated as:

$$df = \frac{p(p-1)}{2} \tag{3.4}$$

The corresponding statistical significance is determined by evaluating the probability of observing a chi-square value greater than χ^2_{observed} under the chi-square distribution with df degrees of freedom:

$$p\text{-value} = P(\chi^2_{df} > \chi^2_{\text{observed}}) \tag{3.5}$$

3.3 Exploratory Factor Analysis

EFA was employed to identify the latent structure underlying self-reported stress from validated psychological questionnaires. As a method designed to model complex and multidimensional constructs, EFA is well-suited for uncovering unobserved factors that explain the correlation patterns among questionnaire items. By capturing shared variance and accounting for item-specific measurement error, EFA facilitates the extraction of latent stress dimensions that reflect individual differences in stress perception and response [42].

EFA was conducted using the maximum likelihood (ML) extraction method [43], which estimates model parameters by maximizing the likelihood of the observed data under the assumption of multivariate normality. ML offers advantages over other extraction methods by allowing statistical tests to check how well the model fits the data. This helps in determining how many factors to retain and which variables are most suitable, thereby making the analysis of the underlying structure more reliable.

The input to the EFA model is the intercorrelation matrix derived from the observed questionnaire items, represented as the vector $\mathbf{X}$. These items reflect various aspects of perceived stress. The model decomposes $\mathbf{X}$ into two components: the common variance explained by latent stress factors $\mathbf{f}$, weighted by the factor loading matrix $\mathbf{\Lambda}$, and the unique item-specific variance $\boldsymbol{\epsilon}$. The mathematical representation of the EFA model in the context of continuous stress modeling is formally expressed in Equation 3.6.

$$\mathbf{X} = \mathbf{\Lambda}\mathbf{f} + \boldsymbol{\epsilon} \tag{3.6}$$

Where:

- $\mathbf{X} \in \mathbb{R}^{p\times 1}$: vector of p observed stress-related questionnaire items,
- $\mathbf{\Lambda} \in \mathbb{R}^{p\times m}$: factor loading matrix indicating the relationship between questionnaire items and latent stress dimensions,
- $\mathbf{f} \in \mathbb{R}^{m\times 1}$: vector of m latent stress factors,
- $\boldsymbol{\epsilon} \in \mathbb{R}^{p\times 1}$: vector of unique variances, capturing questionnaire item-specific error.

The primary output is the factor loading matrix $\mathbf{\Lambda}$, which quantifies the strength and direction of association between each item and each latent factor. Items with high loadings on the same factor are grouped to define a shared underlying dimension of stress. Based on this model, a set of individual factor scores is estimated, representing each respondent's position along the latent stress dimensions.

3.3.1 Factor Retention and Model Fit Assessment

To identify the optimal number of factors to retain in the EFA model using ML estimation, several statistical criteria were applied to ensure that the resulting factor structure effectively captured the underlying latent constructs [42].

Kaiser Criterion

The Kaiser criterion assesses the significance of each factor based on its eigenvalue. Factors with eigenvalues greater than 1 are retained, as they account for more variance than a single observed variable.

Likelihood Ratio Test

The Likelihood Ratio Test evaluates the improvement in model fit associated with increasing the number of factors. A non-significant result ($p > 0.05$) suggests that the additional factor does not provide a statistically meaningful improvement in explaining the data, thereby supporting the retention of a simpler solution with fewer factors.

Model Fit Indices

Model fit was further evaluated using the root mean square error of approximation (RMSEA), which estimates how well the factor model approximates the population covariance matrix. RMSEA values below 0.08 indicate acceptable fit, while values below 0.05 are considered indicative of good fit. The RMSEA is computed as follows:

$$\mathrm{RMSEA} = \sqrt{\frac{\chi^2 - \mathrm{df}}{\mathrm{df} \cdot (N-1)}} \tag{3.7}$$

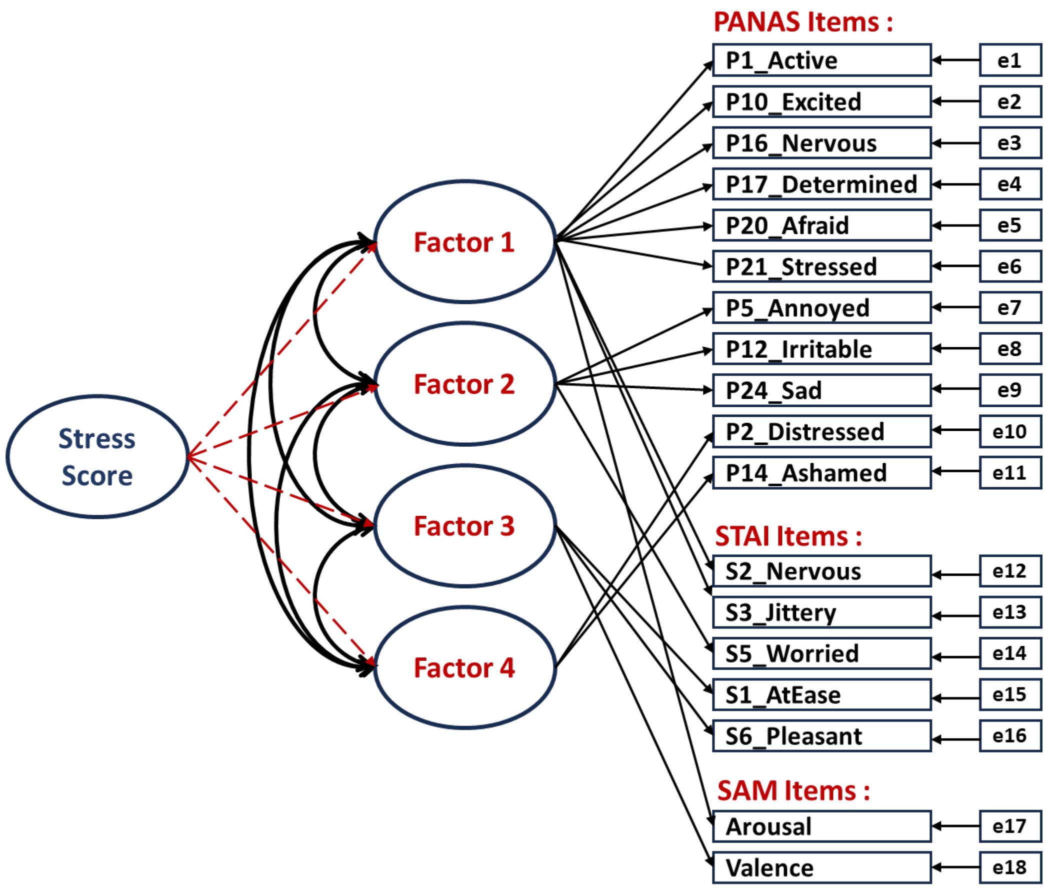


Figure 3.2: Conceptual diagram of the factor structure derived from questionnaire items. The items load onto four latent factors that represent underlying dimensions of stress and are used to derive the continuous stress score.

where χ^2 is the model chi-square statistic, df is the model degrees of freedom, and N is the sample size.

Figure 3.2 illustrates the conceptual structure of the EFA conducted using the ML extraction method on items from the PANAS, STAI, and SAM questionnaires. Each item loads onto one of four latent factors, which represent core psychological dimensions of stress. These factors are used to derive a continuous stress score that captures individual differences in emotional and cognitive responses to stress.

3.3.2 Factor Rotation

Thurstone's simple structure is a foundational principle in EFA, to enhance the interpretability of factor solutions. The goal is to ensure that each observed variable loads strongly on a single factor while exhibiting minimal loadings on other factors, thereby facilitating the identification of distinct latent constructs.

Thurstone outlined five guidelines to achieve simple structure [50]:

1. Each variable should have at least one near-zero loading< 0.3.
2. Each factor should have at least as many near-zero loadings as the number of factors.
3. For every pair of factors, several variables should show near-zero loading on one factor but not the other.
4. In analyses with four or more factors, each factor pair should include several variables with near-zero loadings on both.
5. Each factor pair should have only a few variables with significant loadings on both.

These criteria promote a solution in which each factor is defined by a unique subset of variables, minimizing overlap and improving conceptual clarity. To achieve this structure, Promax rotation was applied. As an oblique rotation method, Promax allows for correlated factors, making it particularly suitable when the underlying constructs are theoretically expected to overlap. The assumption of orthogonal uncorrelated factors would be overly restrictive because psychological stress is a multidimensional phenomenon with interrelated components [42].

The factor correlation matrix, shown in Table 3.2, confirms this assumption by displaying moderate to high inter-factor correlations, with off-diagonal values ranging from 0.463 to 0.602. These results indicate that the latent factors are not orthogonal and share substantial variance, further justifying the use of an oblique rotation method.

Table 3.2: Factor Correlation Matrix

Factor	**1**	**2**	**3**	**4**
1	1.000	0.579	0.588	0.602
2	0.579	1.000	0.602	0.463
3	0.588	0.602	1.000	0.515
4	0.602	0.463	0.515	1.000

3.3.3 Explained Variance-Weighted Aggregation of Factor Scores

Finally, factor scores are computed for each subject across all their sessions using a regression method. Inspired by the approach proposed in [24], a questionnaire-based quantitative stress index ($\mathrm{QSI}_{\mathrm{Quest}}$) is derived by calculating a weighted average of the factor scores. The weights are based on the proportion of variance explained by each factor, allowing the final score to reflect the relative contribution of each latent component. The resulting stress score is expressed on a continuous scale from 0 to 100.

3.4 Results

3.4.1 Preliminary Analysis

The preliminary tests demonstrated high internal consistency among the questionnaire items. Reliability analysis yielded a Cronbach's alpha of 0.851 across 18 selected items: 11 from PANAS, 5 from STAI, and 2 from SAM, indicating strong coherence within the item set. Sampling adequacy was also confirmed by a high KMO value of 0.898, suggesting excellent suitability for factor analysis. Furthermore, Bartlett's Test of Sphericity was statistically significant with a test statistic of $\chi^2_{\text{observed}} = 1148.30$ with $df = 153$ and a significance level of p-value < 0.001, indicating that the correlation matrix significantly differs from an identity matrix, confirming the dataset's suitability for factor analysis.

3.4.2 Determining Number of Factors

Table 3.3 presents the evaluation metrics used to determine the appropriate number of factors to retain in the EFA. The Kaiser criterion supports retaining the first four components, as each has an eigenvalue exceeding one, indicating that they explain more variance than a single observed variable. The fifth factor, with an eigenvalue below this threshold, was excluded.

The likelihood ratio test for the four-factor model yielded a non-significant result ($p = 0.061$), suggesting that adding a fifth factor would not significantly improve model fit. This is further supported by the RMSEA value of 0.058, which lies within the accepted range for good model fit (< 0.08) and approaches the threshold for excellent fit (< 0.05).

Collectively, these criteria support the selection of a four-factor solution as the most parsimonious and statistically robust representation of the latent structure underlying the questionnaire responses.

Table 3.3: Evaluation Metrics for Determining the Number of Factors

	Kaiser Criterion	Likelihood Ratio Test			Model Fit
Factors n.	**Eigenvalue**	χ^2	**df**	**p-value**	**RMSEA**
1	**9.041**	374.461	135	<0.001	0.155
2	**1.830**	221.547	118	<0.001	0.109
3	**1.307**	154.516	102	0.001	0.083
4	**1.056**	108.286	87	**0.061**	**0.058**
5	0.744	75.835	73	0.387	0.023

Table 3.4: Rotated factor loadings of questionnaire items obtained from exploratory factor analysis, conducted using SPSS software.

Item	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**
P17_Determined	0.88	-	-	-
P10_Excited	0.83	-	-	-
S2_Nervous	0.82	-	-	-
P1_Active	0.82	-	-	-
P16_Nervous	0.80	-	-	-
P20_Afraid	0.69	-	-	-
S3_Jittery	0.58	-	-	-
P21_Stressed	0.56	-	-	-
Arousal	0.48	-	-	-
P5_Annoyed	-	1.07	-	-
P12_Irritable	-	0.86	-	-
P24_Sad	-	0.79	-	-
S5_Worried	-	0.62	-	-
S6_Pleasant	-	-	-0.91	-
Valence	-	-	-0.84	-
S1_AtEase	-	-	-0.67	-
P2_Distressed	-	-	-	1.01
P14_Ashamed	-	-	-	0.62

Note: Loadings with absolute values less than 0.3 are considered negligible and were excluded from further analysis.

3.4.3 Factor Loading and Interpretation

The resulting factor loadings are presented in Table 3.4. Each retained item demonstrated a dominant loading on a single factor and negligible loadings (i.e., values ≤ 0.3) on others, supporting a well-defined and interpretable factor structure. The extracted factors were interpreted as follows:

Factor 1 reflects emotional activation, encompassing a range of high-arousal affective states. It includes both positively valenced emotions (e.g., Excited, Active, Determined) and negatively valenced yet physiologically arousing states (e.g., Afraid, Nervous, Stressed, Jittery). The inclusion of both valences suggests this factor captures general emotional intensity or arousal, independent of positive or negative tone, emphasizing arousal as a key dimension of the stress response.

Factor 2 captures negative emotionality, characterized by strong loadings on items such as Sad, Worried, and Irritable. It reflects a generalized experience of emotional strain, encompassing both affective distress and cognitive-emotional tension.

Factor 3 represents emotional tension or the absence of positive affect, characterized by negative loadings on calming and pleasurable states (e.g., AtEase, Pleasant, Valence). It reflects a state of psychological discomfort and low emotional well-being, indicating a withdrawal from affective engagement and reduced access to positive emotional experiences.

Factor 4 is interpreted as internalized distress, reflecting inwardly directed negative emotions, such as Distress and Shame. These items capture self-evaluative components of stress, including guilt, shame, and perceived personal inadequacy.

Collectively, the four factors accounted for 71.264% of the total variance, with individual contributions of 28.600%, 27.639%, 9.029%, and 5.997%, respectively. These proportions were subsequently used to compute a continuous stress score, defined as a weighted combination of the standardized factor scores:

$$\mathrm{QSI}_{\mathrm{Quest}} = 0.401 \cdot F_1 + 0.388 \cdot F_2 + 0.127 \cdot F_3 + 0.084 \cdot F_4 \tag{3.8}$$

3.4.4 Stress Score Distribution Across Sessions

Table 3.5 summarizes the descriptive statistics of QSI_{Quest} across sessions. Stress levels varied systematically, with the highest scores during the TSST Stress session and the lowest during the Fun and Meditation sessions. A one-way ANOVA confirmed a statistically significant difference among sessions ($p < 0.01$), demonstrating that the QSI_{Quest} score effectively distinguishes between different stress states.

Table 3.5: Descriptive statistics of questionnaire-based stress index across sessions

Session	Mean	Std	Min	Max
Fun	11.46	8.03	0.00	29.76
Post-Fun Meditation	13.43	5.54	6.03	29.39
Post-Stress Meditation	20.17	8.90	6.28	34.39
Baseline	28.01	13.09	8.37	48.89
Stress (TSST)	76.87	14.17	46.36	100.00

3.5 Comparison with Questionnaire-Based Methods

To evaluate the discriminative ability of the proposed QSI_{Quest} stress score relative to existing questionnaire-based stress scales, a one-way analysis of variance (ANOVA) was employed. The F-statistic in ANOVA quantifies the extent to which group means differ and is defined as the ratio of the variance between groups to the variance within groups [51]:

$$F = \frac{\text{variance between group means}}{\text{variance within groups}}$$

This statistical approach was used to assess how effectively each stress scale distinguishes between different experimental conditions.

Several stress scales used in the literature were included for comparison. These consist of single-item measures, such as 'Distressed', 'Stressed', 'Valence', and 'Nervous', as well as composite scores. One composite, referred to here as Composite A, was calculated as the average of the items [“upset”, “distressed”, “agitated”, “tense”, “nervous”], while another, Composite B, was based on the average of [“irritated”, “annoyed”, “angry”, “distressed”, “upset”, “hostile”, “stressed”] [39].

One-way ANOVA tests were conducted across the following three comparison conditions to assess the discriminative performance of each stress scale:

- Across all sessions, including Baseline, Fun, Stress, and Meditation.
- Between the stress session (TSST) and the fun session.
- Between the stress session (TSST) and the baseline session.

For each comparison, the F-statistic was computed for all stress measures, including the proposed $\text{QSI}_{\text{Quest}}$ score. The results, presented in Table 3.6, highlight the relative discriminative power of each measure in capturing session-based differences in perceived stress.

Table 3.6 presents the F-statistics obtained from one-way ANOVA for each stress scale across three session comparisons. The proposed $\text{QSI}_{\text{Quest}}$ score demonstrated the highest discriminative performance in all conditions, with a notably strong F-statistic in the TSST vs. Fun comparison.

Notably, Composite A, which simply averages “Distressed” and “Nervous” along with other items, yielded lower F-statistics than the individual item “Nervous” alone. Similarly, Composite B combined “Distressed” and “Stressed” in an unweighted average, resulting in weaker discriminative performance than “Stressed” by itself. These

Table 3.6: F-statistics from one-way ANOVA for each stress scale across different session comparisons

Stress Scale	All Sessions	Base vs TSST	TSST vs Fun
QSI_{Quest}	91.39	86.89	217.54
Nervous	58.62	70.87	126.00
Composite A	44.77	61.13	73.09
Stressed	36.22	43.15	76.66
Composite B	18.97	24.90	32.49
Valence	12.07	19.48	41.57
Distressed	9.42	13.58	16.29

patterns illustrate how aggregating items through simple averaging can dilute the contribution of high-performing indicators when equal weights are assumed. In contrast, the proposed QSI_{Quest} incorporated these same items ("Distressed," "Nervous," and "Stressed") but applied weights derived from exploratory factor analysis, accounting for their differential contributions and error variances. As a result, QSI_{Quest} achieved substantially higher F-statistics than any individual item or composite, underscoring the benefit of a psychometrically-informed approach for capturing latent stress responses across experimental conditions.

Chapter 4

Deep learning quantitative Stress index

This chapter presents the design and implementation of the proposed system model for estimating a continuous quantitative stress index (QSI) from physiological signals collected from wearable devices. As illustrated in Figure 4.1, the framework integrates multimodal physiological data with a continuous questionnaire-based quantitative stress index ($\mathrm{QSI}_{\mathrm{Quest}}$), constructed in Chapter 3. This index serves as the ground truth for supervised regression, allowing the model to learn individualized, real-time stress fluctuations. The system model encompasses signal preprocessing pipelines for electrocardiogram (ECG), blood volume pulse (BVP), and electrodermal activity (EDA), followed by feature extraction tailored to each modality. An optimized feature subset is then selected and used as input for both traditional machine learning and deep learning models. Two deep learning architectures, QSI-MLP and QSI-CNN, are proposed to capture the temporal and nonlinear structure of continuous stress responses.

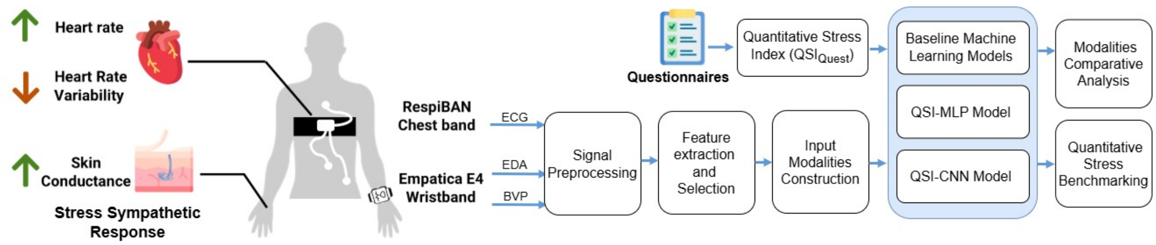


Figure 4.1: Overview of the framework for estimating a Quantitative Stress Index (QSI) by integrating features from physiological signals from wearables with the questionnaire-based ground truth (QSI_{Quest}) from Chapter 3, supporting the development and evaluation of QSI-MLP and QSI-CNN models for stress prediction across input modalities.

4.1 Experimental Datasets Description

4.1.1 WESAD Dataset

The WESAD dataset [22] encompasses physiological data collected from 15 subjects during a controlled laboratory study. Each participant is equipped with two devices: Empatica E4, a wrist wearable device used to capture key physiological signals, including BVP at 64 Hz, EDA and skin temperature (TEMP) at 4 Hz, and the RespiBAN, a chest-worn device used to capture ECG signals at 700 Hz.

4.1.2 StressID Dataset

The StressID dataset [44] is a multimodal dataset designed for stress identification. It includes synchronized recordings of physiological signals, facial videos, and audio collected under various affective conditions from 65 participants. ECG and EDA signals are recorded using the BioSignalsPlux system at 500 Hz. ECG is measured using chest electrodes, and EDA is measured using electrodes placed on the palm.

Participants completed 11 tasks designed to induce a range of emotional responses, including a 3-minute breathing session, emotional video clips, seven 1-minute cognitive and social stress tasks (e.g., Stroop test, mental arithmetic, reading) containing the TSST test [38], and a 2.5-minute relaxation phase. Each task is annotated with self-reported levels of stress, relaxation, arousal, and valence.

In this study, the WESAD dataset serves as the primary dataset for constructing the proposed quantitative stress scale. To improve the generalizability of the physiological stress model, the StressID dataset is incorporated as an auxiliary dataset, increasing the inter-subject variability and enhancing the robustness of the training process.

4.2 Physiological Signals Preprocessing

In our study, we focus on two primary physiological features: EDA features and HRV extracted from either chest-worn ECG signals or wrist-worn BVP signals, both of which are widely used as indicators of autonomic nervous system activity in stress estimation research.

4.2.1 ECG Signal

Signal Cleaning and Filtering

The raw ECG signal from the sensor was converted to SI units (millivolts) using the following scaling formula:

$$\text{ECG (mV)} = \left(\frac{\text{signal}}{\text{chan\_bit}} - 0.5 \right) \times \text{vcc}$$

where signal is the raw digital output, chan_bit $= 2^{16}$ corresponds to the 16-bit

ADC resolution, and vcc = 3 V denotes the reference voltage.

The ECG signal, which captures the heart's electrical activity, is preprocessed using a bandpass finite impulse response (FIR) filter with cutoff frequencies set at 0.67 Hz and 45 Hz to eliminate noise while preserving essential signal components. Subsequently, the DC offset is removed by subtracting the mean to correct baseline drift. These preprocessing steps are implemented using the biosppy toolbox [52].

R-Peak Detection and Artifact Correction

R-peak detection is performed using the Elgendi method [53], an efficient algorithm employing two moving averages optimized within the 8–20 Hz frequency band to enhance QRS complex detection accuracy. ECG artifacts are corrected using the robust method proposed by Lipponen and Tarvainen [54], which applies time-varying thresholds combined with a novel beat classification scheme. Both peak detection and artifact correction, as well as subsequent HRV feature extraction, are implemented through the NeuroKit toolbox [55]. Figure 4.2 presents a sample result of the preprocessing, showing the ECG signal before and after cleaning, along with the detected R-peaks.

4.2.2 BVP Signal

Signal Cleaning and Denoising

The BVP signal captures fluctuations in peripheral blood volume and reflects underlying pulse rate dynamics. To mitigate noise and improve the accuracy of heart rate variability (HRV) estimation, multiple preprocessing strategies were systematically evaluated. The objective was to identify the optimal combination of signal

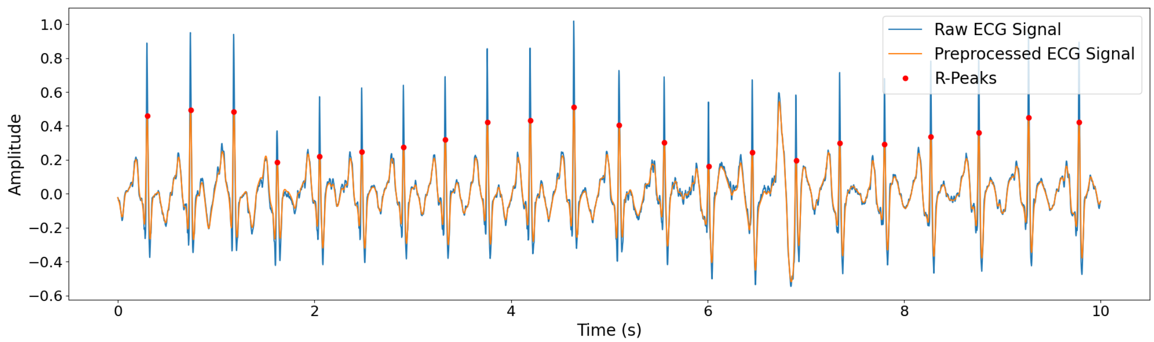


Figure 4.2: Example of ECG signal before and after preprocessing, annotated with the detected R-peaks.

processing techniques that enhances peak detection and increases the reliability of NN interval estimation.

The preprocessing pipelines incorporated the following techniques:

- **Detrending and Trend Removal:** Removes slow-varying baseline drift that may obscure physiological peaks [56].
- **Butterworth Band-Pass Filtering (0.5–9 Hz):** Retains cardiac-relevant frequency components while attenuating low-frequency drifts and high-frequency noise [57, 58].
- **Demodulation via Hilbert Transform:** Computes the analytic signal and extracts the amplitude envelope to enhance peak clarity [56, 59].
- **Empirical Mode Decomposition (EMD):** Decomposes the signal into intrinsic mode functions (IMFs) to capture dominant cardiovascular oscillations [60].

Each pipeline was evaluated on the WESAD dataset [22] by computing the mean absolute error (MAE) between NN intervals derived from BVP signals and those obtained from reference ECG recordings, quantifying the average absolute difference between corresponding NN intervals. It is defined as:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\mathrm{NN}_{\mathrm{BVP},i} - \mathrm{NN}_{\mathrm{ECG},i}| \tag{4.1}$$

where $\mathrm{NN}_{\mathrm{BVP},i}$ and $\mathrm{NN}_{\mathrm{ECG},i}$ represent the i^{th} NN interval derived from the BVP and ECG signals, respectively, and N is the total number of intervals.

Table 4.1: Mean absolute error of NN interval estimation across stress sessions for each BVP preprocessing pipeline For WESAD Dataset

Preprocessing Pipeline	**Baseline**	**Stress**	**Meditation After Stress**	**Fun**	**Meditation After Fun**	**Sessions Mean ± SD**
Raw BVP	0.163	0.382	0.077	0.253	0.058	0.187 ± 0.176
Detrend(D)	0.031	0.082	0.029	0.034	0.025	0.040 ± 0.025
BPF(F)	0.060	0.102	0.070	0.154	0.065	0.089 ± 0.088
Hilbert(H)	0.049	0.103	0.040	0.044	0.044	0.056 ± 0.041
EMD(E)	0.155	0.330	0.126	0.210	0.025	0.169 ± 0.338
D-F	0.032	0.085	0.029	0.039	0.026	0.042 ± 0.026
F-H	0.045	0.091	0.039	0.040	0.037	0.051 ± 0.037
D-H	0.032	0.041	0.029	0.033	0.029	0.033 ± 0.023
D-E	0.031	0.103	0.031	0.034	0.026	0.045 ± 0.073
H-E	0.048	0.098	0.038	0.043	0.043	0.054 ± 0.039
D-F-H	**0.031**	**0.040**	**0.026**	**0.031**	**0.028**	**0.031 ± 0.023**
D-F-H-E	0.031	0.041	0.027	0.031	0.030	0.032 ± 0.031

Among the evaluated pipelines, the combination of detrending, bandpass filtering, and Hilbert transform (D–F–H) consistently achieved the lowest MAE (0.031 ± 0.023) across all sessions, demonstrating robustness under varying physiological and emotional states. While detrending alone showed promising results, it was less effective during stress conditions. Although combining EMD with Hilbert transform offered slight improvements, it provided no additional benefit when used alongside detrending or within the full D–F–H pipeline. This indicates that the key noise components are already effectively removed through detrending and filtering, and the inclusion of EMD may introduce unnecessary variability.

To validate the selected preprocessing pipeline across varying signal characteristics and recording environments, additional experiments were conducted using the PPG-DaLiA dataset [61], as detailed in Appendix A, further supporting its robustness and generalizability.

Pulse Peak Detection and Artifact Correction

Systolic peak detection is conducted using Elgendi's method [62], which uses dual moving averages for accurate identification of pulse peaks. Artifact correction employed the robust method by Lipponen and Tarvainen [54], as previously described in the ECG preprocessing section. These steps, along with HRV feature extraction, are implemented using the NeuroKit2 toolbox [55]. Figure 4.3 presents a sample result of the preprocessing, showing the BVP signal before and after cleaning, along with the detected systolic peaks.

4.2.3 EDA Signal

EDA reflects fluctuations in skin conductance driven by sweat gland activity under sympathetic nervous system control and is widely employed as an objective marker of psychological arousal. The EDA signal consists of two primary components within the 0–2 Hz frequency range: tonic and phasic. The tonic component, also known as the skin conductance level (SCL), represents slow-varying baseline fluctuations

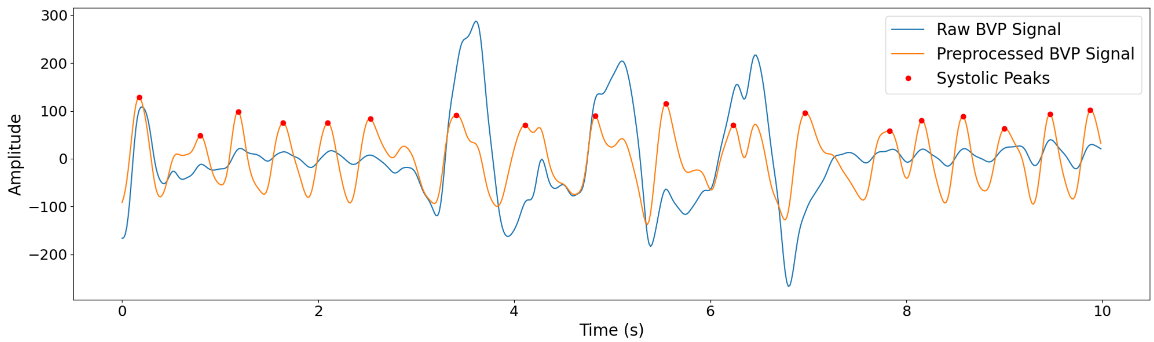


Figure 4.3: Example of BVP signal before and after preprocessing, annotated with the detected systolic peaks.

in conductance that reflect the individual's general psychophysiological state and autonomic regulation. In contrast, the phasic component captures short-term rapid fluctuations in response to discrete stimuli, known as skin conductance responses (SCRs). A typical SCR is characterized by a fast rise from baseline followed by a slower exponential decay, representing transient sympathetic responses [19].

Tonic and Phasic Decomposition

A significant challenge in EDA analysis is the overlap of consecutive SCRs, particularly when stimuli occur in rapid succession, which can lead to potential inaccuracies in decomposing the signal into its underlying components. To address this, we utilized the convex optimization model known as cvxEDA [63], which represents the EDA signal as the output of a linear time-invariant system and explicitly accounts for Gaussian noise in the measurements. This approach allows for the accurate decomposition of the EDA signal into its phasic and tonic components, even when responses overlap. Figure 4.4 presents a sample decomposition of the EDA signal into its tonic and phasic components, obtained using the cvxEDA method.

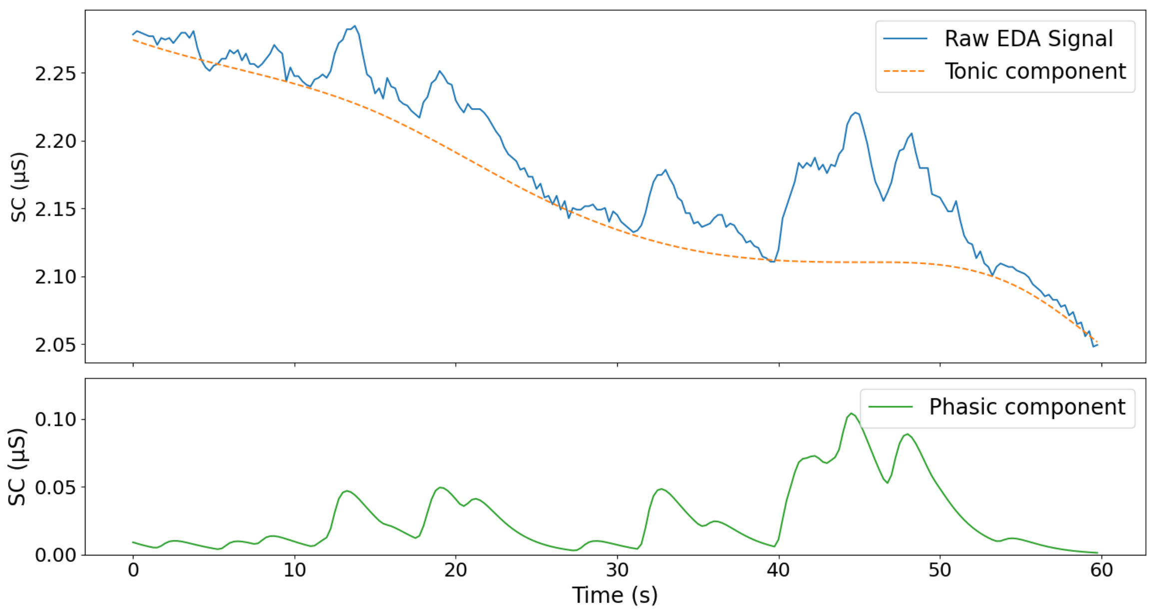


Figure 4.4: Decomposition of the EDA signal into its tonic and phasic components.

4.3 Feature Extraction from Preprocessed Signals

4.3.1 Sliding Window Segmentation and Configuration

To extract meaningful stress-related information from physiological signals, a sliding window approach is adopted. Based on prior studies [64, 65], HRV parameters in both time and frequency domains can be reliably estimated from ultra-short ECG or BVP segments as brief as 50 seconds. Accordingly, our analysis used a window length of 50 seconds with a 1-second step size to balance temporal resolution and feature stability.

4.3.2 HRV Feature Extraction from ECG and BVP

From the ECG and BVP signals, we extracted widely recognized heart rate (HR) and HRV features, which are well-established in the literature as physiological biomarkers associated with stress, as discussed in Chapter 2. These features included the root mean square of successive differences (RMSSD), standard deviation of NN intervals (SDNN), the percentage of interval differences greater than 50 ms (pNN50) and percentage of interval differences greater than 20 ms (pNN20), low-frequency spectral power (LF) and high-frequency spectral power (HF), and the LF to HF ratio (LF/HF), which reflects autonomic nervous system balance.

An individual feature ablation analysis is performed to evaluate the contribution of each feature to the regression model's performance. Beginning with the full feature set, each feature is removed independently, and the model is retrained to assess the resulting change in prediction error. The magnitude of error increase is used to rank feature importance. Based on this ranking, features with minimal impact are iteratively excluded in combination. The process is terminated once further exclusion led to a degradation in model performance, resulting in the selection of an optimal feature subset for the regression task.

4.3.3 EDA-Derived Features

From the EDA signal, we extracted statistical features including the median (medianTonic), standard deviation (stdTonic, stdPhasic), maximum amplitude (maxTonic, maxPhasic), and the area under the curve (aucTonic, aucPhasic) from both tonic and phasic components. Additionally, we included EDASymp, which quantifies

Table 4.2: Description of Extracted Physiological Features

Feature	Description
HR	Number of heartbeats per minute
RMSSD	Root mean square of successive RR intervals
SDNN	Standard deviation of NN intervals
pNN50	Percentage of interval differences greater than 50 ms
pNN20	Percentage of interval differences greater than 20 ms
LF	Power in the low-frequency band (0.04–0.15 Hz)
HF	Power in the high-frequency band (0.15–0.4 Hz)
LF/HF	Ratio of low- to high-frequency power
maxTonic	Maximum of the tonic EDA component
aucPhasic	Area under the curve of the phasic component
maxPhasic	Maximum of the phasic EDA component
stdTonic	Standard deviation of the tonic component
EDASymp	Spectral power of EDA in 0.045–0.25 Hz band

the power spectral density in the 0.045–0.25 Hz band, serving as a marker of sympathetic nervous system activity. The feature set is guided by the findings of Greco et al. [20], who demonstrated the reliability of these parameters for acute stress monitoring.

To mitigate multicollinearity, an intercorrelation analysis is conducted. Features with absolute pairwise correlations exceeding 0.95 are grouped, and within each group, the feature most strongly correlated with $\mathrm{QSI}_{\mathrm{Quest}}$ is retained. Specifically, `aucPhasic` and `stdPhasic` exhibited a 95% correlation—`aucPhasic` is retained. Among `medianTonic`, `maxTonic`, and `aucTonic`, strong pairwise correlations of 99% are observed, and `maxTonic` is selected as the representative feature.

The final EDA feature set comprised `maxTonic`, `aucPhasic`, `maxPhasic`, `stdTonic`, and `EDASymp`. A detailed description of all HRV and EDA features is provided in Table 4.2.

4.4 Quantitative Stress Index Modeling from Physiological Data

To estimate continuous stress levels from physiological signals, a range of supervised regression models is implemented, including both traditional machine learning algorithms and deep learning approaches.

4.4.1 Baseline Traditional Machine Learning Approaches

To establish baseline performance, we implemented traditional machine learning models as reference points for evaluating the effectiveness of deep learning models.

Random Forest

Random forest (RF)[66] is an ensemble learning algorithm that constructs multiple decision trees during training and outputs the average prediction of individual trees. It enhances predictive accuracy, mitigates overfitting, and is well-suited for capturing complex nonlinear relationships in physiological data.

XGBoost

Extreme gradient boosting (XGBoost) [67] is an advanced implementation of gradient boosting that builds additive decision tree models in a sequential manner. It incorporates regularization to prevent overfitting and offers computational efficiency through parallel processing and sparsity-aware learning, making it effective in modeling nonlinear physiological features.

Gradient Boosting

Gradient boosting (GB) [68] is a boosting technique that trains an ensemble of weak learners—typically shallow decision trees—by iteratively minimizing the residual errors of preceding models. It demonstrates strong performance in capturing subtle patterns in complex datasets.

Extremely Randomized Trees

Extremely randomized trees (Extra Trees) [69] is an ensemble method that constructs multiple unpruned decision trees using randomized thresholds for feature splits and bootstrapped samples. This increased randomness reduces variance and enhances generalization, particularly useful for high-dimensional physiological feature sets.

Support Vector Regression

Support vector regression (SVR) [70] is a kernel-based regression technique that maps input features into higher-dimensional spaces to find an optimal hyperplane for prediction within a predefined margin of tolerance. It is capable of capturing intricate nonlinear relationships while maintaining robustness to outliers.

4.4.2 Fully Connected Deep Model: QSI-MLP

Utilizing the large sample size, the fully connected quantitative stress index model (QSI-MLP) model is implemented to explore underlying nonlinear associations between physiological features and continuous stress levels that may not be adequately captured by conventional machine learning approaches. The QSI-MLP regression model is structured as a fully connected network composed of four hierarchical blocks,

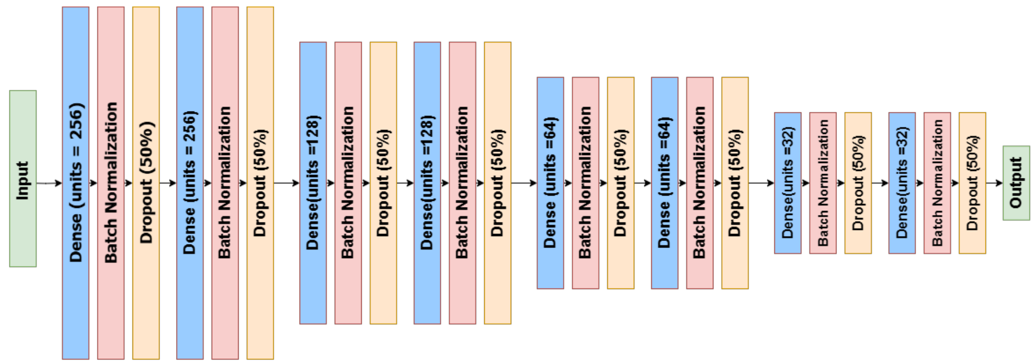


Figure 4.5: The QSI-MLP model: a deep fully connected Quantitative Stress Index model for continuous stress prediction based on HRV and EDA physiological features.

as illustrated in Figure 4.5. Each block includes two dense layers with ReLU activation, followed by batch normalization and a dropout layer with a dropout rate of 50%. The number of units in the dense layers progressively decreases across blocks (256, 128, 64, and 32 units) to extract informative representations while reducing dimensionality. A final dense layer with linear activation outputs continuous stress estimates on a 0–100 scale.

4.4.3 Convolutional Deep Model: QSI-CNN

The convolutional quantitative stress index model (QSI-CNN) model integrates convolutional layers to enhance feature extraction by learning local patterns and feature interactions that may not be captured by fully connected layers alone, followed by the same dense structure as the QSI-MLP model, as shown in Figure 4.6. The input features are reshaped to align with the input requirements of the CNN layers. The network architecture comprises 3 one-dimensional convolutional layers(Conv1D) with ReLU activation, followed by max pooling layers (pool size = 2). These convolutional layers are configured with 256, 128, and 64 filters, respectively, to extract temporal

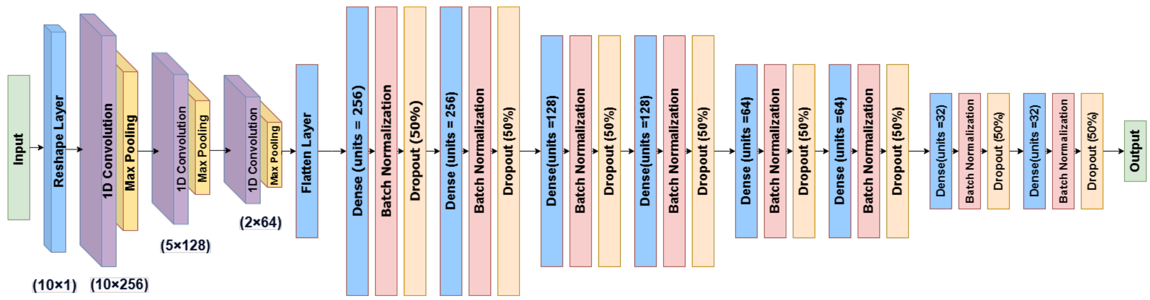


Figure 4.6: The QSI-CNN model: a convolutional quantitative stress index model integrating convolutional layers with fully connected blocks for continuous stress prediction from HRV and EDA physiological features.

features from the input signals progressively. A flattening layer is subsequently transitioned from the extracted feature maps to the dense blocks.

The dense component replicates the QSI-MLP architecture with four hierarchical blocks of dense layers (256, 128, 64, and 32 units), each followed by batch normalization and dropout (50%). A final linear layer produces the continuous stress output.

4.5 Experimental Design and Model Training

The models are trained and evaluated using different physiological input modalities, each comprising a subset of the ten features listed in Table 4.2. These features included HRV metrics derived from either ECG or BVP signals, along with EDA-derived features. The ground truth for supervised learning is the continuous stress score, $\text{QSI}_{\text{Quest}}$, introduced in Chapter 3. To account for inter-individual variability and enhance personalization, all input features are standardized relative to each individual's baseline values.

4.5.1 Training Procedure and Evaluation Metrics

Machine Learning Models Configuration

All ensemble models, RF, XGBoost, GB, and Extra Trees, were used with their default settings, including 100 estimators, a commonly adopted value that balances predictive performance and overfitting risk [71, 72]. The SVR model is implemented using its default configuration, which utilizes a radial basis function (RBF) kernel, well-suited for modeling nonlinear relationships. Regression performance is assessed using the MAE metric.

Models Configuration

Both QSI-MLP and QSI-CNN models are initially implemented as binary classification models to validate the discriminative power of the selected features and network architectures. A softmax activation function is used in the output layer to enable classification of stressed vs. non-stressed states. Model performance is evaluated using accuracy, which measures the proportion of correctly classified stress states, indicating the overall performance of the stress score..

Following classification, both models are extended to regression tasks for continuous stress prediction. In this configuration, the output layer employed a linear activation function. Regression performance is assessed using the MAE, which quantifies the average absolute difference between the predicted stress scores $\hat{y}_i$ and the ground truth scores derived from the questionnaire-based index $\mathrm{QSI}_{\mathrm{Quest},i}$, as defined in Equation 4.2.

Table 4.3: Summary of Hyperparameter Optimization Experiments

Hyperparameter	**Search Space**
Dropout Rate	0.1, 0.2, 0.3, 0.4, 0.5, 0.6
No. of Dense Blocks	1, 2, 3, 4, 5
Optimizer	SGD, Adam, RMSProp
Initial Learning Rate	0.01, 0.001, 0.0001, 0.00001

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y}_i - \mathrm{QSI}_{\mathrm{Quest},i}\right| \tag{4.2}$$

The models are trained using the Adam optimizer with a learning rate of 0.0001. A learning rate reduction callback and model checkpointing are used to preserve the best-performing model and mitigate overfitting.

To ensure robustness and subject-independence, all results are obtained using 15-fold random cross-validation, in which 6 randomly selected subjects are held out for testing in each fold, and the remaining subjects are used for training. Evaluation metrics are reported as the mean across the 15 folds, along with the corresponding standard deviation (SD) to reflect variability across splits.

Hyperparameter Optimization Procedure

A series of experiments are conducted to optimize the model's architecture and training parameters. Each setting is evaluated on validation performance to determine the best configuration. The parameters explored included dropout rate, number of dense blocks, optimizer type, and starting learning rate.

4.5.2 Physiological Input Modalities

To evaluate the individual and combined contributions of different physiological signals to stress prediction, we tested multiple input modalities. This included a comparison between chest-worn ECG sensors and wrist-worn PPG sensors, as well as the use of EDA features. The analysis aimed to determine whether each modality contributes unique predictive information and to assess the impact of multimodal fusion on model performance.

ECG-Derived Input Modality

This modality included HRV features extracted from chest-worn ECG signals, which offer high signal quality and temporal precision. It benchmarks cardiac features' effectiveness, traditionally considered the gold standard in stress research.

BVP-Derived Input Modality

HRV features are computed from wrist-worn BVP signals to assess the utility of photoplethysmography as a more accessible and wearable-friendly alternative to ECG. This modality allowed us to evaluate the trade-off between wearability and signal fidelity.

EDA-Derived Input Modality

This configuration used only EDA features extracted from wrist-worn sensors to determine the predictive value of electrodermal activity in isolation. EDA reflects sympathetic nervous system activity and has been widely used as a marker of emotional and physiological arousal.

Multimodal Fusion Input: ECG and EDA

ECG-derived HRV features are combined with EDA features to assess whether integrating cardiac and electrodermal modalities improves stress prediction. This setup aimed to explore the potential synergy between the two autonomic pathways represented by these signals.

Multimodal Fusion Input: BVP and EDA

BVP-derived HRV features are combined with EDA features to simulate a fully wrist-based wearable monitoring system and evaluate whether this configuration could achieve predictive performance comparable to chest-based setups.

4.5.3 Integration of Auxiliary Dataset

To assess the effect of increased subject variability and enhance the generalizability of model training, the StressID dataset is incorporated as an auxiliary source of physiological data. This dataset comprises 65 additional participants and includes ECG and EDA recordings. Accordingly, analyses involving StressID are limited to input modalities utilizing either or both of these signals.

To generate a continuous stress label aligned with QSI_Quest, a composite score is derived from participants' self-reported ratings of valence, arousal, stress, and relaxation. As an initial step, valence and arousal are used to compute a dimensional stress index, denoted as $\text{stress}_{\text{Dim}}$, which quantifies the affective distance from a relaxed emotional state, typically defined by high valence and low arousal, given that stress is commonly associated with low valence and high arousal [73].

$$stress_{\mathrm{Dim}} = \sqrt{(10 - \mathrm{valence})^2 + (\mathrm{arousal})^2}$$

The final composite stress score is derived through Exploratory Factor Analysis exploratory factor analysis (EFA) applied to three observed variables: self-reported stress, $\mathrm{stress}_{\mathrm{Dim}}$, and the inverse of the relaxation score. This procedure mirrored the analytical approach applied to the WESAD dataset, as detailed in Chapter 3, including preliminary evaluation of internal consistency via Cronbach's alpha. Sampling adequacy and factorability of the data are assessed using the kaiser-meyer-olkin (KMO) statistic and Bartlett's Test of Sphericity.

EFA is conducted using the Maximum Likelihood extraction method without rotation, as a single-factor solution is specified. Due to the limited number of observed variables, the Kaiser Criterion is used as the sole method for determining the number of factors, as the one-factor solution based on only three observed variables resulted in a just-identified model. In such cases, model fit indices cannot be computed because there are no remaining degrees of freedom to evaluate how well the model fits the data.

4.5.4 Data Balancing

Initially, the dataset exhibited a pronounced class imbalance, with stress sessions accounting for only 22.5% of the total 42,300 samples. Such an imbalance can negatively impact model performance, as most machine learning algorithms are optimized for overall accuracy. In this context, models tend to favour the majority class while

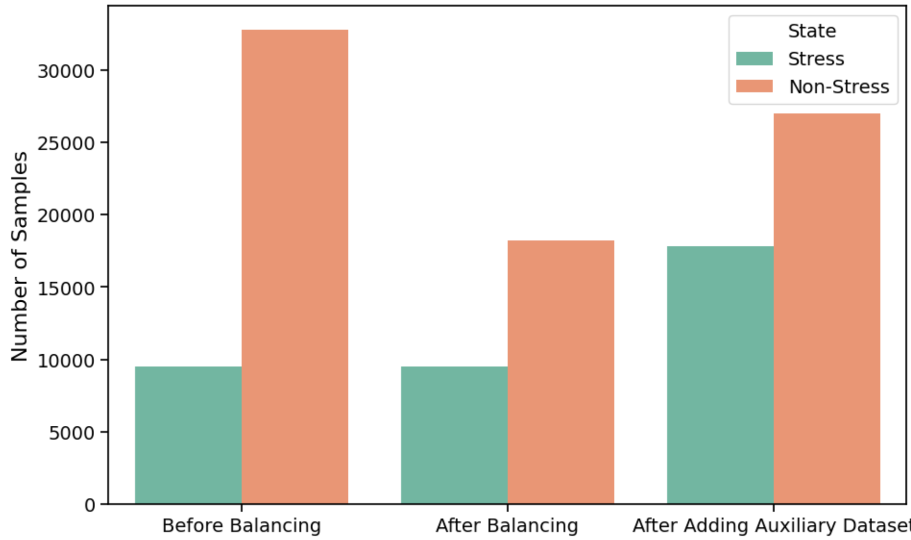


Figure 4.7: Stress and non-stress sample distribution across balancing stages.

failing to adequately learn patterns associated with the minority class [74]. This often results in poor generalization, where the models achieve high accuracy or low error by predominantly predicting the majority class values, while failing to identify critical instances of the minority class, such as high stress levels, which are especially important in stress level estimation or detection applications.

To address this issue, the dataset is first balanced by downsampling the non-stress class, increasing the stress proportion to 35% across 27,700 samples. An auxiliary dataset with a more balanced distribution is integrated, which enhances class representation without discarding valuable data. This resulted in a final dataset of 44,842 samples, with stress sessions comprising 40%. This progressive rebalancing strategy significantly mitigated class imbalance, improving the model's capacity to learn from both stress and non-stress conditions, as illustrated in Figure 4.7.

4.6 Comparative Analysis with Established Stress Indices

4.6.1 Definition of Established Stress Scores

To compare the proposed QSI-CNN against established stress indices, we implemented two benchmark models: the mental stress index (MSI) proposed in [26] and the physiological stress model (PSM) proposed in [27]. The MSI score is computed from ECG as defined in Equation 4.3, where s is a scaling factor set to 100 if $HR > 80$, and 60 otherwise. The PSM score incorporates features from ECG, BVP, EDA, and TEMP, and is computed according to Equation 4.4.

$$\mathrm{MSI} = [0.303 - 0.003 \cdot \mathrm{SDNN}_{\mathrm{ECG}} + 0.028 \cdot \mathrm{LF/HF}_{\mathrm{ECG}}] \cdot \left(\frac{\mathrm{HR}_{\mathrm{ECG}}}{s}\right) \tag{4.3}$$

$$\mathrm{PSM} = 0.5 \cdot \mathrm{mGrad} + 0.4 \cdot \mathrm{HR}_{\mathrm{BVP}} + 0.2 \cdot \mathrm{EHF}_{\mathrm{ECG}} + 0.06 \cdot \mathrm{mPAT} + 0.01 \cdot \mathrm{stdPhasic} \tag{4.4}$$

where mGrad is the mean TEMP gradient, $\mathrm{EHF}_{\mathrm{ECG}}$ denotes the extended high-frequency power of HRV derived from ECG in the (0.15–HR/2) Hz band, and mPAT is the mean pulse arrival time calculated between ECG and BVP.

4.6.2 Evaluation Metrics

Pearson correlation coefficients are used to quantify the linear association between the QSI-CNN and each benchmark model. To further evaluate the discriminative power of each stress score, a Random Forest classifier is trained using 10-fold cross-validation to classify between stress (TSST) and fun sessions, as well as between stress and baseline sessions.

Classification performance is reported using standard metrics, including accuracy, precision, recall, and F1 score. Accuracy reflects the proportion of correctly classified samples, indicating the overall performance of the model. Precision measures the percentage of predicted stress instances that are truly stress, reflecting robustness against false positives. Recall quantifies the proportion of actual stress cases correctly detected, indicating sensitivity. The F1 score balances precision and recall, providing a reliable metric when class distributions are imbalanced.

4.7 Results and Analysis

4.7.1 Auxiliary Dataset Ground Truth

A preliminary analysis assessed the reliability and structure of the candidate variables used to derive a unified stress score for the auxiliary dataset. Using valence and arousal as ground truth in place of $\text{stress}_{\text{Dim}}$ resulted in a low Cronbach's alpha of 0.34, indicating inadequate internal consistency. Similarly, substituting the relaxation score directly ,rather than its inverse, led to a negative alpha value, violating reliability assumptions.

In contrast, the combination of self-reported stress, the inverse of relaxation, and $\text{stress}_{\text{Dim}}$ produced a Cronbach's alpha of 0.813, reflecting high internal consistency. This combination was therefore retained for further analysis.

For the assessment of sampling adequacy and factorability, the KMO measure yielded a value of 0.672, indicating acceptable sampling adequacy. Bartlett's Test of

Table 4.4: Factor loadings for items contributing to $\mathrm{QSI}_{\mathrm{Quest(Aux)}}$

Item	Factor Loading
Self-reported Stress	0.871
Inverse Relaxation	0.848
$\mathrm{stress}_{\mathrm{Dim}}$	0.599

Sphericity was statistically significant, $\chi^2 = 805.81$, $df = 3$, $p < 0.001$, confirming that the correlation matrix was appropriate for factor analysis.

The number of factors was determined using the Kaiser criterion. Only one factor had an eigenvalue greater than 1 (2.185), while the second eigenvalue was substantially lower (0.553) and was therefore excluded. The retained factor accounted for 61.18% of the total variance. As shown in Table 4.4, all three variables loaded highly on this factor.

The factor score was computed using the regression method and served as the quantitative ground truth, denoted as $\mathrm{QSI}_{\mathrm{Quest(Aux)}}$, for the auxiliary dataset in subsequent model analyses.

4.7.2 Feature Selection Results

Pearson's correlation analysis was conducted to assess the linear association between each extracted physiological feature and the $\mathrm{QSI}_{\mathrm{Quest}}$ stress index. Figures 4.8, 4.9a, and 4.9b present the correlation results for features derived from EDA, as well as HR and HRV metrics computed from BVP and ECG signals, respectively

EDA features in Figure 4.8 demonstrated moderate to strong positive correlations with stress, particularly maxTonic, maxPhasic, and aucPhasic, supporting existing evidence that increased sweat gland activity is associated with sympathetic arousal.

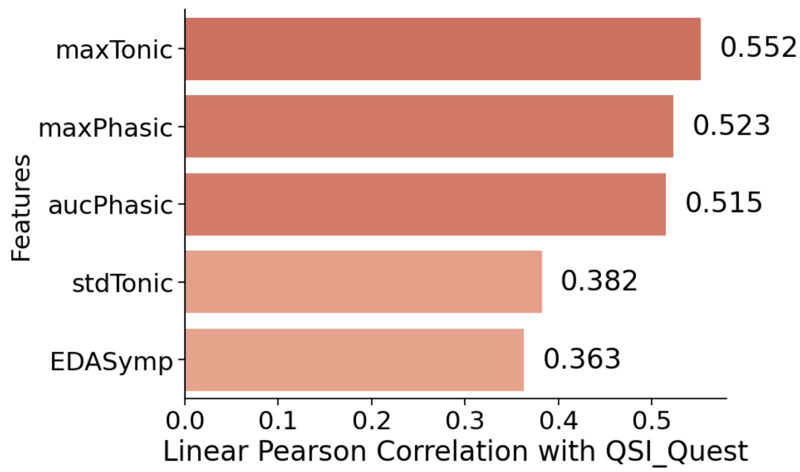
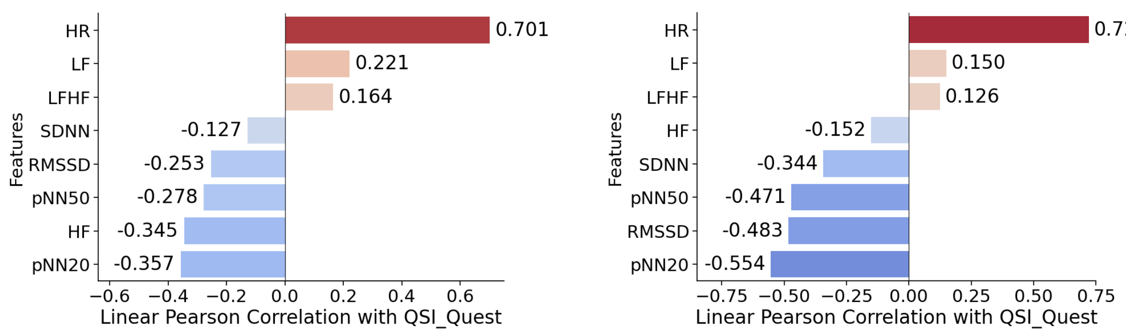


Figure 4.8: Pearson Correlation of EDA features with the Questionnaire-based Quantitative Stress Index (QSI_{Quest})



(a) BVP-derived HRV features (b) ECG-derived HRV features

Figure 4.9: Pearson Correlation of Heart Activity features with the Questionnaire-based Quantitative Stress Index (QSI_{Quest})

HRV features derived from both BVP and ECG signals exhibited the expected patterns: HR showed the strongest positive correlation with stress, reflecting increased sympathetic activation. In contrast, Time-domain HRV metrics such as RMSSD, pNN20, and pNN50 demonstrated negative correlations with stress, consistent with the shift toward more uniform inter-beat intervals during psychological stress. HF, which reflects vagal tone, also showed a negative correlation, supporting its association with parasympathetic withdrawal. LF and the LF/HF ratio showed weaker positive correlations, which reflects the shift toward sympathetic dominance typically observed in stressful conditions, as illustrated in Figure 4.9.

Notably, while the directionality of correlations was consistent across ECG and BVP modalities, the strength of association was generally higher in ECG-derived features. This suggests a slight advantage of ECG in capturing autonomic changes, likely due to its greater signal fidelity.

Individual Ablation Feature Selection

The results below show the selected BVP and ECG features retained after individual ablation, based on their contribution to reducing the MAE. All results were obtained using 15-fold random cross-validation using the QSI-CNN model, and the mean MAE was used to evaluate performance. Only features whose removal degraded model performance were retained and are reported in order of importance.

The initial model trained on all BVP-derived features achieved a mean MAE of 11.19. Among these, HR proved most critical, as its removal significantly increased the error to 18.96. In contrast, excluding SDNN improved performance, reducing the MAE to 10.53. Although removing LF/HF slightly increased the error to 10.76, it was retained due to its overall contribution. The final set of retained BVP features is summarized in Table 4.5.

Using the complete set of ECG-derived features, the initial model yielded a mean MAE of 12.17. Feature ablation showed that removing pNN20 and HF sequentially improved performance, reducing the error to 11.42 and 11.26, respectively. Further

Table 4.5: Retained BVP features sorted by importance based on mean MAE increase after ablation

Feature	MAE after removal	MAE Increase
HR	18.96	8.43
pNN20	11.48	0.95
HF	11.43	0.90
LF	11.36	0.83
pNN50	11.26	0.73
RMSSD	11.24	0.71
LF/HF	11.16	0.63
Aggregated feature set	**10.53**	–

Table 4.6: Retained ECG features sorted by importance based on mean MAE increase after ablation

Feature	MAE after removal	MAE Increase
HR	19.92	9.05
LF	12.61	1.74
LF/HF	12.09	1.22
SDNN	11.95	1.08
RMSSD	11.51	0.64
Aggregated feature set	**10.87**	–

exclusion of pNN50 led to an MAE of 10.87. However, removing RMSSD subsequently increased the error to 11.36, so it was retained. The final set of retained ECG features is presented in Table 4.6.

Although LF, HF, and LF/HF did not exhibit strong linear correlations with $\mathrm{QSI}_{\mathrm{Quest}}$, their inclusion improved model performance within both the BVP and ECG modules. Specifically, LF and LF/HF contributed to error reduction in the ECG-based model, indicating a nonlinear but meaningful association with stress. For this reason, these features were retained despite their limited linear relationship with the stress index.

4.7.3 Hyperparamter optimization Results

The hyperparameter tuning experiments revealed that a dropout rate of 0.5 provided the best regularization effect without underfitting. Increasing the number of dense blocks improved the model's learning capacity, with four blocks offering the best trade-off between complexity and performance.

Among the optimizers tested, Adam and RMSProp yielded similar high performance, while SGD underperformed. A learning rate of 0.0001 proved optimal, offering a balance between convergence speed and training stability. These findings were incorporated into the final configuration for both QSI-MLP and QSI-CNN models.

4.7.4 Classification Performance Evaluation

Table 4.7 demonstrates that across all input modalities, the QSI-CNN consistently outperforms the QSI-MLP in binary classification of stress versus non-stress states. Accuracy improvements, while modest, were consistent across modalities, ranging from 0.1% to 0.9%.

Among the individual modalities, EDA achieved the highest classification accuracy using the QSI-CNN model (94.89%), likely due to its high sensitivity to sympathetic nervous system activation under stress. BVP and ECG followed closely, with accuracies of 94.04% and 93.16%, respectively, indicating that HRV features derived from wrist-worn BVP sensors slightly outperform those from chest-worn ECG in the baseline classification setting. However, with the inclusion of auxiliary data for ECG and EDA, the ECG-based model improved to 94.26%. This improvement suggests that ECG-derived features benefit from increased inter-subject variability, indicating a potential for enhancing BVP-based model performance through the integration of

Table 4.7: Mean accuracy (% ± SD) of initial classification models QSI-MLP and QSI-CNN across input modalities

Input Modality	QSI-MLP	QSI-CNN
BVP	93.56 ± 3.02	**94.04 ± 2.65**
ECG	92.27 ± 4.07	**93.16 ± 5.59**
ECG (Aux)	93.75 ± 2.87	**94.26 ± 1.96**
EDA	94.78 ± 4.03	**94.89 ± 5.10**
EDA (Aux)	94.97 ± 2.41	**95.21 ± 3.29**
BVP-EDA	98.07 ± 1.08	**98.58 ± 0.94**
ECG-EDA	97.42 ± 3.73	**97.94 ± 1.62**
ECG-EDA (Aux)	98.23 ± 1.24	**98.81 ± 1.97**

Note: (Aux) indicates models trained with additional auxiliary data. Underlined values indicate best performance (highest Accuracy) within each row.

similarly diverse auxiliary data.

Multimodal configurations yielded the highest classification performance across all settings. The QSI-CNN model achieved an accuracy of 98.81% using combined ECG and EDA inputs with auxiliary data, the best result among all configurations. Even in the absence of auxiliary data, the combination of BVP and EDA achieved 98.58%, substantially outperforming all single-modality models. These results demonstrate that fusing multiple physiological signals provides complementary information not captured by any individual modality alone. The classification experiments thus provide empirical validation of the model's ability to reliably distinguish between stress and non-stress states, while also confirming the discriminative utility of the extracted features prior to their deployment in subsequent regression tasks.

Table 4.8: Mean MAE (± SD) for traditional machine learning regression models across physiological input modalities, including the effect of auxiliary dataset integration.

Input Modality	**RF**	**XGBoost**	**GB**	**Extra Trees**	**SVR**
BVP	13.67 ± 1.15	14.15 ± 1.07	**12.40 ± 1.11**	13.17 ± 1.07	12.95 ± 1.44
ECG	13.88 ± 1.63	13.92 ± 1.72	**12.27 ± 1.59**	13.46 ± 1.50	12.82 ± 1.50
ECG (Aux)	13.50 ± 0.86	12.79 ± 0.78	**11.69 ± 0.80**	13.32 ± 0.82	11.87 ± 1.13
EDA	14.88 ± 1.91	15.00 ± 1.97	14.23 ± 1.89	14.26 ± 1.78	**13.74 ± 1.64**
EDA (Aux)	14.52 ± 1.60	13.92 ± 1.57	13.04 ± 1.60	14.20 ± 1.84	**12.68 ± 1.66**
BVP-EDA	12.86 ± 0.91	12.84 ± 0.85	**11.84 ± 1.00**	**11.64 ± 0.96**	**11.85 ± 1.14**
ECG-EDA	12.74 ± 1.01	12.21 ± 0.82	**11.30 ± 1.05**	**11.16 ± 0.95**	**11.45 ± 1.06**
ECG-EDA (Aux)	11.59 ± 0.76	11.43 ± 0.89	**10.29 ± 0.87**	**10.34 ± 1.05**	**10.63 ± 1.26**

Note: (Aux) indicates models trained with additional auxiliary data. Underlined values indicate best performance (lowest MAE) within each row.

Table 4.9: Mean MAE (± SD) for proposed deep learning models across physiological input modalities, including the effect of auxiliary dataset integration.

Input Modality	**QSI-MLP**	**QSI-CNN**
BVP	10.75 ± 1.56	**10.53 ± 1.43**
ECG	11.09 ± 1.77	**10.87 ± 1.15**
ECG (Aux)	10.65 ± 1.59	**10.42 ± 1.33**
EDA	12.62 ± 2.00	**12.32 ± 2.20**
EDA (Aux)	12.42 ± 1.97	**12.14 ± 1.85**
BVP-EDA	10.14 ± 0.97	**9.82 ± 0.93**
ECG-EDA	9.80 ± 1.25	**9.58 ± 0.83**
ECG-EDA (Aux)	9.10 ± 1.18	**8.94 ± 0.90**

Note: (Aux) indicates models trained with additional auxiliary data. Underlined values indicate best performance (lowest MAE) within each row.

4.7.5 Regression Performance Evaluation

Model Comparison and Analysis

For single-modality configurations, both GB and SVR demonstrated strong performance in the regression task (Table 4.8). GB achieved the lowest MAE for BVP and ECG (12.55 and 12.77, respectively), underscoring its effectiveness in modeling HRV-derived features. GB's sequential error-correcting mechanism enables it to efficiently capture subtle nonlinear relationships. SVR, on the other hand, achieved the best performance for EDA-based features, likely stemming from its ability to capture smooth, individualized stress-response patterns through margin-based learning and kernel functions, while maintaining robustness to noise within the feature space.

In multimodal configurations, GB, Extra Trees, and SVR exhibited comparable performance, with Extra Trees slightly outperforming the others in most cases, as illustrated in Table 4.8. This advantage can be attributed to Extra Trees' randomized feature selection and ensemble diversity, which did not yield substantial benefits in single-modality setups due to limited input dimensionality. However, in multimodal settings, these characteristics become more advantageous, allowing the model to effectively manage increased feature dimensionality and capture complex interactions across modalities. Its ability to generalize well in the presence of weakly correlated but complementary features likely underpins its superior performance in multimodal regression tasks.

Despite their strong performance, all traditional models were ultimately outperformed by both QSI-MLP and QSI-CNN across all input modalities, as shown in Table 4.9. This highlights the advantage of deep learning in capturing more expressive and complex feature interactions. Across all configurations, QSI-CNN consistently outperformed QSI-MLP, with MAE improvements ranging from 0.12 to 0.32. For example, in the ECG-EDA configuration with auxiliary data, QSI-CNN achieved an MAE of 8.94, compared to 9.10 for QSI-MLP. While these differences are moderate, their consistency indicates that CNNs are more effective at modeling structured feature dependencies and local interactions, particularly when the input features exhibit underlying interdependence.

Single-Modality vs. Multimodal Performance

The addition of EDA led to consistent performance improvements across all machine learning models when transitioning from individual to multimodal configurations. For BVP, the reduction in error ranged from 0.56 to 1.53, while ECG showed greater improvement, with reductions ranging from 0.97 to 2.30 (Table 4.8). Similarly, in the deep learning models, transitioning from individual BVP to BVP-EDA reduced error by 0.61 in QSI-MLP and 0.71 in QSI-CNN. For ECG, the shift to a multimodal ECG-EDA input resulted in a larger reduction of 1.29 in both models (Table 4.9). These results indicate that ECG benefits more from multimodal integration, likely because EDA captures peripheral sympathetic variability that complements the central cardiac signals of ECG. Moreover, the gains were more pronounced in the deep learning models, suggesting that more complex architectures are better able to exploit the added variability and feature richness provided by multimodal inputs.

The added value of combining signals lies in capturing distinct stress-related variabilities that are not fully represented in any individual modality alone. EDA reflects sympathetic arousal, while BVP and ECG provide cardiac-derived HRV measures. Their integration enables the model to detect more complete stress signatures.

However, this performance gain must be balanced against user burden and the practicality of wearable deployment in real-world settings. While ECG-EDA offers the best performance, it requires two separate wearables. In contrast, BVP-EDA, both obtainable from a single wrist-worn wearable, provides strong performance by capturing complementary information without the need for additional wearables.

wrist-worn vs. chest-worn Performance

Although ECG is traditionally preferred for HRV analysis due to its signal stability, the results show that BVP achieved slightly lower prediction error than ECG, with MAEs of 10.53 vs. 10.87 for QSI-CNN and 10.75 vs. 11.09 for QSI-MLP, as illustrated in Table 4.9. This suggests that stress may introduce additional variability in peripheral vascular signals, which is more effectively captured by BVP than by ECG. Such variability, likely stemming from stress-induced changes in vascular tone, may be less evident in cardiac electrical activity. Additionally, effective signal preprocessing may have mitigated common PPG limitations, such as motion artifacts, allowing the model to better exploit informative features from the BVP waveform.

When combined with EDA, the ECG-EDA configuration outperformed BVP-EDA, achieving an MAE of 9.58 vs. 9.82. This indicates that EDA contributes peripheral autonomic variability that enhances ECG-based predictions, given the ECG's central and rhythmically stable nature. In contrast, BVP already captures similar sympathetic responses to those reflected in EDA, resulting in limited additional benefit from combining the two.

However, the performance gap between ECG-EDA and BVP-EDA is minimal.

Given that ECG-EDA requires two separate wearables, while PPG-EDA can be measured from a single wrist-worn device, the latter offers a more practical and minimally obtrusive solution. The comparable results suggest that the additional contribution of ECG in the ECG-EDA setup is not substantial enough to justify the use of a second wearable. Therefore, with suitable preprocessing, the BVP-EDA configuration provides nearly equivalent performance with greater convenience, making it a more preferable option for real-world stress monitoring.

Effect of Auxiliary Dataset Integration

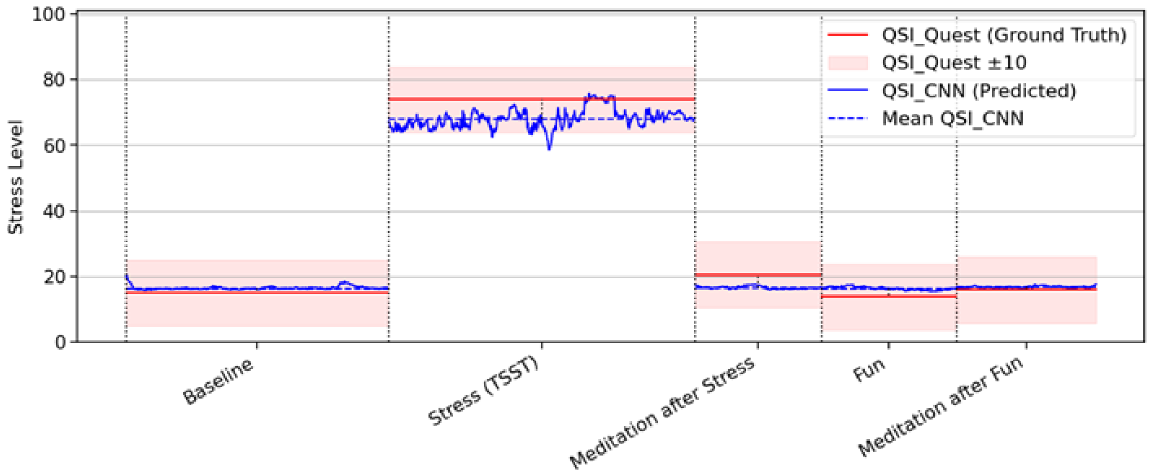
The integration of auxiliary data consistently improved performance across modalities. For example, using Gradient Boosting, the error for ECG was reduced by 0.58, while SVR applied to the EDA modality achieved a larger reduction of 1.06. In multimodal configurations, incorporating auxiliary data into the ECG-EDA setup yielded an average reduction of 0.88 (Table 4.8). Although the performance gains in deep learning models were more modest, both QSI-MLP and QSI-CNN still benefited from auxiliary data. Error reductions of 0.45 for ECG, 0.18 for EDA, and 0.64 for ECG-EDA were observed. Specifically, the ECG-EDA configuration improved from (9.80 ± 1.25 and 9.58 ± 0.83) to (9.10 ± 1.18 and 8.94 ± 0.90) for QSI-MLP and QSI-CNN, respectively, achieving the lowest error across all evaluated modalities and models (Table 4.9). These results underscore the importance of dataset diversity and the benefit of dataset augmentation in enhancing model generalization across both traditional and deep learning approaches. Although BVP was not evaluated under auxiliary integration due to data limitations, similar improvements could be expected if additional data were available, giving a great potential for improvement.

4.7.6 Qualitative Comparison Across Modalities

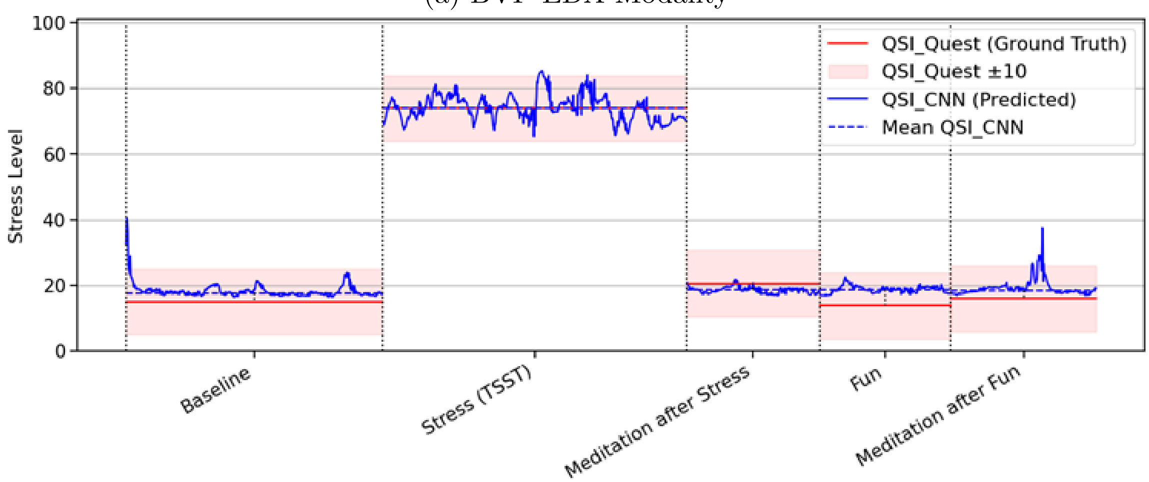
Given that the primary objective of this study is to estimate a continuous quantitative stress index that reflects temporal variations in physiological responses and accounts for individual differences, it is important to examine the model's behaviour beyond aggregate performance metrics. Unlike binary labels, real-life stress fluctuates throughout a session. Visualizing the predicted stress values across sessions enables qualitative evaluation of the model's ability to reflect these dynamic patterns, offering a more realistic representation of actual stress experiences.

The predicted stress levels illustrated in Figure 4.10 from both the BVP-EDA and ECG-EDA models generally follow the same pattern as the $\mathrm{QSI}_{\mathrm{Quest}}$ reference, especially during the stress (TSST) session, where fluctuations are more pronounced. In contrast, during baseline and relaxation periods, the predictions are more stable with fewer fluctuations, reflecting the lower and steadier nature of stress levels in these calmer conditions. This behavior supports the model's ability to reflect dynamic stress levels during emotionally intense periods, while maintaining stability during calm intervals. When auxiliary data is added to the ECG-EDA model 4.10c, the predicted stress signal becomes more uniform and consistent across time. These smoother fluctuations suggest better alignment with the ground truth and reduced prediction noise. The improved signal stability supports the quantitative finding of lower prediction error, as it helps the model capture more reliable stress patterns, particularly by reducing noise and overfitting, resulting in a more stable and accurate prediction trajectory.

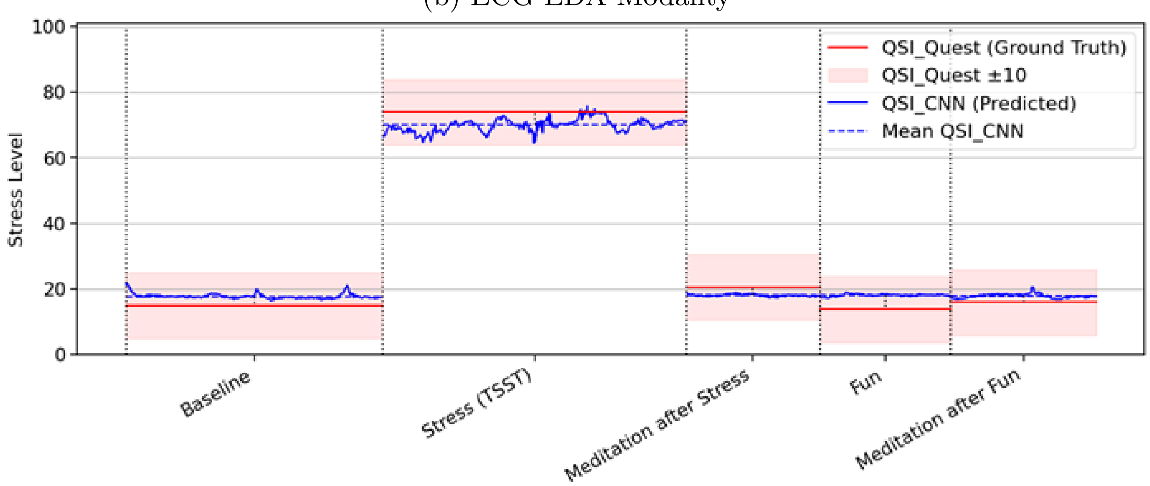
Figure 4.11 depicts that the EDA individual Modality shows a high level of noise across all sessions, particularly during the baseline and stress segments. The BVP



(a) BVP-EDA Modality

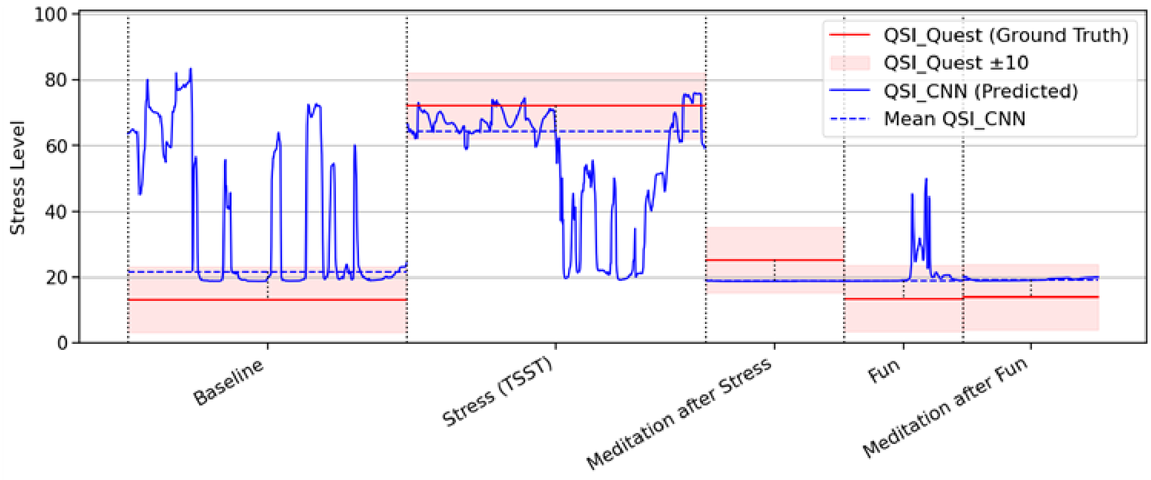


(b) ECG-EDA Modality



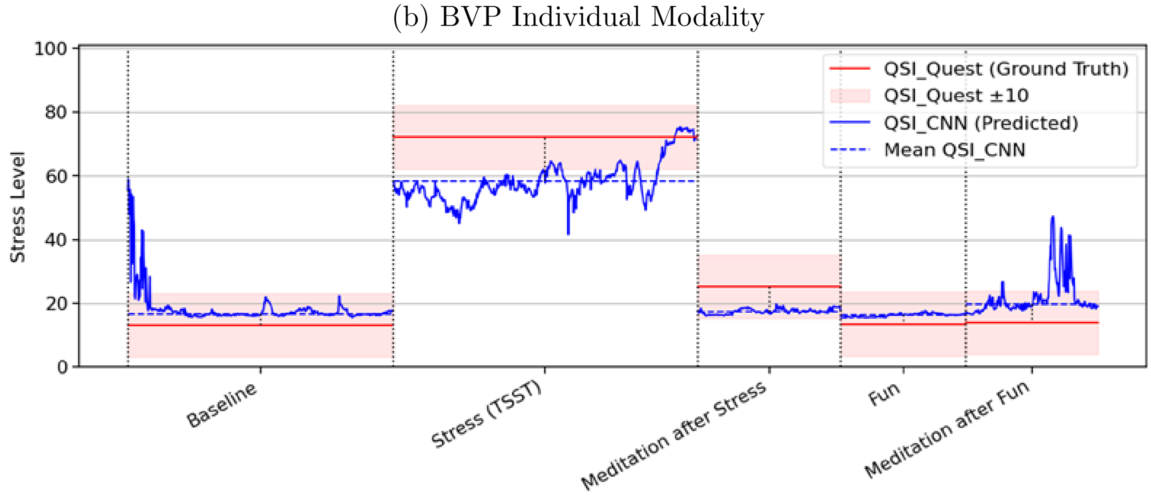
(c) ECG-EDA (Aux) Modality

Figure 4.10: Predicted stress levels by QSI-CNN across different physiological input modalities for Subject 4.



(a) EDA Individual Modality

(b) BVP Individual Modality



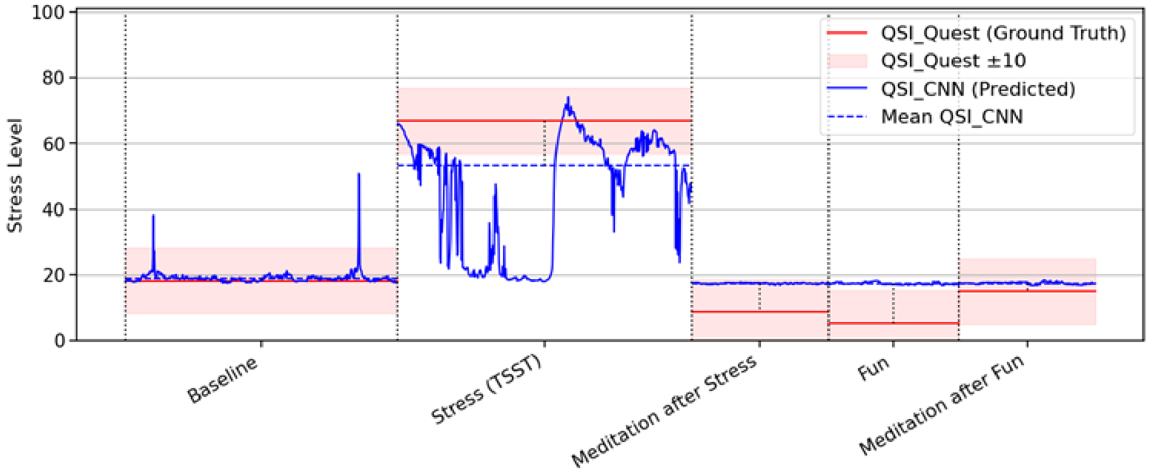
(c) BVP-EDA Multimodality

Figure 4.11: Comparison of the effect of multimodality versus individual modality for Subject 11.

Modality prediction also exhibits noise, especially during the stress session, with frequent irregular spikes. However, when both modalities are combined, the predicted stress level becomes notably smoother and more stable, closely aligning with realistic patterns of physiological stress. The combined model mitigates the spikes and noise seen in the individual modalities, indicating that multimodal inputs help compensate for modality-specific artifacts.

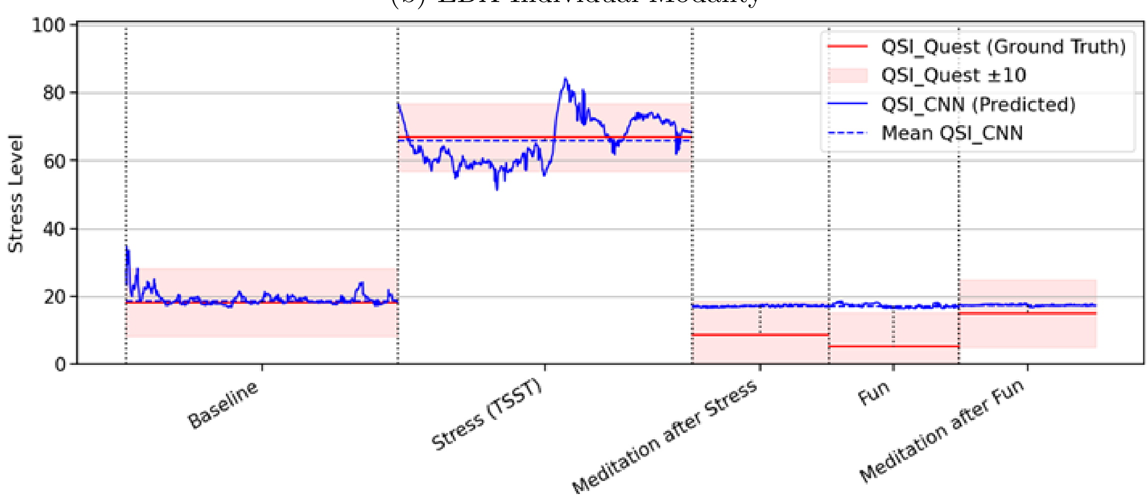
Additionally, a gradual decline in predicted stress is observed at the beginning of the baseline session, leveling off into a stable, low-stress state. This pattern, noted in several participants such as in Figure 4.11c, reflects a natural transition from initial alertness to a relaxed resting state, supporting the model's ability to capture realistic, time-varying stress dynamics.

As illustrated in Figure 4.12, signal noise is not exclusive to wrist-worn modalities such as EDA or BVP but is also present in chest-worn ECG signals. The ECG prediction displays prominent noise spikes throughout the stress sessions, indicating susceptibility to artifacts even in traditionally stable modalities. Although the EDA signal also exhibits some irregularities, the combination of ECG and EDA leads to a notably improved prediction, mitigating the noise present in the ECG input. This suggests that ECG alone may not provide a sufficient or consistent measure of sympathetic nervous system activation. Given the complex and distributed nature of physiological stress responses, multimodal integration enhances robustness by capturing complementary information across systems, resulting in more accurate and stable stress estimates.



(a) ECG Individual Modality

(b) EDA Individual Modality



(c) ECG-EDA Multimodality

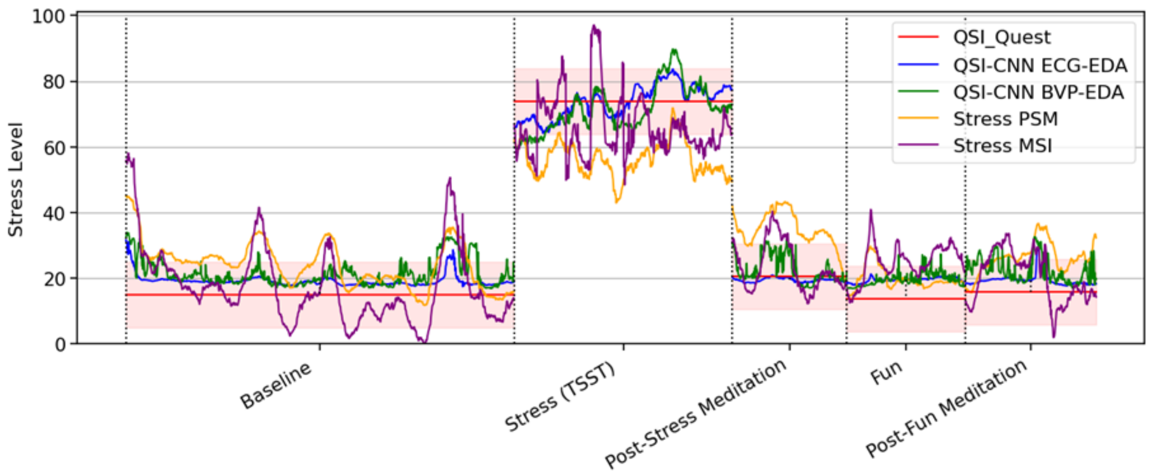
Figure 4.12: Comparison of individual versus combined ECG and EDA modalities for Subject 7, showing the effect of combining chest- and wrist-worn signals.

4.7.7 Performance Comparison with MSI and PSM

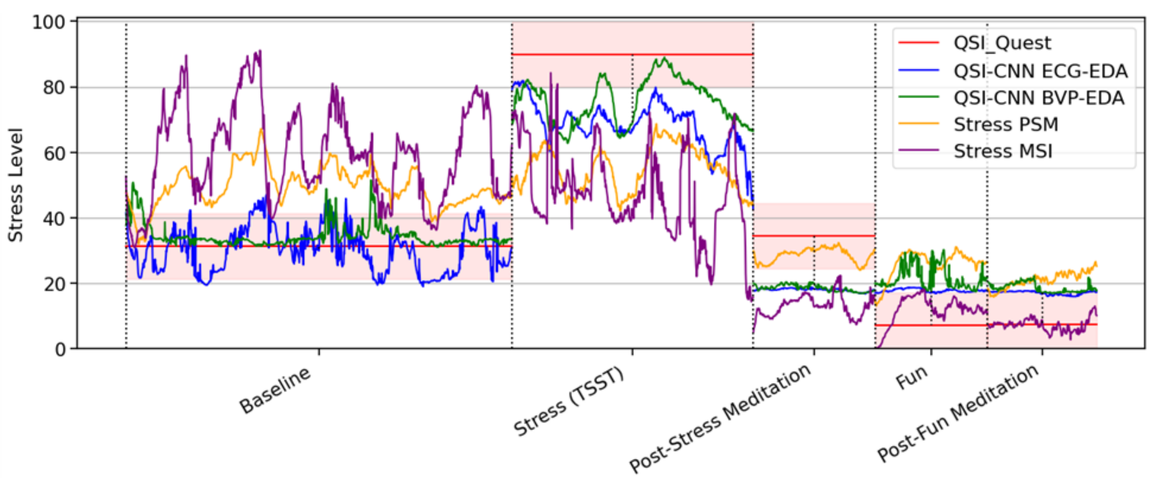
The Pearson correlation results show that both BVP-EDA and ECG-EDA QSI-CNN models exhibit high correlations with the established indices, indicating that they capture similar overall stress trends. The BVP-EDA model exhibits high correlations with the PSM and MSI indices, at 0.82 and 0.75, respectively, while the ECG-EDA model displays strong correlations of 0.79 and 0.81 with the same indices. These results suggest that both modalities effectively reflect the underlying stress dynamics.

Figure 4.13 illustrates the temporal alignment of QSI-CNN models against established indices (MSI and PSM) for two representative subjects. The MSI scores, originally ranging from 0 to 0.448, were rescaled to a 0–100 range. In 4.13a, all models show consistent trends across the task sessions, with distinguishable patterns reflecting baseline, stress, and recovery phases. This suggests that in some individuals, established linear indices may capture similar stress dynamics as the proposed QSI-CNN, reinforcing the model's validity. However, as shown in 4.13b, MSI and PSM exhibit reduced discriminative power between baseline and stress sessions, and often show erratic fluctuations or unrealistic spikes within the same session. These irregularities hinder their interpretability and compromise session-level differentiation. In contrast, QSI-CNN models provide smoother, more stable predictions that better reflect the structure of each session, offering improved robustness to subject variability and demonstrating greater consistency across different experimental phases.

Figures 4.14a and 4.14b compare the classification performance of QSI-CNN models (ECG-EDA and BVP-EDA) with established linear indices (MSI and PSM) in distinguishing stress from fun and baseline sessions. In the stress versus fun task (Figure 4.14a), all models perform comparably, with QSI-CNN models showing a

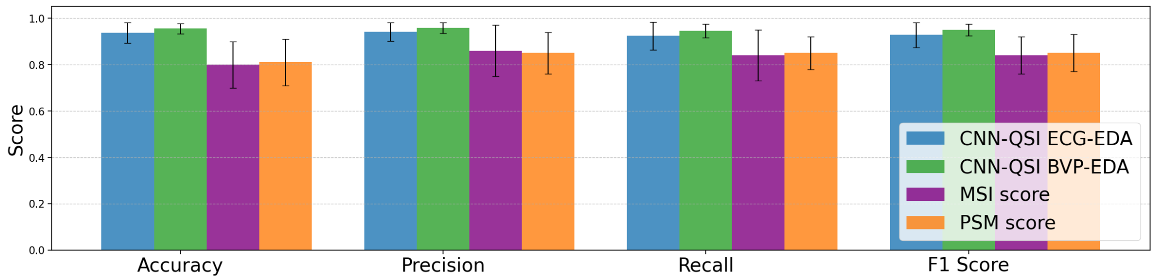


(a)

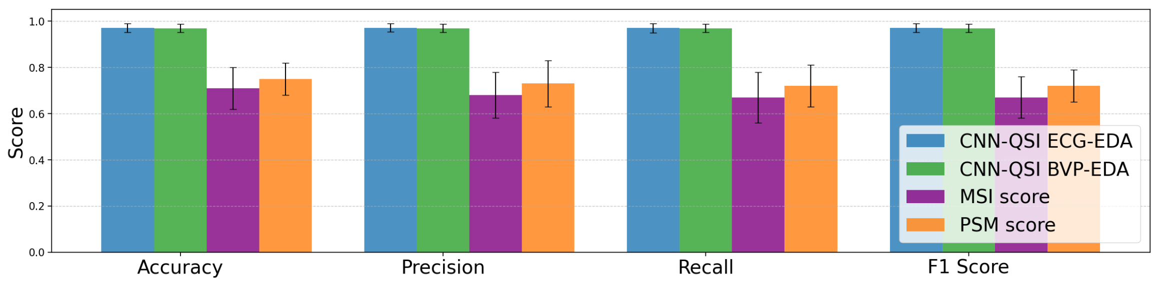


(b)

Figure 4.13: Comparison of QSI-CNN and established stress indices (MSI and PSM) for Subjects 14 and 10. (a) All models follow similar trends with clear session differences. (b) QSI-CNN shows clearer session separation, while MSI and PSM are more variable and less distinct.



(a) Stress vs. fun sessions



(b) Stress vs. baseline sessions

Figure 4.14: Classification performance comparison between QSI-CNN models (ECG-EDA and BVP-EDA) and established indices (MSI and PSM) in distinguishing stress states: (a) stress vs. fun sessions, (b) stress vs. baseline sessions.

slight advantage. This suggests that both approaches effectively capture the pronounced contrast between high- and low-arousal states. However, in the stress versus baseline task (Figure 4.14b), QSI-CNN models significantly outperform MSI and PSM across all metrics. The linear indices exhibit substantial drops in recall and F1 scores, indicating difficulty in identifying stress under more subtle conditions. In contrast, QSI-CNN models maintain high precision, recall, and F1 scores, demonstrating greater sensitivity and robustness. These results underscore the limitations of fixed-form linear indices in detecting stress under minimal contrast and confirm the superior discriminative capability of the proposed deep learning QSI-CNN models, particularly in more challenging classification scenarios.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

The proposed framework for estimating a continuous quantitative stress index (QSI) was developed and validated through a multi-stage process integrating psychometric and physiological data.

A quantitative stress reference score, questionnaire-based quantitative stress index ($\text{QSI}_{\text{Quest}}$), was constructed using exploratory factor analysis on self-report questionnaires, capturing four latent dimensions of acute stress: emotional activation, negative emotionality, emotional tension, and internalized distress. This factor structure offered a more refined and psychometrically sound representation of stress than traditional sum scores. Statistical validation confirmed its discriminative power, showing significant differences across stress, baseline, and fun sessions. Additionally, $\text{QSI}_{\text{Quest}}$ achieved a higher F-statistic compared to other questionnaire-based methods, reflecting greater between-session variance relative to within-session variance and indicating improved sensitivity to stress states. This index serves as a robust ground truth for the supervised physiological models.

Empirical evaluation of blood volume pulse (BVP) preprocessing pipelines revealed that the combination of detrending, Band-Pass Filter, and the Hilbert transform resulted in a notable reduction in NN interval estimation error compared to references derived from electrocardiogram (ECG), demonstrating consistent performance across various datasets. This approach enhanced the accuracy of HRV parameter extraction from BVP signals by reducing noise, thereby facilitating the development of a more reliable stress estimation model.

Following preprocessing, feature extraction, and selection, physiological signals, ECG, BVP, and electrodermal activity (EDA) were used to train both classification and regression models for stress estimation. The proposed deep learning models, fully connected quantitative stress index model (QSI-MLP) and convolutional quantitative stress index model (QSI-CNN), consistently outperformed traditional machine learning baselines. QSI-CNN, in particular, demonstrated superior performance due to its ability to capture temporal and nonlinear dynamics in physiological data, showing improvements in both classification and continuous regression tasks.

The framework was evaluated across multiple sensor modalities, revealing that wrist-based signals (BVP and EDA) performed comparably to chest-worn ECG, especially when used in fused configurations. Notably, the performance gap between ECG-EDA and BVP-EDA was minimal. Given that ECG-EDA requires two separate wearables, while BVP-EDA can be acquired from a single wrist-worn device, the latter presents a more practical and less obtrusive solution. The comparable results suggest that the additional benefit provided by ECG is insufficient to justify the added wearability burden. With suitable preprocessing, the BVP-EDA combination offers near-equivalent performance with greater real-world applicability.

Multimodal fusion led to improved prediction accuracy and reduced error in quantitative evaluations. Qualitative analyses further demonstrated that combining modalities mitigates signal artifacts and noise, such as spikes in individual signals, resulting in smoother and more realistic stress predictions. The integration of an auxiliary dataset also enhanced generalizability across subjects, highlighting the potential for further improving wrist-only configurations.

Qualitative findings highlight the benefits of integrating both psychometric and physiological data within the proposed framework. The $\mathrm{QSI}_{\mathrm{Quest}}$, derived from validated self-report questionnaires, serves as a personalized ground truth that enables the model to learn subject-specific stress baselines and accurately distinguish between different stress states. Concurrently, physiological signals introduce intra-session variability, reflecting the natural fluctuations in stress responses over time, resulting in more realistic, personalized, and temporally sensitive stress estimations.

Comparative evaluation against established linear stress indices (MSI and PSM) further validated the proposed framework. QSI-CNN models demonstrated higher temporal consistency, stronger session-level discriminability, and improved classification performance, particularly in distinguishing stress from baseline sessions, a task where linear indices often failed to capture subtle physiological variations. These findings reinforce the advantage of deep learning models in modeling complex stress dynamics.

5.2 Future Work

Future research can build upon the proposed framework through the following directions:

- **Development of standardized datasets:** There is a critical need for publicly available datasets that combine continuous, standardized stress questionnaires with multimodal physiological signals. Similar to the WESAD dataset, but with a larger number of participants and greater variability, such datasets would support reproducibility, enhance model benchmarking, and improve generalizability, particularly for wrist-worn configurations.

- **Extension with biochemical markers:** Incorporating biochemical indicators such as cortisol or salivary alpha-amylase presents a promising direction for enriching the current psychophysiological model. Adding this third dimension could further strengthen the Quantitative Stress Index by capturing hormonal stress responses, potentially improving its accuracy and biological interpretability.

- **Incorporation of personal history and context:** Including personalized information such as baseline stress levels, prior sessions, or lifestyle factors could improve individual model calibration and enable more accurate and context-aware stress predictions.

- **Real-time implementation:** Deploying the proposed models on embedded or mobile platforms would enable real-time, on-device stress monitoring for practical use in personal health, workplace applications, and adaptive telehealth systems.

Bibliography

[1] S. A. Motzer and V. Hertig, “Stress, stress response, and health,” *Nursing Clinics*, vol. 39, no. 1, pp. 1–17, 2004.

[2] S. M. Smith and W. W. Vale, “The role of the hypothalamic-pituitary-adrenal axis in neuroendocrine responses to stress,” *Dialogues in clinical neuroscience*, vol. 8, pp. 383–395, 2006.

[3] A. Mariotti, “The effects of chronic stress on health: new insights into the molecular mechanisms of brain–body communication,” 2015.

[4] American Psychological Association, “Stress in america 2023: A nation recovering from collective trauma,” https://www.apa.org/news/press/releases/stress/2023/collective-trauma-recovery, 2023, accessed: 2025-04-08.

[5] World Health Organization, *Mental health at work: policy brief.* Genève, Switzerland: World Health Organization, sep 2022.

[6] R. C. MacCallum, S. Zhang, K. J. Preacher, and D. D. Rucker, “On the practice of dichotomization of quantitative variables.” *Psychological methods*, vol. 7, no. 1, 2002.

[7] I. M. Amer, S. M. A. Oteafy, and H. S. Hassanein, "Affective communication of sensorimotor emotion synthesis over urllc," in *IEEE 48th Conference on Local Computer Networks (LCN)*, 2023.

[8] A. M. Colman, *A Dictionary of Psychology.* Oxford University Press, 2015.

[9] M. G. Ziegler, "Psychological stress and the autonomic nervous system," in *Primer on the autonomic nervous system*, 2012.

[10] B. M. Kudielka, D. H. Hellhammer, and C. Kirschbaum, "Ten years of research with the trier social stress test (tsst)-revisited," *Social neuroscience: Integrating biological and psychological explanations of social behavior*, pp. 56–83, 2007.

[11] P. A. Nepomnaschy, T. C. Lee, L. Zeng, and C. Dean, "Who is stressed? comparing cortisol levels between individuals," *American Journal of Human Biology*, vol. 24, no. 4, pp. 515–525, 2012.

[12] N. Ali and U. M. Nater, "Salivary alpha-amylase as a biomarker of stress in behavioral medicine," *International journal of behavioral medicine*, vol. 27, pp. 337–342, 2020.

[13] H.-G. Kim, E.-J. Cheon, D.-S. Bai, Y. H. Lee, and B.-H. Koo, "Stress and heart rate variability: a meta-analysis and review of the literature," *Psychiatry investigation*, vol. 15, no. 3, p. 235, 2018.

[14] C. Ko and S. Oteafy, "Integrative sensing techniques for building non-invasive body sensor networks," in *ICC IEEE International Conference on Communications*, 2023, pp. 6187–6192.

[15] K. M. Dalmeida and G. L. Masala, "Hrv features as viable physiological markers for stress detection using wearable devices," *Sensors*, vol. 21, no. 8, p. 2873, 2021.

[16] Z. Li, H. Snieder, S. Su, X. Ding, J. F. Thayer, F. A. Treiber, and X. Wang, "A longitudinal study in youth of heart rate variability at rest and in response to stress," *International Journal of Psychophysiology*, vol. 73, no. 3, pp. 212–217, 2009.

[17] M. Diarra, J. Theurel, and B. Paty, "Systematic review of neurophysiological assessment techniques and metrics for mental workload evaluation in real-world settings," *Frontiers in Neuroergonomics*, vol. 6, p. 1584736, 2025.

[18] T. Pereira, P. R. Almeida, J. P. Cunha, and A. Aguiar, "Heart rate variability metrics for fine-grained stress level assessment," *Computer methods and programs in biomedicine*, vol. 148, pp. 71–80, 2017.

[19] J.-C. Roy, W. Boucsein, D. C. Fowles, and J. Gruzelier, *Progress in electrodermal research*. Springer Science & Business Media, 2012, vol. 249.

[20] A. Greco, G. Valenza, J. Lazaro, J. M. Garzon-Rey, J. Aguilo, C. de la Cámara, R. Bailón, and E. P. Scilingo, "Acute stress state classification based on electrodermal activity modeling," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 788–799, 2021.

[21] G. Giannakakis, D. Grigoriadis, K. Giannakaki, O. Simantiraki, A. Roniotis, and M. Tsiknakis, "Review on psychological stress detection using biosignals," *IEEE Transactions on Affective Computing*, vol. 13, no. 1, pp. 440–460, 2022.

[22] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in *Proceedings of the 20th ACM international conference on multimodal interaction*, 2018, pp. 400–408.

[23] M. Jaiswal and C.-P. Bara, "Muse: a multimodal dataset of stressed emotion," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020.

[24] A. Arza, J. M. Garzón-Rey, J. Lázaro, E. Gil, R. Lopez-Anton, C. de la Camara, P. Laguna, R. Bailon, and J. Aguiló, "Measuring acute stress response through physiological signals: towards a quantitative assessment of stress," *Medical & biological engineering & computing*, vol. 57, pp. 271–287, 2019.

[25] L. Petrakova, K. Boy, L. Mittmann, L. Möller, H. Engler, and M. Schedlowski, "Salivary alpha-amylase and noradrenaline responses to corticotropin-releasing hormone administration in humans," *Biological Psychology*, vol. 127, pp. 34–39, 2017.

[26] W. Wu, S. Pirbhulal, H. Zhang, and S. C. Mukhopadhyay, "Quantitative assessment for self-tracking of acute stress based on triangulation principle in a wearable sensor system," *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, pp. 703–713, 2018.

[27] E. Garcia Pages, A. Arza, J. Lazaro, C. Puig, T. Castro, M. Ottaviano, M. T. Arredondo, M. L. Bernal, R. López-Antón, C. D. L. Camara *et al.*, "Psychosomatic response to acute emotional stress in healthy students," *Frontiers in physiology*, vol. 13, 2023.

[28] A. A. Schuurmans, P. De Looff, K. S. Nijhof, C. Rosada, R. H. Scholte, A. Popma, and R. Otten, "Validity of the empatica e4 wristband to measure heart rate variability (hrv) parameters: A comparison to electrocardiography (ecg)," *Journal of medical systems*, vol. 44, 2020.

[29] J. P. C. J. Pegueros and J. Rodríguez-Arce, "Physiological datasets in stress and anxiety research: A systematic review," *Biomedical Signal Processing and Control*, vol. 108, p. 107928, 2025.

[30] L. Zhu, P. Spachos, and S. Gregori, "Multimodal physiological signals and machine learning for stress detection by wearable devices," in *IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2022.

[31] M. A. Fauzi and B. Yang, "Continuous stress detection of hospital staff using smartwatch sensors and classifier ensemble," in *pHealth 2021*. IOS Press, 2021, pp. 245–250.

[32] D. S. Benita, A. S. Ebenezer, L. Susmitha, M. Subathra, and S. J. Priya, "Stress detection using cnn on the wesad dataset," in *2024 International Conference on Emerging Systems and Intelligent Computing (ESIC)*, 2024, pp. 308–313.

[33] M. S. Alshareef, B. Alturki, and M. Jaber, "A transformer-based model for effective and exportable iomt-based stress detection," in *GLOBECOM IEEE Global Communications Conference*, 2022, pp. 1158–1163.

[34] N. Chalabianloo, Y. S. Can, M. Umair, C. Sas, and C. Ersoy, "Application level performance evaluation of wearable devices for stress classification with explainable ai," *Pervasive and Mobile Computing*, vol. 87, p. 101703, 2022.

[35] M. Zanetti, T. Mizumoto, L. Faes, A. Fornaser, M. De Cecco, L. Maule, M. Valente, and G. Nollo, "Multilevel assessment of mental stress via network physiology paradigm using consumer wearable devices," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 4409–4418, 2021.

[36] S. Betti, R. M. Lova, E. Rovini, G. Acerbi, L. Santarelli, M. Cabiati, S. D. Ry, and F. Cavallo, "Evaluation of an integrated system of wearable physiological sensors for stress monitoring in working environments by using biological markers," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 8, pp. 1748–1758, 2018.

[37] B. M. Booth, H. Vrzakova, S. M. Mattingly, G. J. Martinez, L. Faust, and S. K. D'Mello, "Toward robust stress prediction in the age of wearables: Modeling perceived stress in a longitudinal study with information workers," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 2201–2217, 2022.

[38] A. P. Allen, P. J. Kennedy, S. Dockray, J. F. Cryan, T. G. Dinan, and G. Clarke, "The trier social stress test: principles and practice," *Neurobiology of stress*, vol. 6, pp. 113–126, 2017.

[39] T. Vaessen, A. Rintala, N. Otsabryk, W. Viechtbauer, M. Wampers, S. Claes, and I. Myin-Germeys, "The association between self-reported stress and cardiovascular measures in daily life: A systematic review," *PloS one*, vol. 16, no. 11, 2021.

[40] A. Diamantopoulos, M. Sarstedt, C. Fuchs, P. Wilczynski, and S. Kaiser, "Guidelines for choosing between multi-item and single-item scales for construct measurement: a predictive validity perspective," *Journal of the Academy of Marketing Science*, vol. 40, pp. 434–449, 2012.

[41] D. McNeish and M. G. Wolf, "Thinking twice about sum scores," *Behavior research methods*, vol. 52, pp. 2287–2305, 2020.

[42] L. R. Fabrigar and D. T. Wegener, *Exploratory factor analysis.* Oxford University Press, 2012.

[43] D. N. Lawley, "The estimation of factor loadings by the method of maximum likelihood," *Proceedings of the Royal Society of Edinburgh*, vol. 60, no. 1, pp. 64–82, 1940.

[44] H. Chaptoukaev, V. Strizhkova, M. Panariello, B. Dalpaos, A. Reka, V. Manera, S. Thümmler, E. Ismailova, M. Todisco, M. A. Zuluaga *et al.*, "Stressid: a multimodal dataset for stress identification," *Advances in Neural Information Processing Systems*, vol. 36, pp. 29 798–29 811, 2023.

[45] V. Markova, T. Ganchev, and K. Kalinkov, "Clas: A database for cognitive load, affect and stress recognition," in *2019 International Conference on Biomedical Innovations and Applications (BIA).* IEEE, 2019, pp. 1–4.

[46] R. M. Sabour, Y. Benezeth, P. De Oliveira, J. Chappe, and F. Yang, "Ubfc-phys: A multimodal database for psychophysiological studies of social stress," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 622–636, 2021.

[47] M. Kalanadhabhatta, C. Min, A. Montanari, and F. Kawsar, "Fatigueset: A multi-modal dataset for modeling mental fatigue and fatigability," in *International Conference on Pervasive Computing Technologies for Healthcare*, 2021, pp. 204–217.

[48] M. Tavakol and R. Dennick, “Making sense of cronbach’s alpha,” *International journal of medical education*, vol. 2, p. 53, 2011.

[49] N. Shrestha, “Factor analysis as a tool for survey analysis,” *American journal of Applied Mathematics and statistics*, vol. 9, no. 1, 2021.

[50] L. L. Thurstone, “An analytical method for simple structure,” *Psychometrika*, vol. 19, no. 3, pp. 173–182, 1954.

[51] D. S. Moore, G. P. McCabe, and B. A. Craig, *Introduction to the Practice of Statistics*. WH Freeman New York, 2009, vol. 4.

[52] P. Bota, R. Silva, C. Carreiras, A. Fred, and H. P. da Silva, “Biosppy: A python toolbox for physiological signal processing,” *SoftwareX*, vol. 26, 2024.

[53] M. Elgendi, M. Jonkman, and F. De Boer, “Frequency bands effects on qrs detection.” *Biosignals*, vol. 2003, p. 2002, 2010.

[54] J. A. Lipponen and M. P. Tarvainen, “A robust algorithm for heart rate variability time series artefact correction using novel beat classification,” *Journal of medical engineering & technology*, vol. 43, no. 3, pp. 173–181, 2019.

[55] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel, and S. H. A. Chen, “NeuroKit2: A python toolbox for neurophysiological signal processing,” *Behavior Research Methods*, vol. 53, no. 4, pp. 1689–1696, feb 2021.

[56] L. Dall’Olio, N. Curti, D. Remondini, Y. Safi Harb, F. W. Asselbergs, G. Castellani, and H.-W. Uh, “Prediction of vascular aging based on smartphone acquired ppg signals,” *Scientific reports*, vol. 10, no. 1, p. 19756, 2020.

[57] S. K. Bashar, D. Han, A. Soni, D. D. McManus, and K. H. Chon, "Developing a novel noise artifact detection algorithm for smartphone ppg signals: Preliminary results," in *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, 2018, pp. 79–82.

[58] S. Hanyu and C. Xiaohui, "Motion artifact detection and reduction in ppg signals based on statistics analysis," in *2017 29th Chinese control and decision conference (CCDC)*, 2017, pp. 3114–3119.

[59] Q. Wang, P. Yang, and Y. Zhang, "Artifact reduction based on empirical mode decomposition (emd) in photoplethysmography for pulse rate detection," in *2010 annual international conference of the IEEE engineering in medicine and biology*. IEEE, 2010, pp. 959–962.

[60] X. Sun, P. Yang, Y. Li, Z. Gao, and Y.-T. Zhang, "Robust heart beat detection from photoplethysmography interlaced with motion artifacts based on empirical mode decomposition," in *Proceedings of 2012 IEEE-EMBS International Conference on Biomedical and Health Informatics*. IEEE, 2012, pp. 775–778.

[61] A. Reiss, I. Indlekofer, P. Schmidt, and K. Van Laerhoven, "Deep ppg: Large-scale heart rate estimation with convolutional neural networks," *Sensors*, vol. 19, no. 14, 2019.

[62] M. Elgendi, I. Norton, M. Brearley, D. Abbott, and D. Schuurmans, "Systolic peak detection in acceleration photoplethysmograms measured from emergency responders in tropical conditions," *PloS one*, vol. 8, no. 10, 2013.

[63] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, "cvxeda: A convex optimization approach to electrodermal activity processing," *IEEE transactions on biomedical engineering*, vol. 63, no. 4, 2015.

[64] H. J. Baek, C.-H. Cho, J. Cho, and J.-M. Woo, "Reliability of ultra-short-term analysis as a surrogate of standard 5-min analysis of heart rate variability," *Telemedicine and e-Health*, vol. 21, no. 5, pp. 404–414, 2015.

[65] J. W. Kim, H. S. Seok, and H. Shin, "Is ultra-short-term heart rate variability valid in non-static conditions?" *Frontiers in physiology*, vol. 12, 2021.

[66] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[67] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[68] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[69] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, pp. 3–42, 2006.

[70] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, pp. 199–222, 2004.

[71] T. M. Oshiro, P. S. Perez, and J. A. Baranauskas, "How many trees in a random forest?" in *Machine Learning and Data Mining in Pattern Recognition: 8th International Conference, MLDM 2012, Berlin, Germany, July 13-20, 2012. Proceedings 8.* Springer, 2012, pp. 154–168.

[72] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, p. 21, 2013.

[73] A. Liapis, C. Katsanos, D. G. Sotiropoulos, N. Karousos, and M. Xenos, "Stress in interactive applications: analysis of the valence-arousal space based on physiological signals and self-reported data," *Multimedia Tools and Applications*, vol. 76, pp. 5051–5071, 2017.

[74] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

Appendix A

BVP Preprocessing Validation

To assess the generalizability and robustness of the findings obtained from the WESAD dataset, the preprocessing pipelines were also applied to the PPG-DaLiA dataset [61]. The goal was to evaluate whether the same signal processing strategies, particularly the combination of detrending, Band-Pass Filter (BPF), and Hilbert transform, which performed best on WESAD, would remain effective across different datasets with varying recording conditions and participant profiles. As in the primary analysis, NN intervals derived from the blood volume pulse (BVP) signal were compared against reference ECG-derived intervals, with mean absolute error (MAE) used as the evaluation metric.

Table A.1 presents the MAE and standard deviation for NN interval estimation across different BVP preprocessing pipelines on the PPG-DaLiA dataset, serving as validation for the results observed in the WESAD dataset. Notably, the D–F pipeline (0.035 ± 0.028) outperformed D–H (0.038 ± 0.027) and performed comparably to D–F–H, suggesting that the combination of detrending and BPF is particularly effective for this dataset. This contrasts with the WESAD results, despite both datasets

Table A.1: Mean ± standard deviation of MAE for each preprocessing pipeline for PPG-DaLiA Dataset.

Pipeline	**MAE (Mean ± SD)**
Raw BVP	0.095 ± 0.039
Detrend (D)	0.041 ± 0.028
Hilbert (H)	0.081 ± 0.033
Detrend-BPF (D-F)	0.035 ± 0.028
Detrend-Hilbert (D-H)	0.038 ± 0.027
Hilbert-EMD (H-E)	0.079 ± 0.039
Detrend-BPF-Hilbert (D-F-H)	**0.034 ± 0.029**
Detrend-BPF-Hilbert-EMD (D-F-H-E)	0.035 ± 0.043

using the same sensor and wearable device, highlighting that differences in experimental settings can influence noise properties and impact the effectiveness of specific preprocessing methods. Despite these differences, the D–F–H pipeline achieved the lowest error (0.034 ± 0.029), confirming its robustness across datasets.

These findings underscore that the contribution of filtering and the Hilbert transform may vary depending on signal conditions. Nevertheless, the full D–F–H pipeline remained consistently effective across both WESAD and PPG-DaLiA, reinforcing its suitability as a general-purpose preprocessing strategy for improving the accuracy of heart rate variability (HRV) parameter estimation from BVP signals. Pipelines incorporating empirical mode decomposition (EMD) introduced greater variability without notable performance gains, supporting the conclusion that EMD does not offer consistent benefits in this context.