POLITECNICO DI TORINO

Master Degree in *Management Engineering:*
ICT & Data Analytics for Management

Master Thesis

Machine Learning for Stress Detection based on Wearable Sensor Data

Supervisor
Dr. Luigi Borzì

Candidate
Matteo Calza-Metre

Academic Year 2024/2025

Abstract

This thesis investigates automatic stress recognition from non-invasive, wearable physiological signals using two complementary paradigms: a *feature–driven* (FD) pipeline based on handcrafted descriptors and classical machine learning, and a *data–driven* (DD) one-dimensional convolutional neural network (1D-CNN) trained end-to-end. Multimodal data include electrodermal activity (EDA), photoplethysmography/BVP (PPG), Accelerometry (ACC), and skin temperature (TEMP). A unified preprocessing and windowing scheme is adopted; the FD branch extracts time-, frequency-, and non-linear features (e.g., heart rate variability indices, Electrodermal Activity (EDA) tonic/phasic markers), followed by model selection among tree ensembles, linear models, and kernels. The DD branch employs a compact convolutional architecture with modality-specific streams and late fusion.

Robustness is assessed through an *in-domain* evaluation and a comprehensive *cross-dataset* protocol spanning four corpora with heterogeneous elicitation protocols and sensors (WESAD, CAMPANELLA, VerBIO, AffectiveROAD). We analyze two operational settings: *zero-shot cross-test* (train on a source dataset and test on a disjoint target without adaptation) and *transfer learning* (fine-tuning on limited target data). Macro-F1 is the primary metric due to label imbalance and deployment relevance.

Empirically, the FD pipeline offers stronger *zero-shot* generalisation: considering the best off-diagonal source for each target, FD outperforms DD on all targets (e.g., WESAD: 0.7192 vs. 0.5370; CAMPANELLA: 0.7675 vs. 0.5875; VerBIO: 0.5852 vs. 0.4941; AffectiveROAD: 0.6678 vs. 0.2787). When light *adaptation* is allowed, the picture reverses: fine-tuning markedly boosts the CNN, surpassing the FD cross-test maxima on three of four targets (up to 0.8030 on WESAD; 0.7713 on CAMPANELLA; 0.6458 on VerBIO), while AffectiveROAD remains marginally in favor of FD (0.6678 vs. 0.6550). In-domain baselines confirm the complementarity of the two families, with CNNs excelling on laboratory-style datasets and FD models remaining competitive where structured, high-SNR descriptors suffice.

Taken together, these results suggest a practical deployment guideline: in *no-label* target scenarios, FD is the safer choice; when even modest target supervision is feasible, *adapted* DD models become preferable. The observed off-diagonal performance gap is traced to cross-corpus divergences in labeling schemes, protocol stressors, device stacks, and temporal priors. Building on these findings, the thesis outlines a roadmap to bridge the gap—self-supervised pretraining on heterogeneous wearables, domain generalisation without target labels, test-time/source-free adaptation, personalised/federated updates for subject idiosyncrasies, and probabilistic label fusion for ecological validity—aimed at reliable stress sensing beyond controlled settings.

Contents

List of Tables

List of Figures

List of Acronyms

ACC	Accelerometry
ACTH	Adrenocorticotropic Hormone
AI	Artificial Intelligence
ANN	Artificial Neural Networks
ANOVA	Analysis of Variance
ANS	Autonomic Nervous System
AUC	Area Under the Curve
BPM	Beats Per Minute
BR	Breath Rate
BP	Blood Pressure
BVP	Blood Volume Pulse
CNN	Convolutional Neural Network
CPT	Cold Pressor Test
DL	Deep Learning
EDA	Electrodermal Activity
ECG	Electrocardiogram
EEG	Electroencephalogram
EMG	Electromyography
FP	False Positive

FFT	Fast Fourier Transform
FIR	Finite Impulse Response
FN	False Negative
FPR	False Positive Rate
FuzzyEn	Fuzzy Entropy
GB	Gradient Boosting
GSR	Galvanic Skin Response
GRU	Gated Recurrent Unit
HF	High Frequency
HR	Heart Rate
HRV	Heart Rate Variability
HPA	Hypothalamic–Pituitary–Adrenal Axis
IBI	Inter-Beat Interval
IIR	Infinite Impulse Respons
IQRNN	Interquartile Range of NN Intervals
LF	Low Frequency
LR	Logistic Regression
LOSO	Leave-One-Subject-Out
LSTM	Long Short-Term Memory
MadNN	Median Absolute Deviation of NN Intervals
MIST	Montreal Imaging Stress Task
ML	Machine Learning
MLP	Multilayer Perceptron
MMST	Mannheim Multicomponent Stress Test
MAST	Maastricht Acute Stress Test

NASA-TLX	NASA Task Load Index
NN	Normal-to-Normal Intervals
PASAT	Paced Auditory Serial Addition Task
PANAS	Positive and Negative Affect Schedule
PCA	Principal Component Analysis
PET	Positron Emission Tomograph
pNN50	Percentage of NN intervals differing by more than 50 ms
PPG	Photoplethysmography
PSS	Perceived Stress Scale
PPT	Pulse Transit Time
RF	Random Forest
RMSSD	root mean square of successive differences
RNN	Recurrent Neural Networks
RR	Interval between successive R-peaks of the ECG
ROC	Receiver Operating Characteristic
SampEn	Sample Entropy
SAM	Sympathetic–Adrenal–Medullary System
SAM Scale	Self-Assessment Manikin
SCL	Skin Conductance Level
SCR	Skin Conductance Response
SCWT	Stroop Color Word Test
SSSQ	Short Stress State Questionnaire
STAI	State-Trait Anxiety Inventory
SVM	Support Vector Machines
SDNN	Standard Deviation of NN Intervals

SKT	Skin Temperature
SNS	Sympathetic Nervous System
SUDS	Subjective Units of Distress Scale
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TPR	True Positive Rate
TSST	Trier Social Stress Test
VFT	Verbal Fluency Task
XGB	eXtreme Gradient Boosting
WESAD	Wearable Stress and Affect Detection Dataset

Chapter 1

Introduction

Stress is broadly defined as a psychophysiological response to internal or external demands that are perceived as exceeding the adaptive resources of an individual. According to Lazarus and Folkman [63], stress arises “when a person perceives the demands of an environmental stimulus to be greater than their ability to meet, mitigate, or alter those demands”. This conceptualization highlights the role of individual perception in modulating the stress experience, distinguishing it from general emotional states such as anxiety or depression.

Stress responses are deeply rooted in homeostasis, the biological imperative to maintain internal stability. As Claude Bernard initially theorized, and later expanded upon by Cannon and Selye, stress represents the disruption of homeostasis and the body’s attempt to restore it. [32] Hans Selye, who pioneered the concept of stress in a medical context, described it as “the non-specific response of the body to any demand for change”, framing it as a general adaptation mechanism.

Stress can be categorized into acute and chronic forms. Acute stress results from identifiable, short-term stressors—like public speaking or job interviews—and typically triggers a well-defined physiological response that is time-limited. Chronic stress, on the other hand, arises from persistent, long-term stressors such as caregiving, financial strain, or discrimination. Chronic exposure to such stressors has been shown to increase systemic inflammation and risk for diseases such as cardiovascular disorders, depression, and immune dysfunction [29].

The physiological mechanisms underlying stress responses primarily involve the Hypothalamic–Pituitary–Adrenal Axis (HPA) and the Sympathetic–Adrenal–Medullary System (SAM). The activation of these systems results in the release of cortisol and catecholamines like adrenaline and noradrenaline, which prepare the body for "fight or flight" responses [32]. Prolonged activation, especially in chronic stress, leads to allostatic load, a state of cumulative biological burden that contributes to disease onset and progression.

Stressors are defined as internal or external stimuli that disrupt an individual’s

homeostasis and elicit a physiological and psychological stress response. These stimuli can be acute or chronic, physical or psychological, and vary significantly in origin and intensity. Stressors are generally classified according to their source (exogenous vs. endogenous) and their nature (physical, environmental, social, cognitive, psychological, chronic or traumatic).
A proposed classification of stressors commonly employed in psychophysiological research, based on their intrinsic nature, is presented in Table 1.1.

Table 1.1: Types of stressors used in psychophysiology research [33]

Stressor Type	**Description and Examples**
Physical	Strenuous physical activity, sleep deprivation, tiredness, painful stimuli, acute injury or medical emergency.
Environmental	Extreme temperature conditions, high humidity, low oxygen or high carbon dioxide levels, high noise, earthquake in the surrounding area.
Mental/task-related	Task demands taxing cognitive capacities, inconsistent reward schedules, conflicting instructions.
Social	Social conflict, criticism, unfair treatment.
Psychological/emotional	Divorce, loss of loved one, intense emotions, mental health conditions.
Chronic	Financial stress, chronic illness, job insecurity, poor living conditions.
Traumatic	Past traumatic experiences intruding into current psycho-emotional state.

Exogenous stressors originate from the external environment. They include physical challenges such as injury or sleep deprivation, environmental conditions like temperature extremes or noise, and mental or task-related demands involving cognitive overload or conflicting instructions. Social situations, such as interpersonal conflict or public speaking, are also potent external stressors, particularly due to their activation of both the sympathetic nervous system and HPA axis.
In contrast, endogenous stressors stem from within the individual. These may involve personality traits, maladaptive thought patterns, or low stress resilience. According to Lazarus and Folkman, stress is transactional and depends on the individual's interpretation of an event and their perceived ability to cope. Thus, factors like perfectionism, emotional instability, or self-criticism can elicit stress responses even in the absence of intense external pressure. [33]

1.1 Biomarkers of Stress

Stress activates a variety of physiological and biochemical processes in the human body, many of which produce measurable biosignals. These biosignals are defined as time-varying indicators generated by biological systems that reflect both immediate and cumulative responses to internal or external stressors. According to Giannakakis et al. [33], biosignals play a fundamental role in psychophysiological stress detection, offering objective insights into autonomic, central nervous, and endocrine system reactivity.

A practical distinction can be made between physiological, physical, and biochemical biosignals. Physiological signals include Heart Rate (HR), HRV, EDA, respiration, Skin Temperature (SKT), and brain activity, all of which reflect the real-time modulation of the Autonomic Nervous System (ANS). Physical signals, such as facial expressions, body posture, head movements, pupil dilation, and eye blinks, provide observable manifestations of stress-related muscular and behavioral responses [58, 48]. These signals, though not always directly linked to internal physiology, offer valuable, non-invasive indicators of stress-induced behavior and have been increasingly used in wearable or ambient systems.

Biochemical signals, on the other hand, include hormone concentrations like cortisol, alpha-amylase, and cytokines. These markers reflect slower systemic responses mediated by HPA axis and immune system and are often used in laboratory or clinical settings [50].

As highlighted by Iqbal et al., physiological signals are better suited for continuous and real-time monitoring, while biochemical markers are more appropriate for baseline assessments and clinical diagnostics. Furthermore, Ladakis and Chouvarda [48] emphasize that combining multiple signal types improves detection robustness and generalizability in stress recognition systems.

The integration of wearable and ambient sensing technologies has enabled non-invasive, real-time acquisition of these signals in both experimental and real-world environments. This has positioned biosignal analysis as a key approach in the development of personalized and scalable stress monitoring solutions. A summary of the most common biosignals used in stress detection, including their measurement modality and physiological origin, is presented in Table 1.2 and will serve as a reference for the following subsections.

1.1.1 Electrocardiography

ECG is a diagnosis tool that reports the electrical activity of the heart recorded by skin electrodes. These electrodes detect the electrical signals generated by cardiac cells during depolarization and repolarization and transmit them to a recording device, creating a continuous time-series signal of the heart's activity. A typical

Table 1.2: Physiological biosignals covered in the following subsections and their physiological origin.

Biosignal	**Physiological origin**
Electrocardiography *(ECG)*	Cardiac electrical activity under autonomic control.
Photoplethysmography / Blood Volume Pulse *(PPG/BVP)*	Peripheral blood volume changes driven by cardiac stroke and vascular tone.
Electrodermal Activity *(EDA)*	Sympathetic sudomotor activity of eccrine sweat glands.
Electroencephalography *(EEG)*	Cortical electrical rhythms reflecting central arousal and cognitive load.
Electromyography *(EMG)*	Surface skeletal muscle activation and tension.
Skin Temperature *(SKT)*	Cutaneous perfusion modulated by thermoregulatory vasoconstriction/vasodilation.
Respiration / Breath Rate *(BR)*	Ventilatory control regulated by autonomic and chemoreflex mechanisms.

ECG tracing of normal heart beat consists of a P wave, Q wave, R wave, S wave, T wave and U wave as shown in Figure 1.1. Specifically, the P wave represents atrial depolarization, the QRS complex corresponds to ventricular depolarization, and the T wave indicates ventricular repolarization [70].

The frequency content of these waves varies according to the speed of electrical conduction in cardiac tissue. For instance, the T wave primarily occupies a band from DC to 10 Hz, the P wave ranges between 5–30 Hz, and the QRS complex spans 8–50 Hz. Abnormalities in ventricular conduction, such as those associated with stress or arrhythmias, can introduce high-frequency components above 70 Hz, often visible as notches on the QRS complex [33]. During psychological stress, the Sympathetic Nervous System (SNS) is activated, increasing both the rate and force of cardiac contractions. This autonomic shift ensures that oxygenated blood reaches vital organs and skeletal muscles more rapidly to support the body's stress response [47]. Hormones such as adrenaline and cortisol further contribute by increasing heart rate, Blood Pressure (BP), and cardiac output. These physiological responses can be captured through ECG alterations, including shortened Interval between successive R-peaks of the ECG (RR) and morphological changes

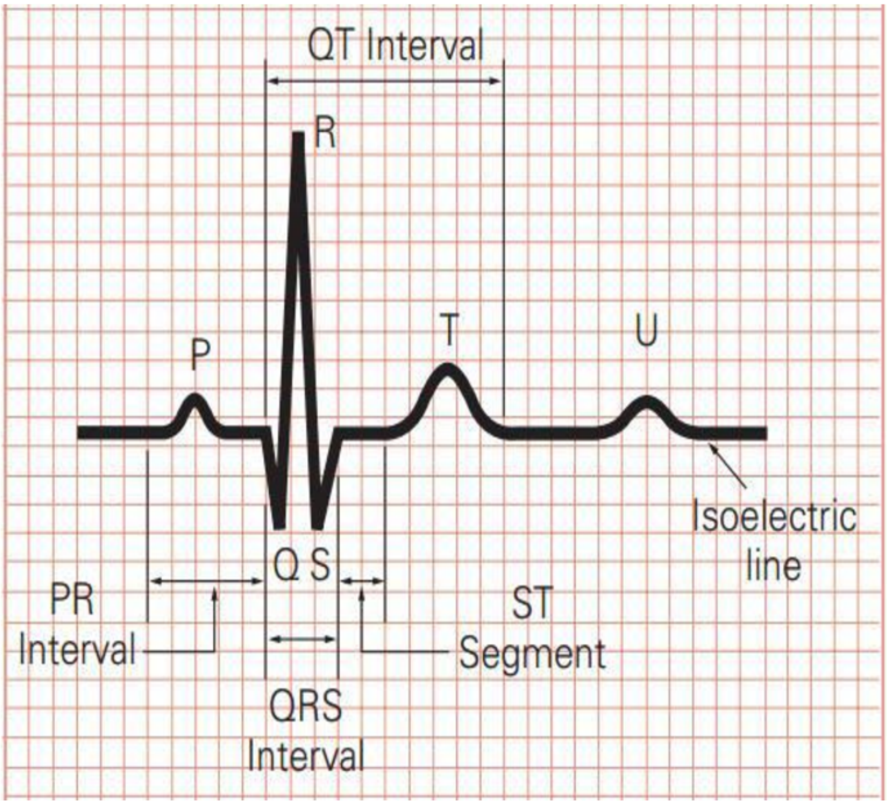


Figure 1.1: Normal ECG signal [85]

in waveforms, such as elevated T-wave amplitudes or shortened QT intervals [45]. Among the metrics derived from ECG, HR is the most widely adopted measure in stress research. HR is defined as the number of Beats Per Minute (BPM) and can be directly calculated from the ECG by measuring the time between successive R peaks, known as the RR interval. The RR interval and HR are inversely related: as heart rate increases, RR intervals shorten. Acute stress typically causes a significant rise in heart rate, which is one of the most immediate and robust indicators of stress in both clinical and experimental settings [3]. Beyond heart rate alone, the variability between successive RR intervals, referred to as HRV, provides deeper insight into the regulation of cardiac activity by the ANS. In HRV analysis, only intervals between normal heartbeats (i.e., those not affected by ectopic beats or artifacts) are considered, and these are referred to as NN intervals (normal-to-normal). HRV can be analyzed using time-domain, frequency-domain, and nonlinear features. Time-domain features such as the Standard Deviation of NN Intervals (SDNN) and the root mean square of successive differences (RMSSD) are commonly reduced during stress. Frequency-domain analysis partitions HRV into low-frequency (LF: 0.04–0.15 Hz) and high-frequency (HF: 0.15–0.4 Hz) components, where High Frequency (HF) is associated with parasympathetic activity and Low Frequency (LF) with a mix of sympathetic and parasympathetic influence. During stress, a decrease in HF power and an increase in the LF/HF ratio are consistently observed, reflecting sympathetic dominance [54]. A concise summary of the ECG-derived features used in this work, grouped by domain and typical modulation under stress, is provided in Table 1.3.

1.1.2 Electroencephalography

Electroencephalogram (EEG) is a non-invasive technique used to measure the brain's electrical activity through electrodes placed on the scalp. These electrodes detect voltage fluctuations caused by ionic current flows in neurons, which are recorded as continuous signals in the time domain. The standard method for electrode placement is the internationally recognized 10-20 system, which ensures consistent positioning over key brain regions such as the frontal, parietal, occipital, and temporal lobes. As shown in Figure 1.2, electrodes are labeled according to their anatomical location: F (frontal), T (temporal), P (parietal), O (occipital), and C (central), with odd numbers representing the left hemisphere and even numbers the right [99].

The raw EEG signal reflects the brain's electrical activity over time, typically sampled at rates between 128 Hz and 1024 Hz and expressed in microvolts (µV). An example of a typical raw EEG signal is shown in Figure 1.3, where the continuous nature of brainwave activity and its oscillatory behavior are clearly visible. This raw signal appears as a complex waveform, but it is actually composed of multiple

Table 1.3: Main ECG-derived features used for stress detection, categorized by domain and reported effect under stress [33].

Feature	**Description**	**Domain**	**Stress Effect**
HR	Heart Rate (BPM)	Time	↑
RR interval	Time between successive R peaks	Time	↓
SDNN	Std. deviation of NN intervals	Time	↓
RMSSD	Root mean square of successive differences	Time	↓
pNN50	% of NN intervals $>$ 50 ms	Time	↓
HF Power	High-frequency power (0.15–0.4 Hz)	Frequency	↓
LF Power	Low-frequency power (0.04–0.15 Hz)	Frequency	↑ / variable
LF/HF Ratio	Sympathovagal balance index	Frequency	↑
SD1	Short-term variability (Poincaré width)	Nonlinear	↓
SD2	Long-term variability (Poincaré length)	Nonlinear	↓
SampEn	Sample Entropy	Nonlinear	↓
ApEn	Approximate Entropy	Nonlinear	↓
DFA α_1	Detrended Fluctuation Analysis (short-term)	Nonlinear	↓ / variable
T-wave amplitude	Height of the T wave	Morphological	↑
QT interval	Duration from Q to T wave	Morphological	↓

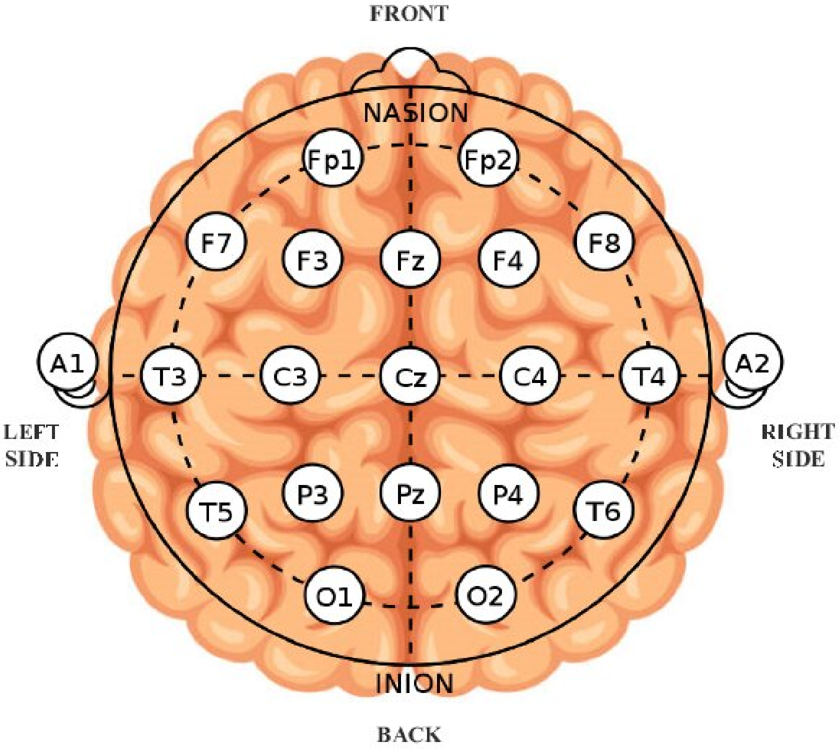


Figure 1.2: Electrode Position on head [71]

overlapping sinusoidal components, each oscillating at a specific frequency. These components can be grouped into five standard frequency bands based on their spectral content, commonly referred to as EEG bands or brainwaves. A clear overview of these bands, along with their corresponding frequencies and associated cognitive states, is provided in Table 1.4. By analyzing the power contained within each band, it is possible to gain insights into the subject's cognitive or physiological state.

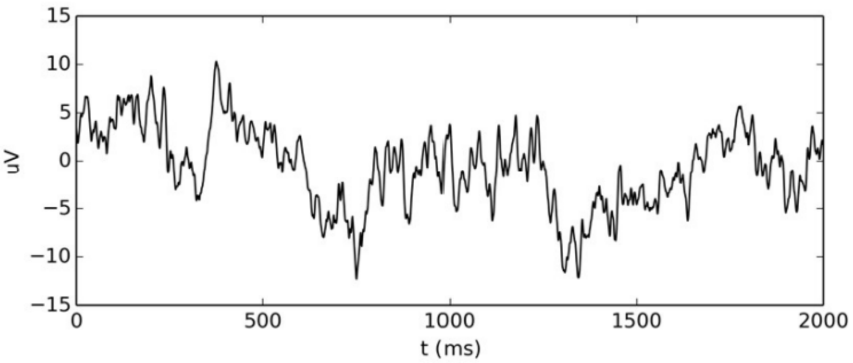


Figure 1.3: Typical Raw EEG Signal

Table 1.4: EEG Frequency Bands and Associated Cognitive States

Band	Frequency (Hz)	Cognitive/Physiological Correlates
Delta	0.5 – 4	Deep sleep, unconscious states
Theta	4 – 8	Drowsiness, light sleep, meditation
Alpha	8 – 13	Relaxed wakefulness, closed eyes
Beta	13 – 30	Active thinking, alertness
Gamma	>30	Higher-order cognitive processes, attention

EEG-based stress assessment relies on extracting discriminative features from specific frequency bands. Studies consistently report increased beta activity and decreased alpha power during stress episodes. Additionally, alpha asymmetry, especially in the frontal lobes, is considered a neurophysiological marker for emotional valence and arousal [6]. Experimental results support the effectiveness of these metrics. As demonstrated by the study conducted by Alonso et al. [6], subjects under stress conditions showed a significant increase in beta power (23–36 Hz) and a reduction in high alpha power (11–12 Hz) across multiple EEG channels. A similar pattern is reported in the study by Malviya et al. [71], where Figure 1.4

and Figure 1.5 illustrate this shift in EEG band power before and after stress exposure across 36 subjects. Frontal alpha asymmetry analysis, in particular, has shown sensitivity to emotional and stress-related brain activity. This asymmetry refers to the difference in alpha power between the left and right frontal regions—typically measured at electrodes F3 and F4. During stress or negative emotional states, studies have observed a relative increase in right frontal activation (reflected by decreased alpha power on the right side) compared to the left, indicating an imbalance in hemispheric processing. A concise summary of the EEG-derived features used in this work, grouped by domain and typical modulation under stress, is provided in Table 1.5.

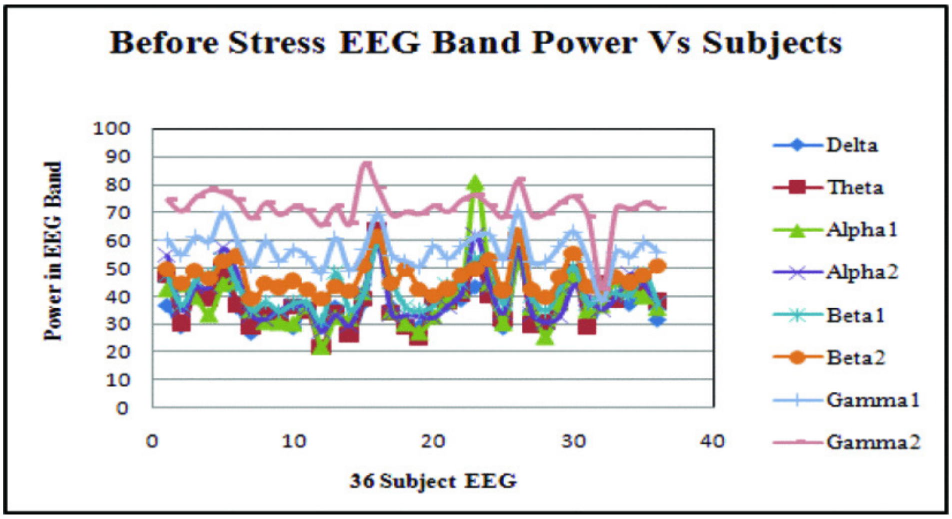


Figure 1.4: Before Stress EEG Band Power Vs Subjects [71]

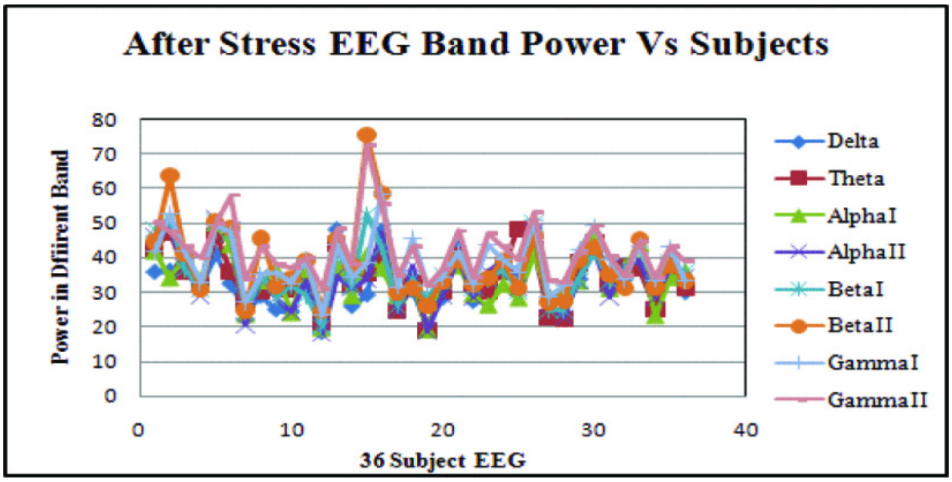


Figure 1.5: After Stress EEG Band Power Vs Subjects [71]

Table 1.5: Main EEG-derived features used for stress detection, categorized by domain and reported effect under stress.

Feature	**Description**	**Domain**	**Stress Effect**
Alpha band power (8–13 Hz)	Band power integrated in the alpha range	Frequency	↓
Beta band power (13–30 Hz)	Band power integrated in the beta range	Frequency	↑
Beta/Alpha ratio (B/A)	Ratio of beta to alpha power	Ratio	↑
Frontal alpha asymmetry (F3–F4)	Log-difference of alpha power (left vs. right frontal sites)	Asymmetry	Right↑ (↓ right)
Spectral / permutation entropy	Frequency-domain irregularity/complexity indices	Complexity	↑
Functional connectivity (coherence)	Inter-channel coupling in selected bands	Connectivity	Var. (↓, ↑)

1.1.3 Electrodermal Activity

EDA, also referred to as Galvanic Skin Response (GSR), measures variations in the electrical conductance of the skin, which are influenced by the activity of the sweat glands. These glands are directly innervated by the SNS, making EDA one of the most sensitive and specific biomarkers for detecting autonomic arousal and stress responses [48]. Unlike other physiological signals, EDA reflects only the sympathetic branch of the ANS, not the parasympathetic one, which enables more targeted monitoring of emotional and psychological arousal [117].

The measurement of EDA is conducted through exosomatic methods, typically using wearable sensors or laboratory-grade electrodes placed on the fingers or palms—regions with a high density of eccrine sweat glands. These sensors apply a constant voltage and measure resulting changes in skin conductance (SC), which typically ranges from 1 to 20 µS (microsiemens). The recorded EDA signal consists of two primary components: the tonic component, or Skin Conductance Level (SCL), and the phasic component, or Skin Conductance Response (SCR). Additionally, non-specific SCRs (NS.SCRs) can occur in the absence of identifiable stimuli, reflecting internal arousal states [116]. Skin Conductance Level (SCL) represents the baseline level of skin conductance over longer periods and reflects general autonomic arousal. Skin Conductance Response (SCR) captures short-term, stimulus-evoked changes in conductance. As shown in Figure 1.6, key SCR features include amplitude, latency, rise time, recovery time, and slope. NS.SCR refers to spontaneous changes in skin conductance without any obvious stimulus, often used as an indicator of general arousal or anxiety.

During stressful situations, the SNS activates eccrine sweat glands, increasing skin hydration and thus electrical conductance. This physiological response leads to elevated SCL values and a higher frequency and amplitude of SCRs, which reflect the body's heightened arousal state. Stress-related stimuli—such as cognitive load, emotional distress, or anticipation of an event—can provoke significant phasic responses, with rapid SCR peaks following the onset of the stressor. The timing and magnitude of these peaks are closely linked to the perceived intensity and immediacy of the stress. Experimental evidence, such as the findings from the BiLoad Test (Stroop and N-Back tasks), confirms that periods of increased mental workload coincide with noticeable surges in SCR amplitude and frequency, along with a sustained elevation in SCL levels [64]. Similar findings were reported by Setz et al. [93], where participants exposed to challenging cognitive tasks exhibited a marked increase in EDA levels. These changes are visually confirmed in Figure 1.7, which shows higher SCR frequency and elevated SCL values during the stress condition Montreal Imaging Stress Task (MIST) compared to baseline.

Beyond the primary components, several derived features are commonly extracted from EDA signals for stress analysis. These features, calculated from SCL,

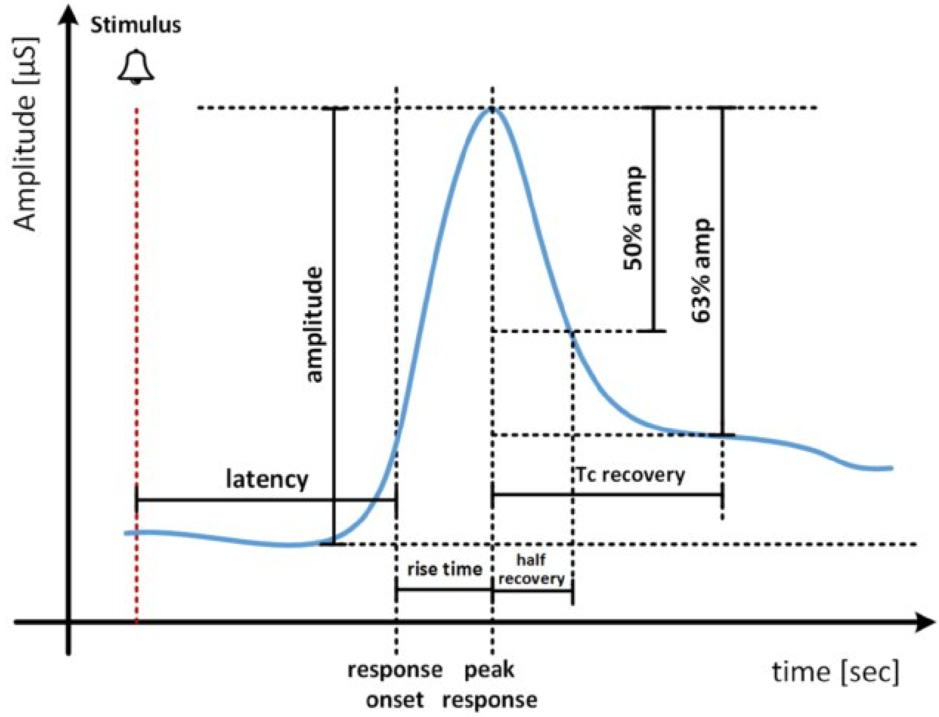


Figure 1.6: A typical skin conductance response variation and common extracted features [33].

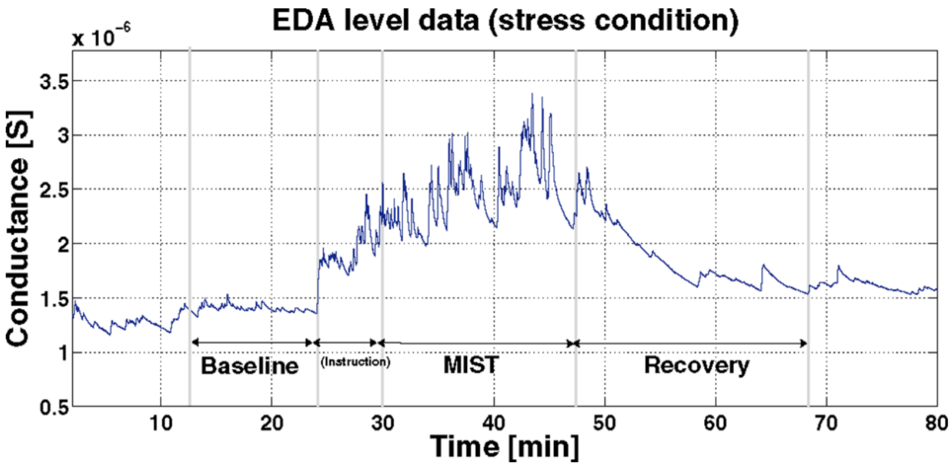


Figure 1.7: Example of EDA signals in baseline and stress conditions [93].

SCR, and NS.SCR, provide quantitative indicators of sympathetic nervous system activity. Table 1.6 summarizes the most widely used features in psychophysiological research.

Table 1.6: Main EDA-derived features used for stress detection, with domain and typical effect under stress.

Feature	**Description**	**Domain**	**Stress Effect**
SCL (Skin Conductance Level)	Baseline conductance over the analysis window	Time (tonic)	↑
SCR amplitude	Height of the stimulus-evoked peak above baseline	Time (phasic)	↑
SCR frequency (nSCR/min)	Number of SCRs per minute (or per window)	Time (phasic)	↑
SCR rise time	Time from onset to peak	Time (phasic)	↓
SCR recovery time	Time to return to 50% of peak	Time (phasic)	↑
SCR slope	Amplitude divided by rise time	Time (phasic)	↑
SCR area (AUC)	Integral of SCR over the window	Time (phasic)	↑

1.1.4 Electromyography

EMG is a widely used technique for recording the electrical activity produced by skeletal muscles during contraction. This electrical signal originates from the action potentials of muscle fibers activated by motor neurons, and is a reflection of both the neural drive and the mechanical activity of the muscles [77]. The signal is captured non-invasively through surface electrodes placed over the skin, typically on large muscles such as the trapezius, which has shown strong correlations with stress-related muscular tension [90].

The raw EMG signal is a time-varying, oscillatory waveform that reflects the summation of multiple motor unit action potentials. It typically appears as a noisy, bipolar signal fluctuating rapidly around zero. The waveform displays spikes of varying amplitudes and durations, which correspond to the recruitment and firing of motor units. Its amplitude, usually in the range of 50–500 µV, reflects the intensity of muscle activation, while its frequency content can span from 10 Hz to over 500 Hz, depending on the task and muscle type [77, 82]. Given the presence of noise and movement artifacts, the raw signal is typically preprocessed using a band-pass filter (commonly between 20 and 450 Hz) to retain relevant physiological information while attenuating movement artifacts and high-frequency noise. In addition, notch filters centered at 50, 100, 150, 200, 250, and 350 Hz are applied to remove powerline interference harmonics [1, 98].

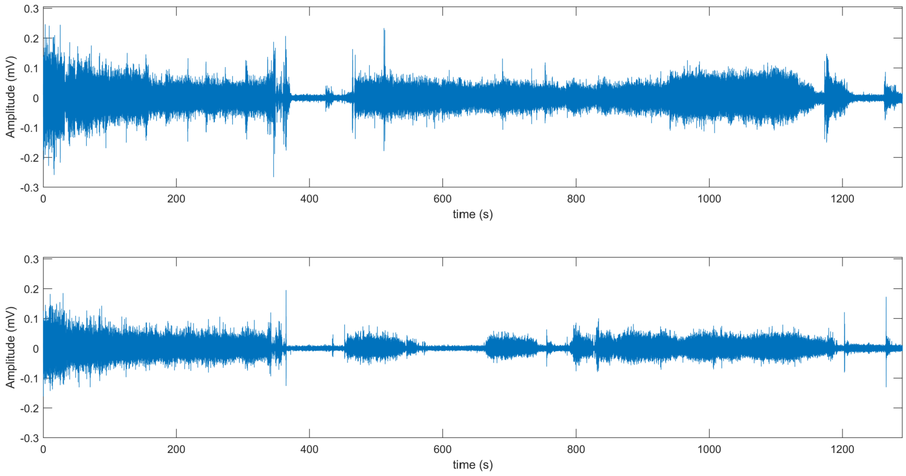


Figure 1.8: Raw EMG Signal (top) and Filtered Signal (Bottom) [21]

Increased muscle activity is a recognized physiological correlate of both emotional and cognitive stress. In particular, surface EMG has been shown to reliably capture the escalation of neuromuscular tension that typically accompanies heightened psychological demand. This is especially apparent in postural muscles such as the trapezius, which are prone to involuntary contraction under mental load. Experimental findings confirm that under stress, EMG signals recorded from the trapezius muscle exhibit increased amplitude, prolonged activation periods, and reduced relaxation intervals, suggesting a state of sustained neuromuscular engagement [21].

To evaluate the effects of emotional or cognitive stress on muscular activity, the raw EMG signal is processed to extract a set of quantitative features that capture relevant physiological changes. These features are generally categorized into two major groups based on their analytical approach: time-domain and frequency-domain measures.

Time-domain features are derived directly from the amplitude and structure of the EMG signal over time. They provide valuable insights into the magnitude, variability, and persistence of muscle activation, capturing the overall intensity and modulation of neuromuscular activity during stress responses.

Frequency-domain features, on the other hand, are extracted by transforming the signal into the frequency domain, typically using the Fast Fourier Transform (FFT). This analysis enables the identification of dominant frequency components and the distribution of spectral power, which are indicative of motor unit recruitment strategies and potential fatigue or stress-induced alterations in neuromuscular dynamics [82].

Table 1.7 provides a summary of the most relevant EMG features commonly used to characterize muscle activation patterns.

Table 1.7: Main EMG-derived features used for stress detection, with domain and typical effect under stress.

Feature	**Description**	**Domain**	**Stress Effect**
RMS (Root Mean Square)	Energy-related magnitude of the rectified EMG	Time	↑
MAV (Mean Absolute Value)	Average of the rectified EMG amplitude	Time	↑
IEMG (Integrated EMG)	Area under the rectified EMG curve	Time	↑
VAR (Variance)	Power of the EMG signal	Time	↑
SSI (Simple Square Integral)	Total energy (sum of squared samples)	Time	↑
WL (Waveform Length)	Sum of absolute sample-to-sample differences	Time	↑
PSD (Power Spectral Density)	Distribution of power over frequency	Frequency	↑ (total)
MF (Median Frequency)	Frequency below which 50% of power is contained	Frequency	Variable

Empirical studies have validated the use of these EMG features for stress detection. For instance, Schleifer et al. [90] reported increased EMG activity and a significant reduction in EMG-gap frequencies during high workload computer tasks, suggesting continuous muscle activation and reduced recovery. Similarly, Luijcks et al. [68] demonstrated that anticipatory stress leads to increased EMG amplitudes in the pre-stimulus phase, with clear differentiation across individual stress reactivity profiles. Moreover, the analysis of both low- and high-frequency components in the EMG spectrum has shown potential for distinguishing between stress and meditative states [1].

These findings underscore the importance of EMG as a robust, non-invasive tool for detecting stress-induced neuromuscular changes. Its ability to reflect both subtle and pronounced variations in muscle activity makes it a valuable signal for real-time cognitive and emotional monitoring in experimental and applied settings.

1.1.5 Photoplethysmography

PPG is a non-invasive optical technique used to detect volumetric changes in blood circulation within the microvascular bed of tissue. It operates by illuminating the skin with a light source—commonly red or near-infrared LEDs—and capturing variations in light intensity reflected or transmitted through the tissue using a photodetector. These variations reflect dynamic changes in blood volume and vessel wall motion [5].

The detected PPG signal consists of two primary components: the pulsatile (AC) component and the tonic (DC) component. The AC component is synchronous with the cardiac cycle, generally centered around 1 Hz, and represents the rhythmic expansion and contraction of arteries due to heartbeat. The DC component is a slowly varying baseline influenced by respiration, vasomotor activity, thermoregulation, and other physiological factors.

A typical PPG waveform includes a steep rising phase called the anacrotic phase, which corresponds to the systolic upstroke resulting from ventricular contraction. This is followed by a falling phase, or catacrotic phase, representing diastole and wave reflections from the peripheral vascular system. A small dip known as the dicrotic notch often appears during the catacrotic phase and indicates the brief backflow of blood following aortic valve closure. The PPG waveform's morphology correlates closely with that of the ECG. In particular, the R-peak of the ECG precedes the systolic peak of the PPG by a brief delay known as the Pulse Transit Time (PPT), which represents the time required for the pressure wave to travel from the heart to the peripheral site of measurement. Figure 1.9 illustrates a representative PPG waveform along with its temporal relationship to the ECG.

PPG-derived features can be grouped into two main categories: time-domain and frequency-domain features. Time-domain features, many of which are graphically illustrated in Figure 1.10, are extracted from individual pulse waves and describe the temporal morphology of the signal. Additional frequency domain features computed from the whole signal are summarized in Table 1.8.

PPG has been widely employed in stress research due to its sensitivity to changes in ANS activity. Under stress, alterations in cardiovascular dynamics can significantly impact the PPG waveform. As shown by Halim et al. [36], a notable increase in the amplitude of the systolic peak has been observed under stress conditions, as summarized in Table 1.9.

1.1.6 Skin Temperature

SKT is a widely recognized physiological indicator and is commonly used in psychophysiological studies due to its strong association with ANS activity. The measurement of SKT is typically performed using surface thermistors or thermocouples applied to peripheral regions such as the fingers, hands, or face. More

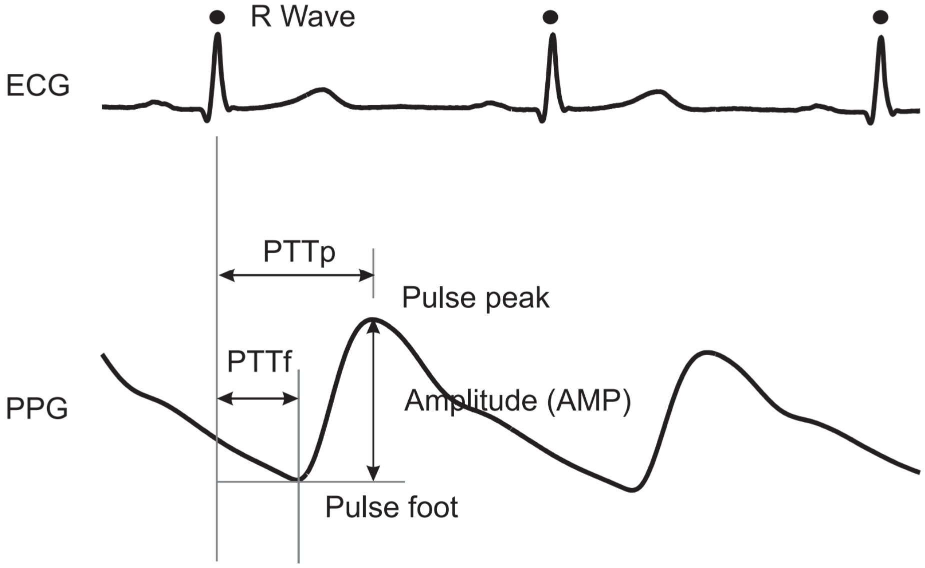


Figure 1.9: Representative PPG waveform and ECG showing the systolic peak, dicrotic notch, and the R-peak to PPG peak delay (Pulse Transit Time – PTT). Adapted from [5].

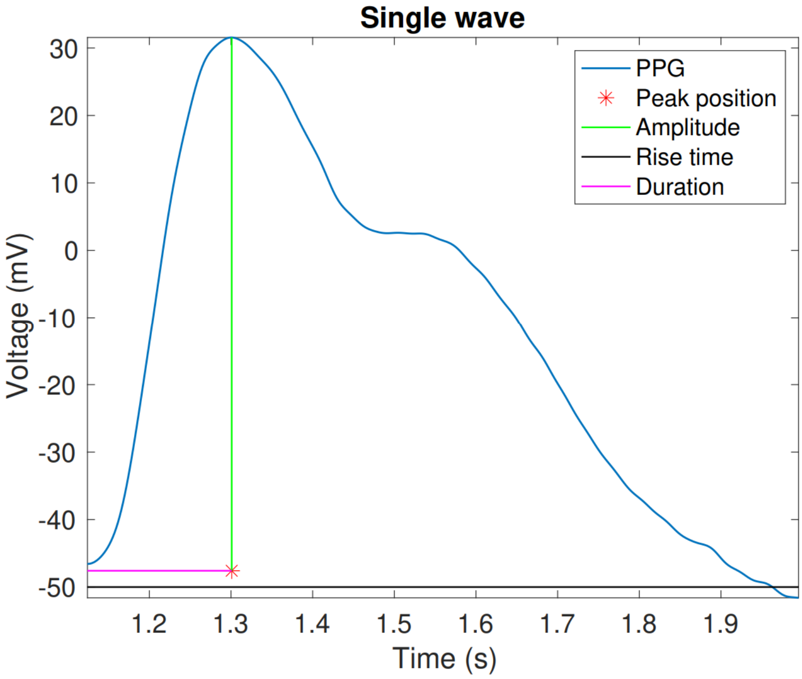


Figure 1.10: PPG wave features from one of the subjects' signals. [64]

Table 1.8: Main PPG-derived features used for stress detection, with domain and typical effect under stress.

Feature	**Description**	**Domain**	**Stress Effect**
IBI (Inter-Beat Interval)	Time between consecutive systolic peaks	Time	↓
BPM	Heart rate computed from Inter-Beat Interval (IBI)	Time	↑
pNN50	Percentage of successive IBIs differing by >50 ms	Time	↓
Power content (total)	Total spectral energy across the considered band	Frequency	Variable
LF norm (0.04–0.15 Hz)	Normalized power in the low-frequency band	Frequency	↑
HF norm (0.15–0.40 Hz)	Normalized power in the high-frequency band	Frequency	↓
LF/HF ratio	Balance between sympathetic and parasympathetic activity	Ratio	↑

Table 1.9: Difference of systolic peak between stress and baseline conditions across subjects [36].

Subject	**Stress Peak (mV)**	**Normal Peak (mV)**	**Difference (mV)**
1	510	481	29
2	923	780	143
3	856	736	120
4	572	537	35
5	701	626	75

recently, non-contact techniques such as thermal infrared imaging (TII) have been adopted, offering high spatial resolution and enabling measurement of localized temperature changes without direct contact with the skin [43].

The physiological relevance of SKT in stress research stems from the sympathetic control of cutaneous microcirculation. Under conditions of acute psychological stress, the sympathetic nervous system induces vasoconstriction in peripheral arterioles, particularly in areas like the fingertips and nose. This vasoconstriction reduces blood flow to the skin and leads to a measurable drop in surface temperature [33, 108]. In contrast, regions such as the forehead and cheeks may exhibit an increase in temperature due to elevated perfusion or sweating responses [28].

Experimental findings support the spatial variability in SKT responses under stress. For instance, as shown by Herborn et al. [43], acute stress is reliably associated with a significant temperature drop in the nasal area, measured using infrared thermography, and this change correlates with stress intensity. Similarly, Kistler et al. [78] found that emotional stress elicited a rapid decrease in fingertip temperature during exposure to aversive stimuli.

To quantify stress-induced thermal responses, several statistical features are commonly extracted from the SKT signal. These include the mean temperature over a given period, the minimum and maximum values, and the standard deviation, which reflects thermal variability. Some studies also consider the slope of the temperature variation to capture transient dynamics.

1.1.7 Respiration and Breath Rate

The respiration signal refers to the physiological process of breathing, which involves inhaling oxygen-rich air and exhaling carbon dioxide-rich air. This process is regulated by the respiratory centers located in the medulla oblongata of the brainstem, which integrate feedback from chemoreceptors and mechanoreceptors to maintain homeostatic gas levels and pH in the blood. The respiratory signal captures this cyclical activity through distinct waveforms representing the inhalation (upward deflection) and exhalation (downward deflection) phases.

Respiration rate, also referred to as Breath Rate (BR), is defined as the number of breaths taken per minute and is a fundamental indicator of respiratory function and overall physiological state. Typically, the respiration rate for a healthy adult at rest ranges between 12 and 20 breaths per minute.

The raw respiration signal, often recorded using thoracic belts or nasal thermistors, is characterized by a quasi-sinusoidal waveform, where each cycle corresponds to a full breath. This signal allows the identification of key respiratory phases, including the peaks of inhalation and exhalation, as well as the minimum point of the cycle. These annotated landmarks can be used to extract meaningful features for physiological analysis, as illustrated in Figure 1.11. Common time-domain features include breath rate, breath amplitude, inter-breath intervals (IBI), inhalation/exhalation duration, and slope of respiratory phases. Frequency-domain features include power spectral density (PSD), and non-linear indices such as sample entropy and permutation entropy, which describe the complexity and variability of the breathing pattern [105].

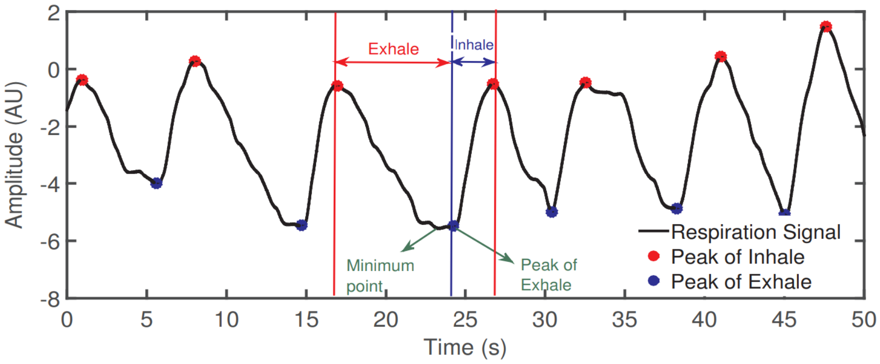
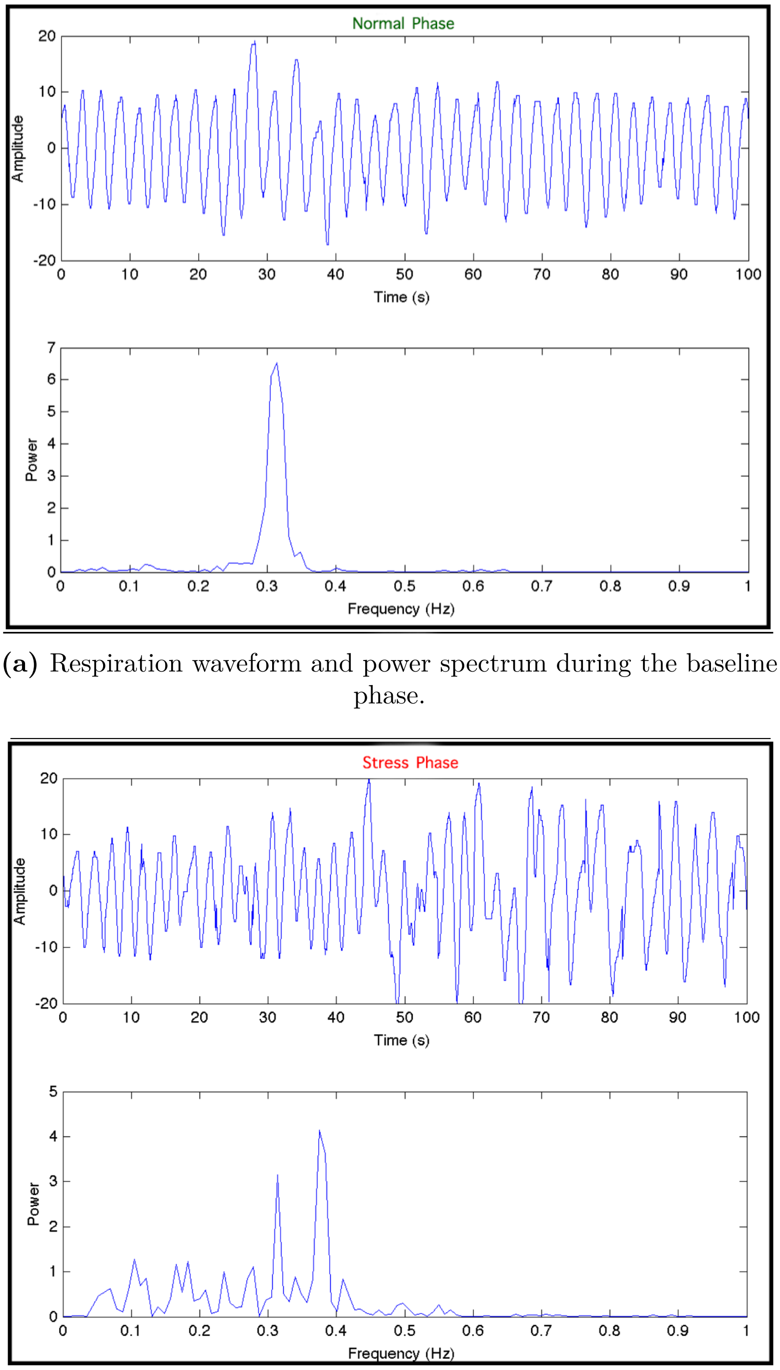


Figure 1.11: Example of a typical respiration waveform under baseline conditions, illustrating inhalation and exhalation phases (Adapted from [115]).

Under stress, the ANS shifts toward sympathetic dominance, resulting in distinct alterations in the respiration signal. The body enters a state of physiological arousal designed to prepare for immediate action (the "fight-or-flight" response), which leads to faster, more variable, and less regular breathing patterns. Specifically, respiration rate increases, the signal becomes less rhythmic, and variability rises significantly across time. A comparative analysis of the respiration signal during baseline and stress phases, as shown in Figure 1.12, highlights these changes. In the baseline phase, respiration is typically slow and regular, with consistent waveform morphology and a dominant frequency peak in the power spectrum. In contrast, under stress conditions, the signal becomes irregular and arrhythmic, and the power spectrum becomes dispersed over a broader frequency range, lacking a clearly defined dominant peak. These changes reflect the body's attempt to increase oxygen intake and adapt to heightened metabolic demands, leading to non-rhythmic breathing patterns and increased respiratory complexity [31].

1.2 Stress-inducing tests

The study of stress in controlled laboratory conditions requires the development of standardized paradigms capable of reliably eliciting psychophysiological and emotional responses. Stressors can be broadly classified into *naturalistic* and *laboratory-induced* conditions. While real-life stressors, such as natural disasters,



(a) Respiration waveform and power spectrum during the baseline phase.

(b) Respiration waveform and power spectrum during the stress phase.

Figure 1.12: Comparison of respiration signal and spectral features under baseline (top) and stress (bottom) conditions (Adapted from [31]).

war exposure or academic examinations have high ecological validity, they lack experimental control and reproducibility, which makes them unsuitable for systematic investigation of stress dynamics [23]. Laboratory protocols, on the other hand, allow for the manipulation of stimuli and conditions, enabling the observation of stress responses under standardized and replicable environments.

Stress induction methods are generally designed to activate the two major physiological systems involved in stress regulation: SAM system and HPA axis. The SAM system is responsible for rapid autonomic responses, such as increases in heart rate, BP, and electrodermal activity, whereas the HPA axis regulates slower endocrine responses, such as cortisol secretion [8]. Different paradigms elicit these systems to varying degrees, depending on whether the stressor is primarily physical, psychological, or social in nature.

According to the literature, laboratory stressors can be grouped into several categories: *cognitive stress tasks*, which impose high mental workload (e.g., Stroop test, mental arithmetic); *psychosocial stress tasks*, which involve social evaluation and performance pressure (e.g., Trier Social Stress Test); *physiological stress tasks*, which rely on physical discomfort or pain (e.g., Cold Pressor Test); and *multicomponent paradigms*, which combine several modalities to maximize stress responses (e.g., Maastricht Acute Stress Test, Mannheim Multicomponent Stress Test) [8, 23, 8].

The choice of a stress induction protocol depends on multiple factors, including the type of stress response to be elicited (psychological, autonomic, or endocrine), the target population (children, adults, clinical groups), and the need for reproducibility across repeated sessions. Importantly, meta-analyses have shown that paradigms including elements of uncontrollability and social evaluative threat are the most effective at eliciting robustHPAaxis activation and cortisol release, highlighting the critical role of social context in stress induction [23].

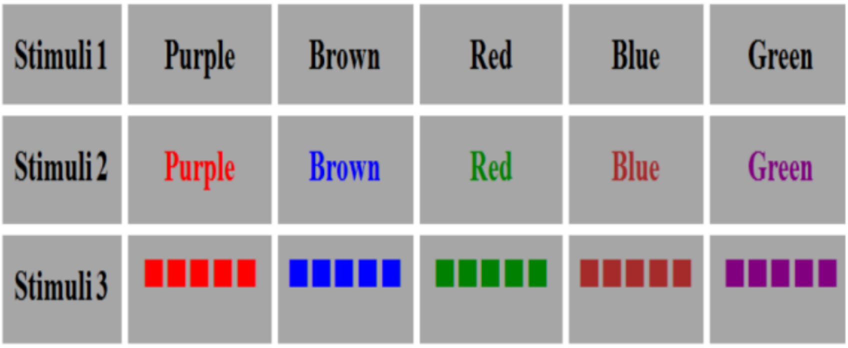
In the following sections, we present an overview of the most commonly employed stress-inducing tasks, organized according to their underlying mechanisms (cognitive, psychosocial, physiological, or multicomponent). For each protocol, we describe its structure, rationale, and stress-inducing potential, together with its main advantages and limitations in research applications.

1.2.1 Stroop color-word test

Among cognitive stress tasks, the Stroop Color Word Test (SCWT) is one of the most widely used paradigms to induce acute psychological stress in laboratory settings. Originally developed by John Ridley Stroop in 1935, the test exploits the interference effect that arises when participants are asked to name the ink color of a word that denotes a different color (e.g., the word “BLUE” printed in red ink) [51]. The first Stroop experiment, often referred to as the classic version, is structured into three tasks: (i) a word-reading task, where participants read color names printed in black ink; (ii) a color-naming task, where participants identify the color of colored squares; and (iii) the incongruent color-word task, where participants must name the ink color of words printed in a conflicting color. This progression is illustrated in Figure 1.13, and the incongruent task in particular generates a strong cognitive conflict between automatic word reading and controlled color naming, which increases reaction times and error rates, thereby inducing acute stress [37].

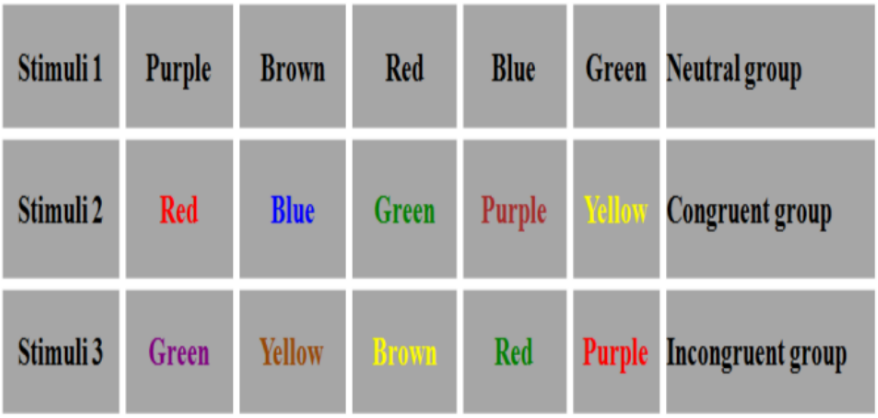
Over time, a more popular and frequently used version of the SCWT has emerged, which also consists of three conditions but with a slightly different structure: (i) a neutral condition, in which color words are displayed in black ink; (ii) a congruent condition, in which the ink color matches the word meaning (e.g., the word "RED" printed in red ink); and (iii) an incongruent condition, in which the ink color and the word meaning are mismatched (e.g., the word "GREEN" printed in yellow ink). This variant is considered particularly effective in eliciting stress because the incongruent condition produces strong interference effects and demands greater cognitive control. An example of this structure is shown in Figure 1.14 [84].

Moreover, several adaptations of the Stroop test have been introduced in the literature, including auditory Stroop, emotional Stroop, and computerized versions, as well as protocols that vary in duration, number of stimuli, or feedback mechanisms. These modifications allow researchers to adjust task difficulty and stress intensity depending on the experimental goals.



Stimuli 1	Purple	Brown	Red	Blue	Green
Stimuli 2	Purple	Brown	Red	Blue	Green
Stimuli 3	🟥🟥🟥🟥🟥	🟦🟦🟦🟦🟦	🟩🟩🟩🟩🟩	🟫🟫🟫🟫🟫	🟪🟪🟪🟪🟪

Figure 1.13: Illustration of the first Stroop experiment: word reading, color naming, and incongruent color-word task [84].



Stimuli 1	Purple	Brown	Red	Blue	Green	Neutral group
Stimuli 2	Red	Blue	Green	Purple	Yellow	Congruent group
Stimuli 3	Green	Yellow	Brown	Red	Purple	Incongruent group

Figure 1.14: Illustration of the more popular Stroop experiment: neutral, congruent, and incongruent conditions [84].

1.2.2 Paced auditory serial addition task

The Paced Auditory Serial Addition Task (PASAT) is a cognitive stress paradigm originally developed by Gronwall and Wrightson in 1974 to assess information processing and working memory performance following traumatic brain injury. In its standard form, participants listen to a continuous sequence of single-digit numbers and are required to add each new number to the one presented immediately before it, reporting the sum aloud. The task is performed under strict time constraints, with digits presented at fixed interstimulus intervals (ranging from 1.2 to 3.6 seconds), and errors are not corrected during the trial [103].

The PASAT induces stress by combining three elements: (i) sustained attention and working memory load; (ii) increasing task pace, which reduces the time available for cognitive processing; and (iii) the impossibility of revising previous answers, which generates frustration and a sense of uncontrollability. This combination reliably evokes acute psychological stress and ANS activation.

Experimental evidence confirms that the PASAT provokes measurable physiological stress responses. In the study by Tanosoto et al., task performance was associated with a significant increase in heart rate, accompanied by a reduction in both low-frequency (LnLF) and high-frequency (LnHF) components of heart rate variability, reflecting sympathetic activation and vagal withdrawal. Conversely, electromyographic (EMG) activity in the masseter muscles did not vary significantly between baseline, task, and recovery phases. These findings, reported in Table 1.10, demonstrate the acute autonomic and cardiovascular effects elicited by PASAT [102].

Table 1.10: HRV and EMG activity before, during, and after the PASAT [102].

Variable	Unit	Baseline	PASAT	Recovery 1	Recovery 2
HR	bpm	71.5 (7.3)	83.8 (13.8)*	72.4 (7.3)	71.2 (7.0)
LnLF	ms^2 Hz^{-2}	5.8 (1.1)	5.0 (0.9)*	6.0 (1.0)	5.8 (0.9)
LnHF	ms^2 Hz^{-2}	5.3 (0.6)	4.6 (1.1)*	5.2 (0.7)	5.1 (0.6)
EMG-LM	μV	75.7 (49.6)	75.7 (56.7)	74.0 (45.2)	67.7 (46.6)
EMG-RM	μV	75.3 (43.8)	78.6 (44.0)	75.2 (44.9)	71.0 (45.2)

1.2.3 Montreal imaging stress task

MIST is a cognitive and psychosocial stress paradigm specifically developed to investigate acute stress responses in neuroimaging environments. It is derived from the Trier Mental Challenge Test and consists of computerized mental arithmetic tasks combined with social evaluative threat. The protocol includes three conditions: (i) a *rest condition*, in which participants view a static screen; (ii) a *control condition*, where arithmetic problems are presented without time pressure or social evaluation; and (iii) an *experimental stress condition*, in which tasks are set just beyond the participant's capacity, with strict time limits and continuous negative feedback comparing their performance with an alleged group average [26].

The combination of high cognitive load, uncontrollability, and social evaluative threat makes the MIST a robust tool for inducing acute stress in functional imaging studies. In particular, the task is effective at activating both the SAM system and HPA axis, as evidenced by changes in cardiovascular parameters and increased cortisol secretion.

Experimental validation confirmed the stress-inducing potential of the MIST. In a Positron Emission Tomograph (PET) study, Dedovic et al. observed that salivary cortisol levels rose significantly under the stress condition compared with the rest condition. As shown in Figure 1.15, cortisol concentrations increased shortly after the onset of the stress condition, peaking around 20–30 minutes, while no such changes were observed in the rest condition. This finding highlights the effectiveness of the MIST in eliciting robust HPA axis activation in controlled laboratory settings.

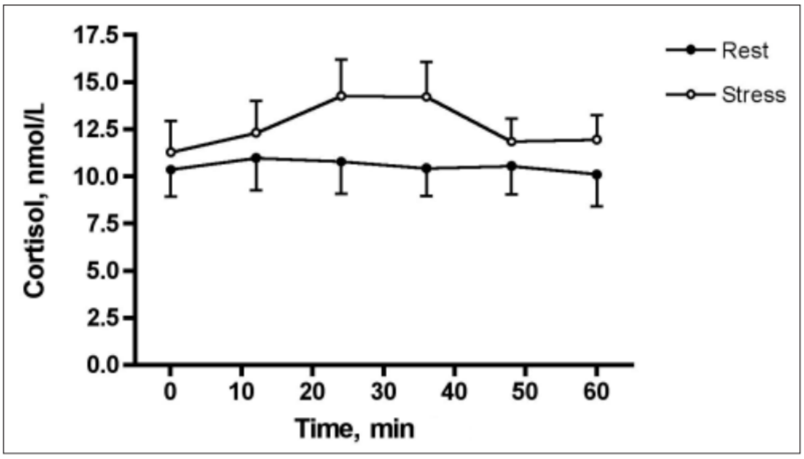


Figure 1.15: Cortisol response to the MIST compared with the rest condition. Data from Dedovic et al. (2005) [26].

1.2.4 Trier social stress test

The Trier Social Stress Test (TSST) is one of the most widely used psychosocial stress paradigms in experimental research. It was originally developed by Kirschbaum and colleagues in 1993 as a standardized laboratory protocol to reliably activate the HPA axis [55]. The TSST consists of three sequential phases: (i) an anticipation period (about 5 minutes), during which participants are told they must deliver a speech; (ii) a public speaking task (5 minutes), often framed as a mock job interview, performed in front of a panel of neutral evaluators; and (iii) a mental arithmetic task (5 minutes), typically involving serial subtractions, also under observation. The evaluators are trained to provide no encouragement, thereby maintaining a strong element of social evaluative threat.

The combination of uncontrollability and social judgment makes the TSST a robust stressor. It has consistently been shown to elicit marked increases in cortisol, Adrenocorticotropic Hormone (ACTH), heart rate, and BP, reflecting strong activation of both the HPA axis and the sympathetic-adrenal-medullary system [4]. Its reproducibility across populations and contexts, along with well-documented endocrine and autonomic responses, have made the TSST the gold standard in psychosocial stress induction. Moreover, several adaptations exist, including the *TSST for Children (TSST-C)* and group-based versions, which preserve the essential features of social evaluation and uncontrollability while adapting the procedure to different populations and research designs [4].

1.2.5 Cold pressor test

The Cold Pressor Test (CPT) is one of the most widely used physiological stress protocols in experimental research. It involves immersing a hand, forearm, or foot into ice-cold water (typically maintained between 0–4°C) for a fixed duration, usually 1–3 minutes [109]. The CPT reliably elicits acute autonomic and endocrine responses, as participants experience both physical discomfort and uncontrollability of the situation.

The CPT induces stress primarily by activating the SAM system, leading to rapid increases in heart rate, BP, and cardiac output, as well as by stimulating the HPA axis, reflected in elevations of salivary cortisol [16]. These combined physiological effects make the CPT a robust paradigm for examining cardiovascular reactivity, autonomic regulation, and pain-stress interactions in both clinical and research contexts.

Experimental evidence confirms its effectiveness as a stressor. Von Baeyer et al. provided methodological guidelines to ensure standardization of water temperature, immersion duration, and measurement of psychophysiological parameters, thereby guaranteeing reproducibility across studies [109]. More recently, Bullock et al. demonstrated that repeated CPT exposure induces a cascade of stress-related changes, including increased heart rate, elevated mean arterial pressure, and decreased high-frequency heart rate variability, alongside heightened self-reported pain and discomfort. Notably, repeated trials showed evidence of habituation in autonomic and subjective responses, offering insight into stress adaptation mechanisms [16].

1.2.6 Maastricht acute stress test

The Maastricht Acute Stress Test (MAST) is a standardized laboratory protocol developed to induce acute stress responses by combining elements of physical and psychosocial stress. It was introduced by Smeets and colleagues in 2012 as a simple and efficient alternative to the TSST, with the advantage of requiring less personnel and being suitable for a wider range of settings [96]. The protocol consists of repeated cycles of two tasks: (i) immersion of the hand in ice-cold water (0–4°C), which elicits strong autonomic activation, and (ii) mental arithmetic tasks performed under social evaluative threat, where participants are pressured to solve problems quickly while receiving negative feedback.

The MAST is designed to activate both the SAM system and the HPA axis, thereby inducing robust physiological and psychological stress responses. Experimental evidence confirms its effectiveness: in the original validation study, participants showed significant increases in systolic and diastolic BP as well as marked elevations in salivary cortisol [96].

Further evidence was provided by Shilton et al., who demonstrated that the MAST reliably elevates cardiovascular reactivity and subjective anxiety levels. In particular, systolic and diastolic BP peaked immediately after the task, while self-reported anxiety increased significantly compared to baseline. These findings, summarized in Figure 1.16, highlight the robust stress-inducing potential of the MAST across autonomic and psychological domains [95].

1.2.7 Verbal fluency task

The Verbal Fluency Task (VFT) is a cognitive paradigm commonly employed to induce acute stress by taxing executive functions and lexical retrieval processes. In its standard form, participants are instructed to produce as many words as possible within a limited time frame (usually 1–3 minutes) that belong either to a semantic category (e.g., animals, fruits) or that begin with a specific letter. Performance is typically monitored and evaluated in real time, which increases the cognitive load and pressure to perform [10].

The VFT induces stress through the combination of time pressure, working memory demands, and continuous lexical search, which together increase perceived workload and frustration. Unlike tasks that primarily rely on external stressors, the VFT generates an internally driven stress response, as participants struggle to maintain fluency under strict temporal and cognitive constraints.

Experimental evidence has shown that the VFT reliably activates the HPA axis. Becker and colleagues demonstrated that performing a VFT significantly increased salivary cortisol levels, peaking approximately 10 minutes after task completion, while participants also reported heightened subjective stress, effort, and fatigue [10]. Interestingly, this effect was not accompanied by sympathetic activation:

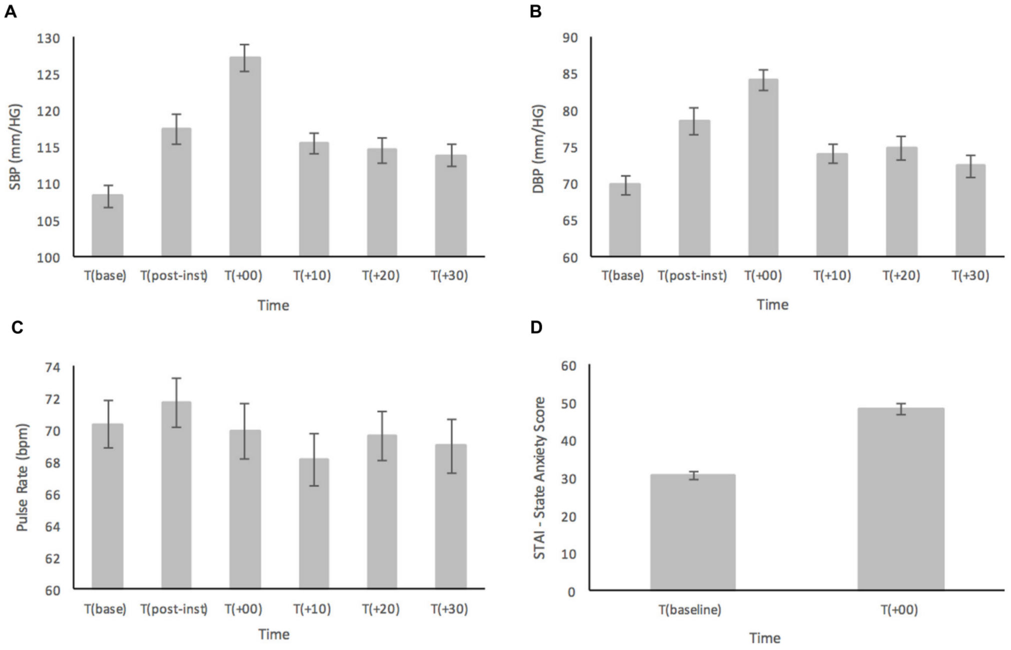


Figure 1.16: Physiological and subjective responses to the MAST. (A) systolic blood pressure (SBP), (B) diastolic blood pressure (DBP), (C) pulse rate (PR), and (D) state anxiety (STAI-Y) before and after the task [95].

measures such as salivary α-amylase, BP, and heart rate did not show significant changes compared to baseline [10]. These findings suggest that the VFT elicits a selective endocrine stress response, making it a valuable protocol for studying acute cognitive stress in controlled laboratory settings.

1.2.8 Mannheim multicomponent stress test

The Mannheim Multicomponent Stress Test (MMST) is a modern stress induction paradigm designed to elicit robust physiological and psychological stress responses by combining several types of stressors within a single experimental protocol. Unlike classical paradigms that focus on a single stress modality, the MMST integrates cognitive, emotional, acoustic, and motivational stressors simultaneously. This multicomponent design aims to maximize ecological validity and to reliably activate both the ANS and the HPA axis [86]. The test typically includes challenging mental arithmetic tasks under time pressure, exposure to aversive sounds such as white noise or explosion-like stimuli, presentation of emotionally negative images, and motivational components such as the threat of monetary loss. The combination of uncontrollability, performance pressure, and emotional discomfort makes the MMST a comprehensive laboratory tool to investigate stress reactivity.

Experimental validation of the MMST by Reinhardt and colleagues demonstrated significant increases in salivary cortisol levels, confirming strong HPA axis activation. In addition, participants exhibited elevated heart rate and electrodermal activity, alongside heightened subjective stress ratings during the task [86]. These converging findings indicate that the MMST reliably provokes acute stress responses across endocrine, autonomic, and psychological domains.

Taken together, the MMST represents a versatile alternative to traditional protocols such as the TSST or CPT. By integrating multiple modalities of stress within a single standardized procedure, it offers a powerful experimental tool for studying stress reactivity in controlled laboratory environments.

1.2.9 Custom and mixed stress protocols

Although standardized stress induction paradigms such as the TSST, PASAT, MIST, CPT, and MAST are widely used in psychophysiological research, many studies have developed *custom or mixed protocols* to tailor stress induction to specific experimental contexts. These approaches often combine cognitive, social, and physiological stressors into a single experimental framework, thereby enhancing ecological validity and allowing the investigation of stress responses in settings closer to real-life scenarios.

For example, Campanella et al. proposed a laboratory protocol specifically designed for workplace-related stress, where participants alternated between different tasks such as manual assembly with and without instructions, mathematical calculations, and a short oral presentation. Each task was separated by rest intervals to establish clear baselines. The combination of cognitive load, manual performance, and social evaluative threat was used to elicit multidimensional stress responses, which were then analyzed through physiological signals (EDA, PPG) acquired with wearable devices [19].

Similarly, other authors have developed custom stress induction tasks by combining Stroop variations with additional challenges such as auditory alarms or multitasking under time constraints [100], or by simulating professional environments, such as emergency departments, where physicians' stress was objectively monitored during real shifts using wearable sensors [53]. These studies demonstrate that while standard protocols remain the cornerstone of experimental stress research, custom-designed paradigms can provide valuable insights, particularly when the goal is to replicate domain-specific stressors or to evaluate stress in applied contexts.

1.3 Stress evaluation methods

The evaluation of cognitive load and stress after experimental induction is a central aspect of psychophysiological research. Once subjects are exposed to standardized stress protocols, it becomes necessary to quantify both their physiological activation and subjective perception of stress in order to establish reliable ground truth. This process is typically based on three complementary approaches: *physiological and neurological measurements*, *behavioral indicators*, and *self-reported questionnaires*. While the first provide objective markers of autonomic and central nervous system activity, the latter capture the individual's perception and appraisal of the stressful condition, which may not always align with biological responses [40]. Behavioral analysis, including speech, facial expressions, and motor activity, offers an additional and unobtrusive source of information.

A key challenge in post-hoc evaluation lies in the definition of ground truth. Stress responses are inherently heterogeneous: some individuals exhibit strong autonomic activation without reporting subjective distress, whereas others may perceive high strain in the absence of clear physiological changes. For this reason, recent studies emphasize the need for multimodal integration, combining physiological data, behavioral cues, and subjective ratings to reduce biases and improve reliability [94]. This multidimensional framework not only enhances the ecological validity of experimental stress studies but also provides robust labeling strategies for machine learning models aimed at automatic stress recognition.

In the following sections, this chapter will review in detail the different methods used for post-hoc assessment of stress, focusing on physiological and neurological

Table 1.11: Summary of stress-inducing protocols, their main features, and physiological responses.

Protocol	Stressor Type	Task Structure	Activated Systems	Measured Responses
SCWT[a]	Cognitive	Word reading, color naming, incongruent color-word naming	SAM (sympathetic arousal)	HR, EDA, EEG (theta/-beta), errors, RT
PASAT[b]	Cognitive	Continuous mental arithmetic under time pressure	SAM, HPA	HR, BP, HRV (↓LnLF, ↓LnHF), EMG, cortisol
MIST[c]	Cognitive + Social	Mental arithmetic under time pressure with negative feedback	SAM, HPA	Cortisol (↑), HR, BP, subjective stress
TSST[d]	Social (evaluative)	Anticipation + speech + arithmetic in front of evaluators	SAM, HPA	Cortisol (↑), ACTH, HR, BP
CPT[e]	Physiological	Hand/arm immersion in ice-cold water (0–4°C)	SAM, HPA	HR (↑), BP (↑), cortisol, pain ratings
MAST[f]	Cognitive + Physiological + Social	Cold pressor + arithmetic under social evaluation	SAM, HPA	Cortisol (↑), HR, BP, EDA, anxiety (↑)
VFT[g]	Cognitive	Word generation (semantic or phonemic) under time pressure	HPA	Cortisol (↑), stress; no clear SAM activation
MMST[h]	Multicomponent	Arithmetic, aversive sounds, negative images, monetary loss threat	SAM, HPA	Cortisol (↑), HR, EDA, subjective stress
Custom	Mixed	Cognitive, manual, and social tasks (e.g., assembly, math, presentations)	SAM, HPA	HR, BP, EDA, PPG, cortisol, subjective stress

[a] SCWT: Stroop Color–Word Test.
[b] PASAT: Paced Auditory Serial Addition Test.
[c] MIST: Montreal Imaging Stress Task.
[d] TSST: Trier Social Stress Test.
[e] CPT: Cold Pressor Test.
[f] MAST: Maastricht Acute Stress Test.
[g] VFT: Verbal Fluency Task.
[h] MMST: Mannheim Multicomponent Stress Test.

measurements, behavioral indicators, self-report questionnaires, and strategies for ground truth establishment in stress research.

1.3.1 Self-reported questionnaires

Self-reported questionnaires are psychometric tools that rely on participants' own evaluations of their emotional, cognitive, or physiological states. They are widely used in stress and cognitive load research because they provide direct insight into the individual's subjective appraisal of stress, thereby complementing objective measurements such as physiological or behavioral signals [38, 25]. Unlike biomarkers, which capture biological reactivity, self-reports reflect the perception and interpretation of the stressor, which may differ considerably across individuals.

Different types of self-report instruments exist. Some focus on the assessment of perceived stress (for example, the Perceived Stress Scale (PSS)), others measure state or trait anxiety (for example, the State-Trait Anxiety Inventory (STAI)), while additional questionnaires are designed to capture affective dimensions (for example, Positive and Negative Affect Schedule (PANAS), Self-Assessment Manikin (SAM Scale)) or task-related workload (for example, NASA-TLX) [25]. These instruments differ in scope, ranging from general stress appraisal to domain-specific evaluation of cognitive effort or emotional activation.

The evaluation of self-reported questionnaires relies on psychometric properties such as reliability (e.g., internal consistency measured by Cronbach's alpha, test–retest stability) and validity (content, construct, and criterion validity). Responses are typically collected on Likert-type scales, assessing frequency or intensity of perceived stressors, and then aggregated into standardized scores that can be compared across individuals or experimental conditions. Despite inherent limitations, such as memory bias and social desirability effects, self-reports remain indispensable because they capture the subjective dimension of stress that cannot be fully inferred from physiology or behavior [38].

Perceived stress scale

PSS is among the most widely applied self-report instruments for assessing perceived stress. It measures the degree to which individuals appraise situations in their lives as stressful, focusing on feelings of unpredictability, uncontrollability, and overload during the past month [7]. The original version includes 14 items, while shorter forms with 10 and 4 items have also been validated. Respondents rate each item on a 5-point Likert scale, with positively worded items reverse-scored before computing a total score; higher values reflect greater perceived stress.

Evidence from the Greek validation study confirmed satisfactory psychometric properties for both the 14- and 10-item versions, including good internal consistency (Cronbach's alpha around 0.82) and a stable two-factor structure distinguishing perceived helplessness from perceived self-efficacy [7]. Moreover, PSS scores showed strong correlations with measures of stress, anxiety, and depression, as well as with the number of self-reported stress-related symptoms, supporting its construct validity. Due to its brevity, reliability, and cross-cultural adaptability, the PSS remains a standard instrument for post-hoc evaluation of stress in both research and clinical contexts.

State-trait anxiety inventory

STAI is a widely used self-report questionnaire designed to measure two distinct components of anxiety: state anxiety, which refers to a temporary emotional condition characterized by feelings of tension and apprehension, and trait anxiety, which represents a stable predisposition to experience anxiety across situations [22]. The original version, known as Form X, was developed in the 1970s and consists of 40 items, equally divided between state and trait subscales. Each item is rated on a 4-point Likert scale, with responses tailored to capture either the intensity of momentary anxiety (state) or the frequency of general anxious tendencies (trait).

A revised version, Form Y, was later introduced to improve the clarity and specificity of items, removing those that overlapped with depressive symptoms and refining the measurement of anxiety. This updated form preserves the same structure of 20 state and 20 trait items, but provides more accurate discrimination between the two constructs. Today, Form Y is the most widely used version of the STAI and is considered a standard tool in both clinical practice and experimental research on stress and anxiety.

Positive and negative affect schedule

PANAS is a widely used self-report questionnaire that measures two independent affective dimensions: Positive Affect (PA), reflecting the extent of experiencing enthusiastic, active, and alert states; and Negative Affect (NA), reflecting distress, upset, and nervousness [112]. The original version contains 20 items, with 10 items for each dimension. Respondents rate how they feel "right now," over the past few days, or in general using a 5-point Likert scale ranging from "very slightly or not at all" to "extremely."

Scoring consists of summing the responses for the PA and NA items separately; each scale ranges from 10 to 50, with higher scores indicating higher levels of positive or negative affect. The PANAS has demonstrated good internal consistency, discriminant validity (PA and NA are largely uncorrelated), and temporal stability, which make it useful for post-hoc evaluations of affect in stress studies when emotional states may vary before, during, or after stress induction.

Despite these strengths, the PANAS has limitations: the meaning of affect adjectives may differ across cultures, time-frame instructions can influence responses, and NA may overlap somewhat with constructs like anxiety or distress in certain contexts. Thus, precise specification of time instructions and translation/cultural adaptation is important when using PANAS in empirical research.

Self-assessment manikin

SAM Scale is a nonverbal pictorial scale developed to measure affective states along three fundamental dimensions: valence (pleasure–displeasure), arousal (activation–deactivation), and dominance (sense of control versus lack of control). Each dimension is represented by a series of simple figures, allowing participants to indicate their feelings quickly and intuitively without relying on verbal descriptions. This makes SAM particularly suitable for cross-cultural research and for studies involving participants with different language backgrounds [13].

Responses are recorded by selecting the figure that best matches the current emotional state for each dimension, and numerical values are then assigned for analysis. Because of its simplicity, SAM has been extensively applied in experimental stress and emotion research, often in combination with physiological or behavioral measures. Its strengths lie in its ease of administration, minimal cognitive load, and cross-linguistic applicability, although the dominance dimension has sometimes been reported as less reliable compared to valence and arousal. An illustration of the SAM is provided in Figure 1.17.

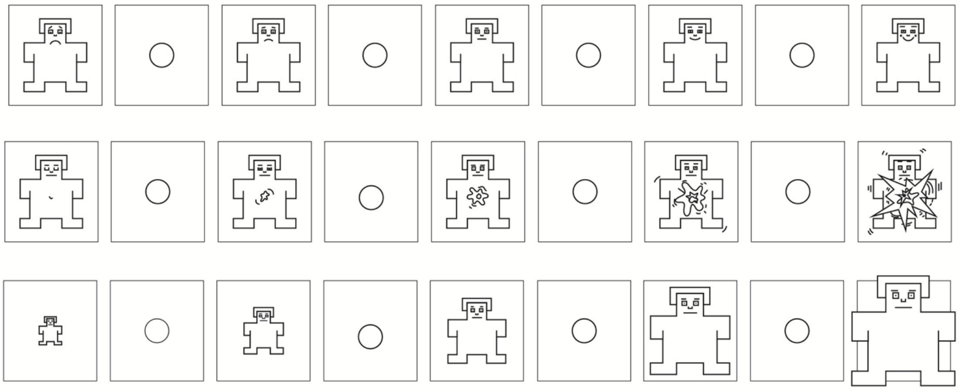


Figure 1.17: The SAM, adapted from Bradley and Lang (1994). The figure depicts the three dimensions of the scale: valence (top), arousal (middle), and dominance (bottom).[13].

NASA task load index

NASA-TLX is a multidimensional tool designed to capture subjective workload during or immediately after task performance. It was developed to provide a standardized way of quantifying perceived workload across diverse domains, addressing the fact that workload is a complex construct influenced by cognitive, physical, and emotional demands [39]. The instrument consists of six subscales: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. Participants rate each subscale after task completion, and an overall workload score is computed either through a weighted or unweighted combination of the ratings.

Originally introduced in aviation, NASA-TLX has been widely adopted in fields ranging from driving and healthcare to human-computer interaction. Its popularity derives from its sensitivity to experimental manipulations, relative ease of use, and diagnostic value provided by the subscales. Over the years, simplified variants such as the Raw TLX (without weighting) have also been used to reduce administration time, but the core six-dimensional structure has remained consistent. NASA-TLX is therefore considered a benchmark tool for subjective workload assessment in both laboratory and applied research. A graphical representation of the six subscales and the overall workload score is provided in Figure 1.18.

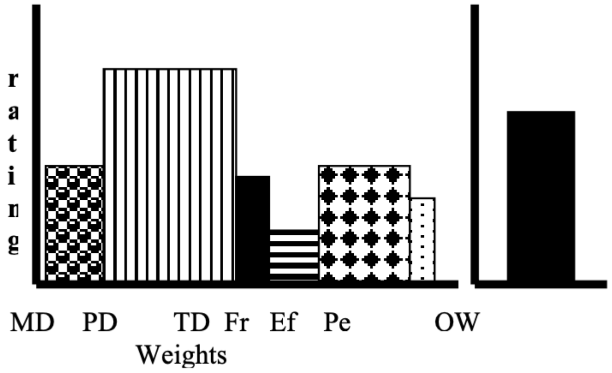


Figure 1.18: Graphic representation of the six NASA-TLX subscales (MD = Mental Demand, PD = Physical Demand, TD = Temporal Demand, Fr = Frustration, Ef = Effort, Pe = Performance) and the computed overall workload (OW), adapted from Hart (2006).[39].

Subjective units of distress scale

The SUDS is a single-item self-report tool used to measure the current level of distress or negative affective intensity, frequently in clinical settings during exposure or emotion-eliciting tasks. Respondents rate their distress on a numerical scale (for example, 0 = calm/no distress to 100 = maximal distress), with descriptive anchors reported in Table 1.12, providing a momentary snapshot rather than a trait measure [72].

Due to its simplicity and immediacy, SUDS is widely adopted for monitoring change within sessions, guiding exposure pacing, or determining whether an intervention task is tolerable. However, recent work raises concerns about its psychometric robustness: issues related to construct underrepresentation (distress may include anxiety, anger, frustration etc.), lack of clearly defined measurement occasions, structural limitations in interpretation, and potential variability in how individuals understand "distress" [72]. As such, while SUDS remains valuable as a rapid measure of subjective distress, researchers should interpret its scores with caution and, when possible, complement it with more comprehensive self-report or physiological measures.

Table 1.12: Anchors for the SUDS, adapted from Wolpe and Wolpe (1981).

Value	**Description**
0	No anxiety at all; complete calmness
1–10	Very slight anxiety
10–20	Slight anxiety
20–40	Moderate anxiety; definitely unpleasant feeling
40–60	Severe anxiety; considerable distress
60–80	Severe anxiety; becoming intolerable
80–100	Very severe anxiety; approaching panic

1.3.2 Ground truth establishment

In stress research, the term *ground truth* refers to the reliable labeling of stress levels against which subjective, physiological, or behavioral data can be validated. Establishing ground truth is essential for both experimental studies and the development of machine learning models, as it provides the benchmark needed to evaluate the accuracy of stress detection systems [106]. Without a robust definition of ground truth, physiological responses may be misinterpreted, and predictive models risk overfitting to noise rather than learning meaningful stress patterns.

Ground truth can be derived from several sources, including self-reported questionnaires, clinician ratings, physiological markers, or contextual data. However, stress is inherently multidimensional, and no single measure can fully capture its complexity. Consequently, researchers often rely on triangulating multiple sources to approximate a reliable label [9]. The importance of ground truth extends beyond methodological rigor: it determines the validity, reproducibility, and translational value of stress research findings. Inaccurate labeling not only undermines scientific conclusions but also limits the applicability of stress detection systems in real-world contexts.

Sources of Ground Truth

Ground truth in stress research can be derived from multiple sources, each contributing a different perspective on the individual's response to stress. The most common source is represented by self-reported questionnaires, which provide a direct but subjective assessment of perceived stress levels. While widely adopted, self-reports are prone to recall bias, social desirability effects, and temporal inconsistencies, making them insufficient as a sole labeling method [106].

Physiological and neurobiological markers, such as heart rate variability, electrodermal activity, or cortisol, offer objective indicators of stress reactivity. These biomarkers capture autonomic or endocrine responses, yet they are influenced by individual variability, health status, and environmental conditions, which complicates their interpretation when used in isolation [106].

Behavioral indicators, including voice modulation, facial expressions, motor activity, and contextual information (e.g., task difficulty, environmental stressors), provide an additional layer of evidence. These measures are particularly valuable in real-world applications where continuous monitoring is required. However, they too are susceptible to confounding factors such as cultural norms or voluntary regulation of behavior [9].

Consequently, no single modality can serve as a flawless ground truth. Current research emphasizes the integration of multiple sources—combining subjective reports, physiological markers, and behavioral observations—to approximate a more reliable and ecologically valid representation of stress.

Challenges and Limitations

Defining ground truth in stress research is inherently challenging because stress is a multidimensional and highly individual phenomenon. One of the main limitations lies in the mismatch between subjective and objective measures: participants may underreport or overestimate their stress levels, while physiological responses can indicate activation in the absence of conscious awareness, a condition often referred to as "silent stress" [106]. Such discrepancies complicate the interpretation of stress labels and raise questions about the validity of single-source ground truth.

Another key challenge is the variability of stress responses across individuals and contexts. Factors such as personality traits, coping strategies, cultural background, and environmental influences can shape both the perception and the physiological expression of stress. This variability undermines the generalizability of labeling schemes developed in controlled laboratory settings, especially when applied to real-world environments [9].

Practical limitations further include the burden of frequent labeling through self-reports, which may be intrusive and reduce ecological validity, and the difficulty of aligning physiological markers with temporally precise ground truth labels. These issues highlight that ground truth in stress research is never absolute, but rather an approximation that requires careful consideration of methodological trade-offs and the integration of multiple complementary sources.

Strategies to Improve Ground Truth

Given the challenges associated with defining ground truth in stress research, several strategies have been proposed to enhance its accuracy and ecological validity. A first approach is the normalization of stress responses relative to individual baselines, which allows researchers to account for inter-individual variability in physiological and subjective measures. By comparing changes from resting states rather than relying on absolute thresholds, it becomes possible to identify stress responses more reliably across heterogeneous populations [106].

Another widely recommended strategy is the integration of multimodal data sources. Combining self-reports, physiological markers, and behavioral indicators enables the triangulation of evidence and reduces the biases inherent in single-method labeling. This multimodal perspective is especially useful in real-world scenarios, where contextual information can play a decisive role in interpreting stress responses [9].

To reduce the burden of intensive labeling, recent research has also explored semi-supervised and weakly supervised approaches, where a limited number of ground truth labels are complemented by algorithmic inference. These methods can mitigate participant fatigue and support continuous monitoring without sacrificing reliability. Finally, the use of standardized stress induction protocols, such as the Trier Social Stress Test or the Cold Pressor Test, provides well-controlled environments for establishing reference conditions, thereby strengthening the consistency of ground truth across studies.

Chapter 2

Technologies for Multimodal Stress Recognition

2.1 Wearable devices for multimodal stress monitoring

Wearable technologies have become central to stress research thanks to portability, non-invasiveness, and the ability to capture multimodal physiological signals in everyday conditions. Unlike traditional laboratory equipment that constrains movement, wearables enable continuous monitoring and improve ecological validity of stress assessments [44, 104]. Their unobtrusive design supports longitudinal observation and seamless integration in experimental protocols focused on real-life stress.

A major advantage is simultaneous acquisition of multiple biosignals—electrodermal activity, photoplethysmography for blood volume pulse, heart rate and heart rate variability, electrocardiography, skin temperature, and inertial measurements—already outlined in Section 1.1 (Biomarkers of Stress). Multimodal fusion within a single device increases robustness and accuracy of stress detection [15, 101]. Research-grade wristbands such as Empatica E4 demonstrated feasible cardiac and electrodermal monitoring outside the lab, although motion and interaction limit data quality [74]. These characteristics help bridge the gap between controlled experiments and real-world monitoring.

Wearables for stress monitoring are commonly designed as wristbands and smartwatches, chest straps, finger rings, adhesive patches, headbands, and smart garments. Each form factor trades off comfort, stability under motion, and battery life [44, 104]. The Empatica E4 is shown in Figure 2.1 as an example of a multimodal wrist device integrating photoplethysmography, electrodermal activity, skin temperature, and tri-axial accelerometry [74].

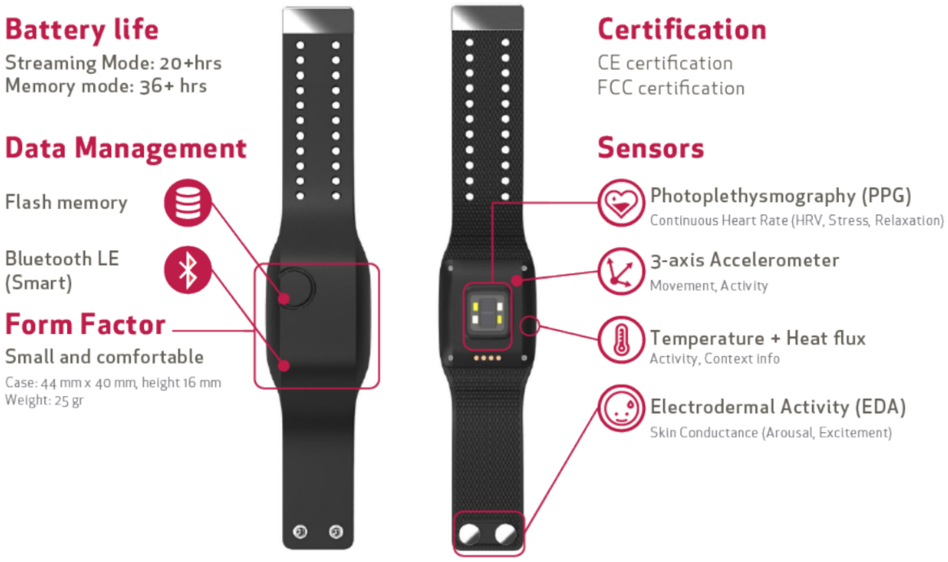


Figure 2.1: Empatica E4 as a representative multimodal wrist device integrating PPG, EDA, skin temperature, and accelerometry [74].

Table 2.1 summarizes representative devices used in the literature, with body location, sensing hardware, and measurable signals. Wrist devices prioritize portability and broad multimodal coverage, whereas chest straps and patches generally offer more stable ECG and respiration measurements.

Table 2.1: Representative wearable devices for stress monitoring: body location, sensors, and measurable signals.

Device	Description	Body location	Measurements / Signals
AutoSense	Wireless system for psychosocial stress	Chestband	2-lead ECG, GSR, RIP respiration, SKT, ambient temp., ACC
Biobeat (watch/patch)	Continuous vital signs	Wrist, chest	HR, HRV, SKT, BP,ECG
Caretaker Medical + ETCO$_2$	Wireless monitoring	Finger-cuff + wrist	CO$_2$, BP, HR, respiration rate
Dexcom G6/G7	Continuous glucose monitoring	Upper arm/abdomen	Glucose

Continued on next page

Device	**Description**	**Body location**	**Measurements / Signals**
Empatica E4	Research-grade wristband	Wrist	EDA, HR (BVP), HRV, IBI, SKT, ACC, gyro
EMOTIV EPOC/Insight/MN8	Brain research headsets	Head	EEG (2–14 ch)
Fitbit Sense 2	Consumer wellness	Wrist	SpO_2, HR, HRV, EDA scan, SKT, breathing rate
Flowtime	Meditation headband	Headband	2-ch EEG, HRV
GlucoMen areo 2K	Glucose monitoring	Wrist	Glucose oxidase, ketonemia
GraphWear	Needle-free glucose	Wrist/abdomen patch	Glucose
HealthPatch MD	Adhesive patch	Chest	ECG, HR, HRV, respiration, SKT, posture
Awario	Arrhythmia detection	Finger jewelry	1-chECG
Hipee Xiaomi	Posture trainer	Neck–shoulder	Posture, movement
ImecECG necklace	NecklaceECG	Neck/chest	ECG, HRV
Muse 2 / S	Meditation and stress	Headband	EEG, PPG, ACC, gyro
K'Watch	Glucose smartwatch	Wrist	Glucose
Lab-on-Skin Xsensio	Biochemical sensors	Arm patch	Cortisol, pH, Na^+, K^+
Microsoft Band 2	Health/fitness	Wristband	Optical HR, ACC, GSR, SKT, UV, others
Mindfield eSense	Conductance sensor	Fingers	Skin conductance
MindWave mobile	Attention/relaxation	Headband	EEG
Movesense	Sports/medical	Chestband	ECG, HR, HRV, ACC
NIRSport 2	Optical spectroscopy	Head	fNIRS, 9-axis ACC
O2Ring	Oxygen monitoring	Finger	SpO_2, HR
Oura Ring	Wellness ring	Finger	HR, HRV, SpO_2, SKT, ACC

Continued on next page

Device	**Description**	**Body location**	**Measurements / Signals**
Prana	Breathing/posture	Waist	Respiration, posture
Sentio Feel	Stress/wellness	Wrist	HR, HRV, EDA, SKT, activity
Shimmer3	Modular platform	Wrist/chest/fingers	ECG, EMG, PPG, GSR, respiration, IMU
Shimmer Verisense	Multi-sensor platform	Wrist/fingers/arm	ACC, gyro, PPG, GSR
Stanford wearable	Academic prototype	Arm/wrist patch	Cortisol (sweat)
HandWave Bluetooth	Conductance sensor	Hand	Skin conductance
UCLA Smartwatch	Cortisol smartwatch	Wrist patch	Cortisol (sweat)
Zephyr	Body straps/shirts	Torso	ECG, HR, HRV, respiration, core temp., ACC

Reliability depends on signal quality across contexts. Typical degradations include motion artefacts affecting PPG and EDA, variability in skin–sensor contact, slow tonic or thermal drift in EDA and skin temperature, and PPG saturation under low perfusion or vigorous motion. Wrist devices maximize ecological validity but are more exposed to movement and contact issues, whereas chest straps and patches usually provide cleaner ECG and respiration at the cost of comfort [104]. Validation studies indicate that research-grade devices such as Empatica E4 reach acceptable agreement for EDA and HR/HRV in controlled and interactive settings, with sensitivity to motion and contact remaining a key limitation; careful windowing and artefact screening are recommended, especially prior to HRV computation [74]. Reviews also note that chest-worn ECG and respiration belts often serve as reference systems when benchmarking wrist platforms; a pragmatic setup combines the ecological advantages of wrist-based multimodal sensing with pilot comparisons against chest references and redundancy across modalities to mitigate single-channel failures [104].

Multimodal acquisition requires attention to synchronization and sampling heterogeneity. Native rates differ across sensors (for example, 4 Hz EDA, 64 Hz BVP, high-rate ECG) and must be resampled to a common timeline; clock drift should be corrected using timestamps and post-hoc alignment, including linear drift compensation or cross-correlation on reference channels such as accelerometry or respiration. Analysis benefits from fixed-length windows with overlap to balance temporal continuity and boundary losses. Event synchronization combines real-time markers with offline matching of experimental logs to physiological timestamps; robust workflows in Wearable Stress and Affect Detection Dataset (WESAD) and cStress highlight the importance of alignment, drift correction, and windowing for reproducible labeling [91, 46].

Deployment in experimental protocols entails correct mounting and calibration, verification of electrode–skin contact, and battery checks to avoid gaps in laboratory settings. Ecological studies benefit from standardized participant instructions and automated quality checks to limit motion artefacts and data loss [46]. Interoperability with stimulus software and event logging is essential, since wireless latencies and dropouts often require post-hoc correction and careful ground-truth alignment. Ethical and privacy safeguards are mandatory in continuous monitoring: informed consent, anonymization, GDPR-compliant storage, and clear data-ownership policies are recommended, particularly for in-the-wild studies [24].

Physiological signals acquired by wearables exhibit low signal-to-noise ratio and are affected by internal sources such as respiration, muscle activity, and cardiac motion, as well as external sources including electrode displacement, power-line interference, and ambient vibrations. Effective preprocessing enhances data quality prior to feature extraction or end-to-end learning. Typical approaches include band-limited digital filtering, empirical mode decomposition, wavelet-based time–frequency denoising, blind or semi-blind source separation with independent or principal component analysis, and adaptive filtering driven by reference channels. The subsequent chapter details the preprocessing and quality-control steps adopted for ECG, EDA, PPG, skin temperature, and motion signals within the unified pipeline used throughout the thesis.

2.2 Signal processing of wearable signals

Physiological signals acquired from wearable devices are affected by motion, variable skin–sensor contact, thermal drift, ambient interference, quantization noise, and sampling inconsistencies. These factors degrade morphology, distort baseline, and contaminate the spectral content of channels such as EDA, PPG/BVP, ACC, and skin temperature. Signal processing is therefore essential to stabilize the baseline, attenuate out-of-band components, preserve salient temporal landmarks, and deliver representations suitable for feature extraction and machine learning. A principled preprocessing stage also improves reproducibility by standardizing how artefacts and interference are handled across datasets and recording contexts.

2.2.1 Digital filtering

Digital filtering plays a crucial role in the preprocessing of biosignals, aiming to reduce noise and artifacts that compromise signal quality. Despite efforts in noise prevention, completely eliminating interference at the source is impractical. Therefore, digital filters are employed to either separate useful signal components from interference (signal separation) or to restore distorted signals (signal restoration) [89].

Digital filters are categorized primarily into two classes: Finite Impulse Response (FIR) and Infinite Impulse Respons (IIR) filters. FIR filters apply a convolution between the input signal and a finite-duration impulse response, defined by

$$y[n] = \sum_{k=0}^{M} h[k]\, x[n-k],$$

where $y[n]$ is the output at time n, $x[n]$ the input at time n, $h[k]$ the filter coefficients, and M the filter order. FIR filters are inherently stable and exhibit linear phase characteristics, which is ideal for preserving waveform shapes such as ECG and EEG signals [75]. Their coefficients can be obtained by approximating a desired frequency response through windowing of the ideal impulse response, frequency sampling with inverse transforms, or optimization-based algorithms such as Parks–McClellan. Once designed, the filter operates by forming a weighted sum of the current and previous input samples, yielding a non-recursive process with linear phase.

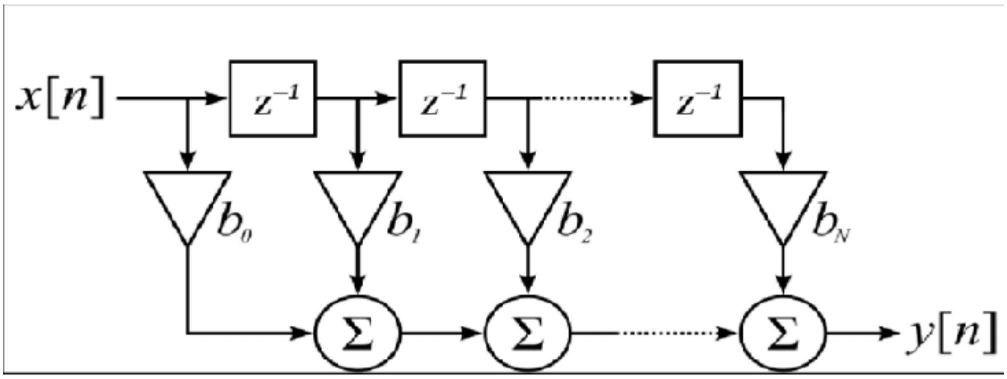


Figure 2.2: FIR block diagram [75].

IIR filters operate via a recursive difference equation that includes both current and past inputs and outputs,

$$y[n] = \sum_{k=0}^{N} b[k]\, x[n-k] - \sum_{k=1}^{M} a[k]\, y[n-k],$$

with feedforward coefficients $b[k]$ and feedback coefficients $a[k]$. IIR filters are computationally efficient because sharp transitions can be achieved with fewer parameters. Design typically starts from an analog prototype such as Butterworth, Chebyshev, or Elliptic, followed by a mapping to the digital domain via bilinear transformation or impulse invariance. The recursive structure enables compact implementations but introduces potential instability and nonlinear phase if coefficients are poorly chosen. In practice, the output at each step is computed as a weighted combination of current and past inputs minus a weighted combination of past outputs, a mechanism that confers efficiency at the expense of stricter stability and quantization constraints.

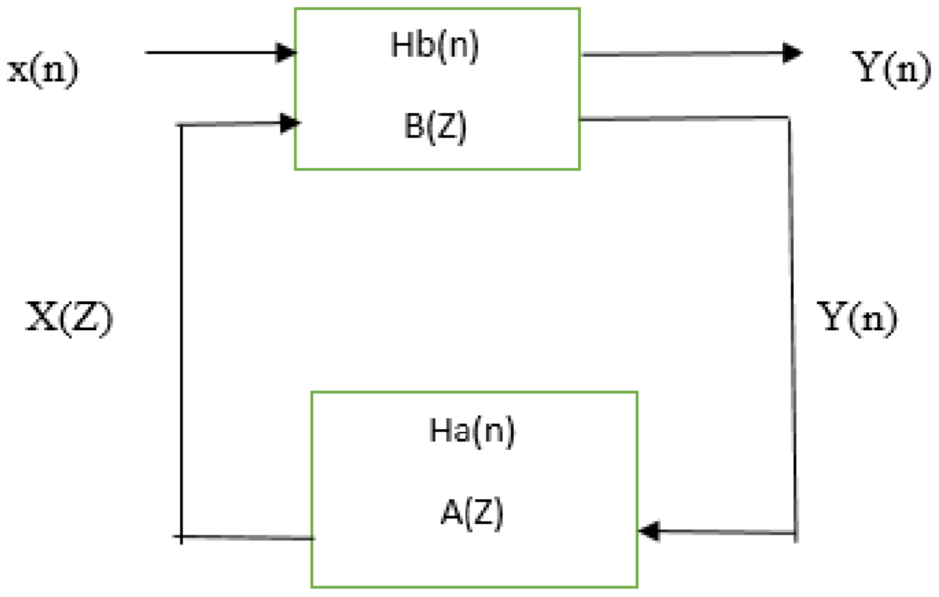


Figure 2.3: Representation of Infinite Impulse Response [75].

Depending on the signal and interference characteristics, filters are configured to emphasize or suppress frequency regions of interest. Low-pass responses attenuate high-frequency noise; high-pass responses remove baseline wander and very slow drifts; band-pass responses isolate physiological bands; narrowband notch responses reject mains interference at 50/60 Hz. Prototype families offer different trade-offs: Butterworth ensures a maximally flat passband, Chebyshev yields steeper transitions with controlled ripple, and Elliptic achieves the sharpest roll-off at the cost of ripple in both passband and stopband. While analog filters rely on physical components, digital implementations are algorithmic and therefore flexible, reproducible, and adaptable, including adaptive variants that adjust coefficients in real time using auxiliary references such as accelerometry. Selecting an appropriate filter requires balancing signal properties, noise profiles, computational constraints, and the need to preserve waveform morphology.

2.2.2 Noise and Artifact Removal in PPG Signal

Photoplethysmography is an optical technique that tracks pulsatile changes in arterial blood volume by measuring light intensity variations at the skin surface. The resulting blood volume pulse provides timing information for heart rate and supports derivative measures such as inter-beat intervals and heart rate variability [92]. In wearable settings the waveform is vulnerable to motion of the sensor relative to the skin, ambient light leakage into the photodetector, low-frequency physiological modulations from respiration and vasomotor tone, and high-frequency electronic interference [42, 92]. These disturbances distort amplitude and morphology, shift the baseline, and corrupt the spectral band used for cardiac timing.

Digital filtering is employed to stabilize the baseline and isolate the cardiac component. A band-pass response centered on the physiological heart-rate band (typically about 0.5–5 Hz for resting to moderate activity) suppresses very slow drifts and high-frequency contamination while preserving systolic upstrokes and diastolic decays important for peak detection. When power-line interference is present, a narrow notch at 50 or 60 Hz reduces residual mains components without substantially affecting the passband. Mild low-pass smoothing after band-pass filtering can further attenuate quantization noise and small oscillations that hinder robust peak picking. Filter design trades off transition sharpness, passband flatness, and phase behavior; linear-phase FIR implementations preserve pulse morphology, whereas IIR realizations provide compact, efficient responses if stability and phase distortion are controlled. With appropriate band-limiting and optional notch rejection, the PPG/BVP signal becomes suitable for reliable beat detection and subsequent temporal feature extraction [42, 92].

2.2.3 Noise and Artifact Removal in EDA Signals

EDA signals are highly sensitive to various noise sources and artifacts, which can mask meaningful skin conductance responses and compromise the reliability of stress detection. These disturbances are generally categorized into two types: *intrinsic noise* and *extrinsic noise*.

Intrinsic noise originates from physiological processes that overlap with the frequency range of true electrodermal responses. The most common example is respiration-related noise, which introduces low-frequency oscillations that can resemble genuine EDA activity. Since traditional filtering techniques cannot reliably distinguish between respiration and actual electrodermal reactivity, more advanced approaches have been proposed. Among them, *noise reference signal–based denoising methods* have gained attention. In this approach, an auxiliary signal correlated with respiration (e.g., respiration belt or photoplethysmography-derived respiration) is collected simultaneously with EDA. This reference signal is then used within an adaptive filtering framework to identify and remove the respiration-induced components from the EDA signal. Such strategies have shown promising results in suppressing respiration noise while preserving the integrity of true electrodermal responses [66]. Figure 2.4 shows an example of respiration noise attenuation in EDA, while Figure 2.6 summarizes the main steps of the noise reference signal–based denoising method.

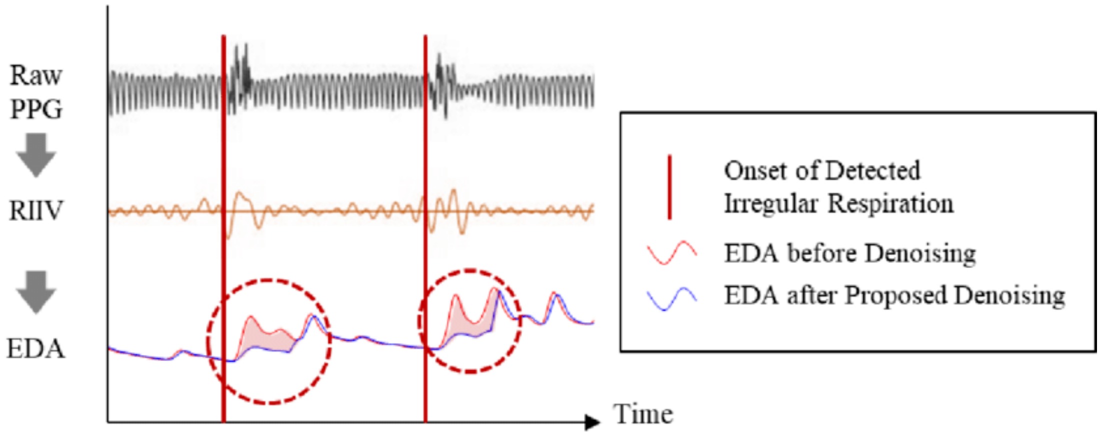


Figure 2.4: Attenuation of respiration noise in EDA signals.

Extrinsic noise arises from external interferences during signal acquisition, including powerline interference, electrode displacement, and motion artifacts caused by hand or wrist movements. These disturbances are broadband and irregular, often resulting in abrupt fluctuations in skin impedance that can obscure phasic responses. Basic filtering techniques remain the most widely used strategies to reduce extrinsic noise, with high-pass and low-pass filters applied to suppress slow drifts and high-frequency contamination, respectively. Additionally, moving average filters are commonly employed to smooth the signal and attenuate transient disturbances. In more complex cases, wavelet-based denoising techniques have also been applied, leveraging the non-stationary nature of EDA to selectively suppress artifacts while preserving tonic and phasic components. The main steps of extrinsic noise attenuation are illustrated in Figure 2.5.

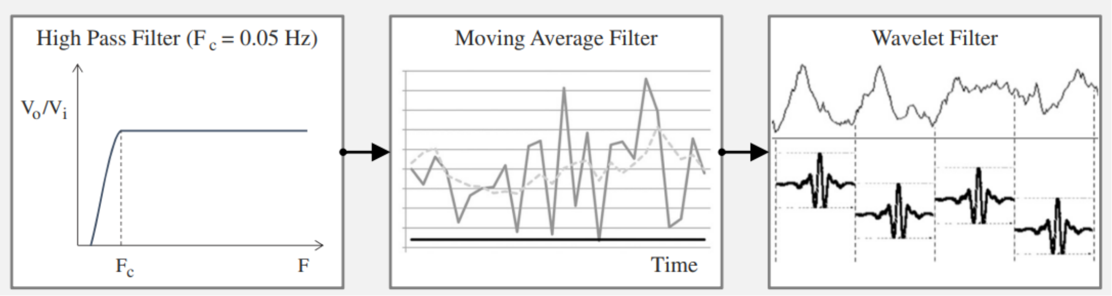


Figure 2.5: Extrinsic noise attenuation steps applied to EDA signals.

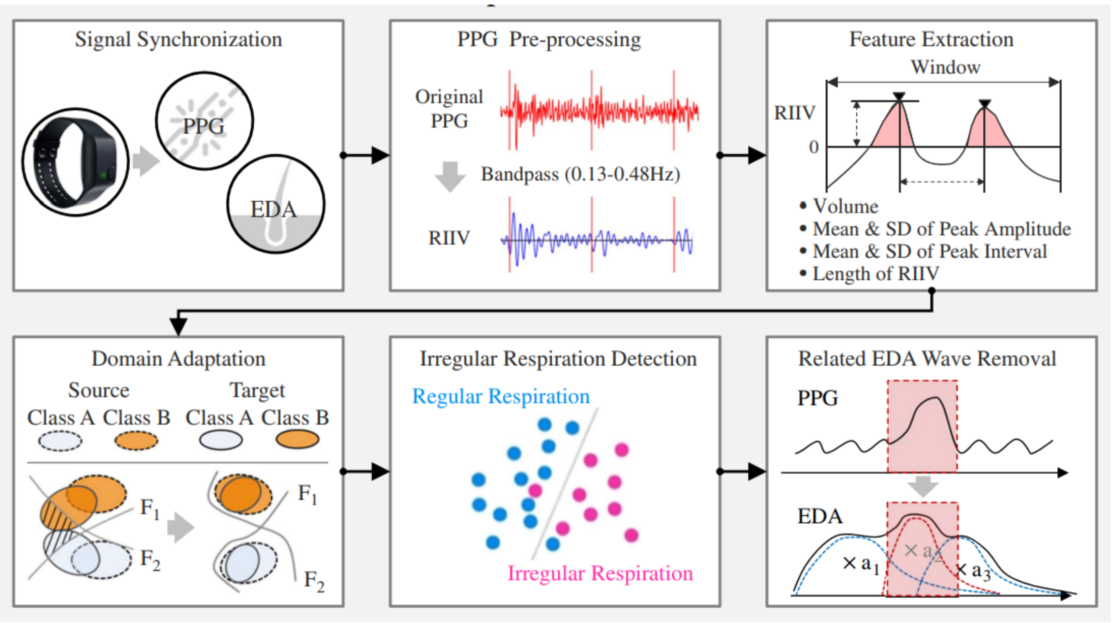


Figure 2.6: Intrinsic respiration noise removal process using reference-based denoising.

2.2.4 Noise and Artifact Removal in ACC Signal

Raw triaxial acceleration combines body motion, gravity, and sensor noise. For downstream analysis, a first step is to remove the quasi-static gravitational component and attenuate out-of-band noise using digital filtering. A practical approach applies a low-cut, zero-phase Butterworth high-pass filter to each axis at a cutoff in the sub-hertz range, so that posture-related acceleration (near 0 Hz) is retained in the low-frequency branch while dynamic movement is preserved in the high-frequency branch. The Euclidean norm of the high-pass outputs provides a gravity-reduced magnitude signal that is less sensitive to sensor orientation. When activity is vigorous or electronic interference is present, the high-pass stage can be complemented by a gentle upper cutoff to limit high-frequency noise, yielding a narrow movement-focused passband. Empirical evaluations on wrist and hip accelerometry have shown that fourth-order Butterworth implementations with cutoffs around 0.2–0.5 Hz effectively separate movement from gravity and yield stable summary metrics for free-living behavior, with band-pass variants further reducing high-frequency artefacts when necessary [107].

2.3 Machine Learning for Stress Detection

AI refers to the broad scientific field devoted to the creation of computational systems capable of replicating or mimicking human cognitive functions such as perception, reasoning, and decision-making. Within this domain, ML represents a methodological approach that enables machines to learn from data rather than being explicitly programmed with rule-based logic. Instead of relying on static instructions, ML algorithms iteratively improve their performance as more data becomes available, thus allowing the automatic discovery of hidden patterns and the generation of predictive models. This paradigm is particularly powerful in biomedical applications, where the variability and complexity of physiological processes make it challenging to define rigid analytical rules [113].

ML can be broadly categorized into three learning paradigms: *supervised learning*, where the model is trained with labeled data to perform tasks such as classification (e.g., stress vs. non-stress) or regression (e.g., stress intensity prediction); *unsupervised learning*, which aims to identify latent structures in unlabeled data (e.g., clustering subjects by stress reactivity profiles); and *reinforcement learning*, where agents learn to optimize sequential decisions through interaction with an environment. In the context of stress research, supervised classification is the most common approach, but unsupervised clustering has also been explored to uncover natural groupings of physiological responses [2].

Within the ML framework, DL has emerged as a transformative subfield. DL relies on artificial neural networks with multiple layers (deep architectures) that allow hierarchical representation learning. Unlike traditional ML models that typically require handcrafted feature engineering, DL models automatically extract increasingly abstract features from raw input data, making them particularly effective in dealing with complex and high-dimensional biosignals such as ECG, EEG, or EDA. CNN, for instance, are widely applied to time-series stress-related data because of their ability to capture local temporal dependencies, while Recurrent Neural Networks (RNN) and their variants (e.g., Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU)) are employed to model sequential dependencies and long-term temporal dynamics [15].

Figure 2.7 illustrates the hierarchical relationship between AI, ML, and DL. AI encompasses the overarching goal of intelligent behavior, ML represents the subset dedicated to learning from data, and DL constitutes the most advanced class of ML methods that leverage large-scale neural networks to achieve unprecedented performance in pattern recognition tasks.

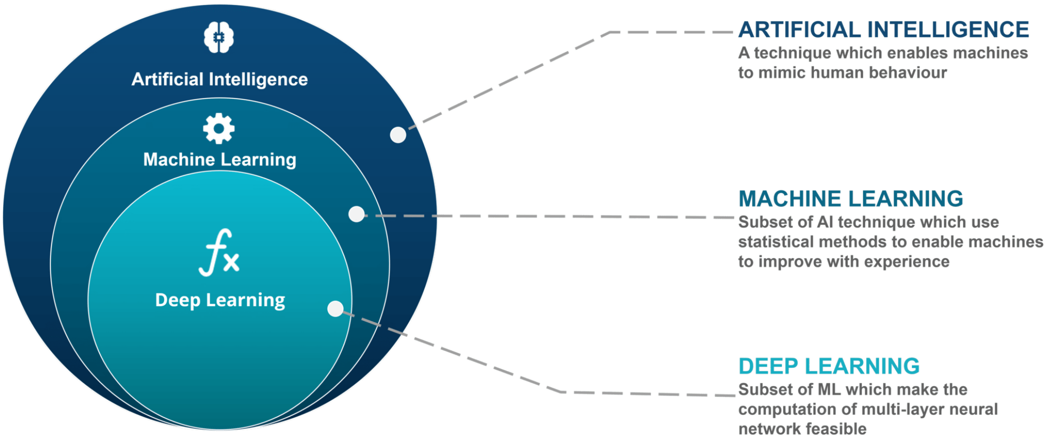

Artificial Intelligence
Machine Learning
f_x
Deep Learning
ARTIFICIAL INTELLIGENCE
A technique which enables machines to mimic human behaviour
MACHINE LEARNING
Subset of AI technique which use statistical methods to enable machines to improve with experience
DEEP LEARNING
Subset of ML which make the computation of multi-layer neural network feasible

Figure 2.7: Schematic representation of the relationship between AI, ML, and DL.

2.3.1 Learning Paradigms

Machine learning encompasses different paradigms that define how algorithms interact with data, depending on the availability of labels and the nature of the task. These paradigms—supervised, unsupervised, semi-supervised, and reinforcement learning—represent the foundations of modern computational intelligence [52]. Supervised learning relies on labeled datasets, where each input is associated with a known output. The goal is to learn a mapping function from inputs to outputs, which can be applied to new unseen data. Within this paradigm, two main tasks can be distinguished:

- **Classification**: the output variable is categorical, for example distinguishing between “stress” and “no stress” conditions, or predicting multiple stress levels.

- **Regression**: the output variable is continuous, such as a stress score or an index of workload.

Supervised methods such as decision trees, support vector machines, and neural networks have proven effective in solving both linear and non-linear problems. Their performance depends strongly on the quality and size of the labeled dataset [88]. In unsupervised learning, no labels are available, and the algorithm seeks to uncover hidden structures within the data. The primary task in this paradigm is clustering, which groups data points into clusters based on similarity. For instance, clustering can be used to discover natural groupings of physiological responses or to identify subject-specific stress profiles. Common techniques include k-means, Gaussian mixtures, and self-organizing maps. Additionally, dimensionality reduction methods such as Principal Component Analysis (PCA) or autoencoders are often employed to project high-dimensional data into lower-dimensional spaces, facilitating visualization and analysis [88]. Semi-supervised learning bridges supervised and unsupervised approaches by exploiting a small amount of labeled data together with a large quantity of unlabeled data. This paradigm is especially valuable when labeling is expensive, but unlabeled data are abundant. Recent advances include semi-supervised generative models (e.g., variational autoencoders and GANs) and teacher–student models, which leverage unlabeled data to improve model robustness and generalization [81]. Semi-supervised learning is increasingly relevant in domains where large-scale annotation is prohibitive but reliable predictions are still required. Reinforcement learning (RL) differs from the previous paradigms, as it focuses on learning through interaction with an environment. An agent iteratively selects actions to maximize a cumulative reward, using strategies that balance exploration and exploitation. RL has been applied successfully in control systems, robotics, and adaptive decision-making tasks. While conceptually distinct, reinforcement learning is often implemented using function approximators such as neural networks, connecting it to the broader ML ecosystem [52].

2.3.2 Popular Algorithms in Stress Detection

Several machine learning algorithms have gained popularity in the context of stress detection due to their ability to model complex and nonlinear relationships between physiological signals and stress states. The following section briefly introduces the most common algorithms, highlighting their principles, strengths, and limitations.

Logistic Regression

Logistic Regression (LR) models the probability of a categorical outcome by mapping a linear combination of input features to the unit interval through the logistic function, making it a simple and interpretable baseline for stress classification [14]. For a feature vector $\mathbf{x}$, the model estimates

$$P(y = 1 \mid \mathbf{x}) = \frac{1}{1 + \exp\Big(-(\beta_0 + \sum_{i=1}^{n} \beta_i x_i)\Big)}. \tag{2.1}$$

Parameters are learned by maximizing the (regularized) log-likelihood; ℓ_2 (Ridge) or ℓ_1 (LASSO) penalties improve generalization in high-dimensional physiological feature spaces, and multinomial variants extend the formulation beyond binary labels [49]. Strengths include low computational cost, calibrated probabilities, and coefficient-level interpretability; limitations arise when decision boundaries are strongly nonlinear or when complex feature interactions dominate, conditions under which kernel methods or neural networks generally provide higher accuracy at the expense of transparency [14].

Decision Trees and Random Forests

Decision Trees partition the feature space by recursive splits on variables to produce a hierarchical structure of decision rules and terminal leaves, offering transparent and easily interpretable models for stress classification from physiological features [49]. Splitting criteria such as entropy or Gini impurity guide node selection, while depth limits and pruning control overfitting. Strengths include handling mixed data types and capturing nonlinear thresholds; limitations arise from high variance and instability when trees grow deep [76]. Random Forest (RF) mitigate these issues by aggregating many decorrelated trees trained on bootstrap samples and random feature subsets, improving generalization and robustness to noise at the expense of reduced interpretability and higher computational cost [49, 76].

Gradient Boosting

Gradient Boosting (GB) builds a strong predictor by fitting shallow decision trees in sequence, each correcting the residual errors of the current ensemble, thus optimizing a chosen loss function in a stage-wise fashion [14, 49, 76]. Compared with bagging ensembles, this approach typically yields higher accuracy on structured, tabular physiological features but is more sensitive to hyperparameters and overfitting, which are controlled through learning rate, tree depth, subsampling, and regularization. Modern implementations such as XGBoost and LightGBM improve scalability and generalization via explicit regularization, histogram-based

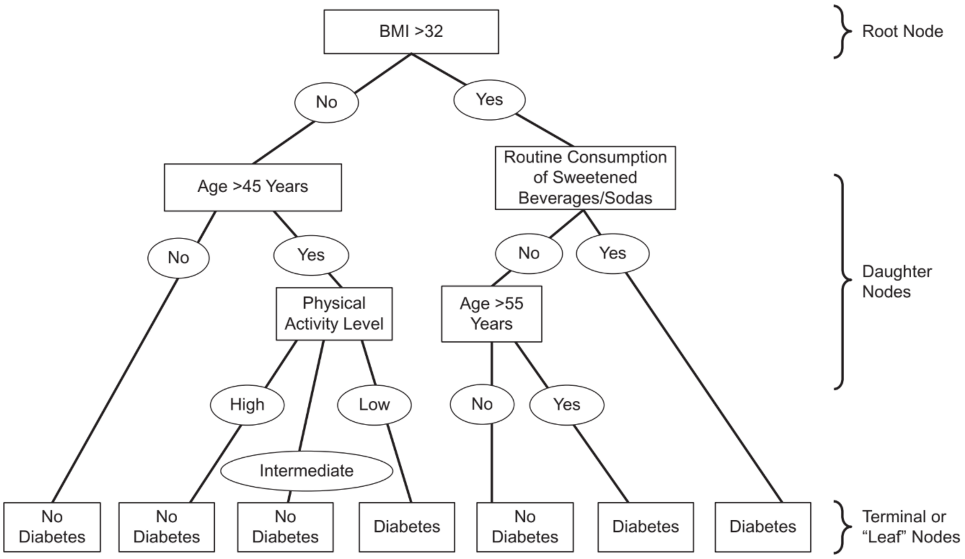


Figure 2.8: Schematic representation of a decision tree classifier. The model recursively partitions data into subsets based on feature thresholds, producing a hierarchical structure of decisions.

split finding, and efficient handling of sparsity, making them effective for multimodal stress recognition where nonlinear interactions among features are prevalent [14].

Support Vector Machines

SVM learn a separating hyperplane that maximizes the margin between classes, yielding robust decision boundaries for binary and multi-class problems [14, 49]. Soft-margin formulations control the trade-off between margin size and training errors through a regularization parameter, improving resilience to noise typical of physiological data. Nonlinear relationships are handled via kernel functions that implicitly map inputs to higher-dimensional spaces; the radial basis function is a common choice when complex, nonlinearly separable patterns are present [76]. SVMs perform well on high-dimensional and small-sample settings and rely only on support vectors, which helps limit overfitting. Limitations include computational cost on very large datasets and reduced interpretability when nonlinear kernels are used [14].

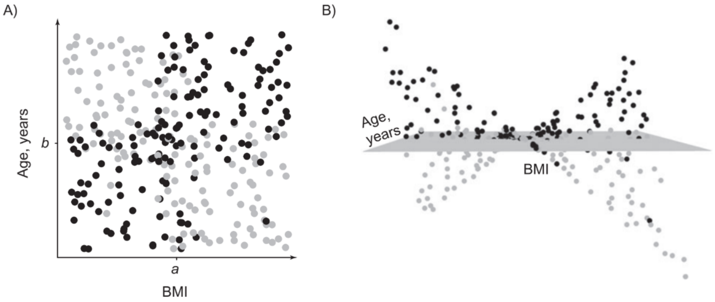


Figure 2.9: SVM decision boundary separating two classes. The margin is maximized between the support vectors of each class, defining the optimal separating hyperplane.

Artificial Neural Networks

ANN are layered computational models that learn nonlinear mappings from inputs to outputs through trainable weights and activation functions [76, 14]. Feedforward architectures with one or more hidden layers are trained by backpropagation to minimize a task-specific loss, enabling the automatic capture of complex relationships that are difficult to express with linear models. In stress detection from physiological signals, ANNs integrate heterogeneous features (e.g., HRV, EDA, PPG) and model their interactions, often improving accuracy over simpler classifiers when sufficient labeled data are available. Main drawbacks include the need for careful hyperparameter tuning, increased data and computational requirements, and limited interpretability compared with transparent models such as LR or shallow trees [76, 14].

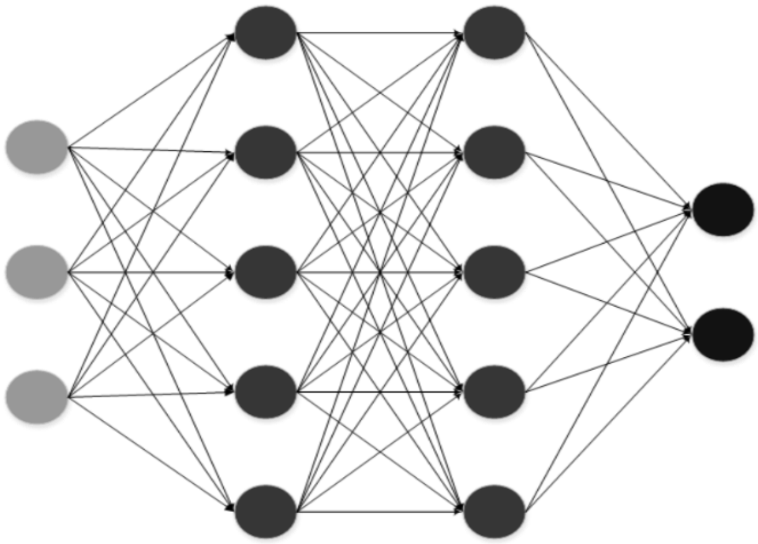


Figure 2.10: Structure of a simple ANN with input, hidden, and output layers. Each neuron transforms the weighted input via an activation function to capture nonlinear relationships between physiological features.

Convolutional Neural Networks

CNN are deep models that learn hierarchical, local patterns through convolutional filters, making them well suited to structured inputs such as time series and spectrograms from physiological sensing [65, 76]. In stress detection, 1D-CNNs can operate directly on raw sequences to capture morphology and short-range temporal dependencies, while 2D-CNNs process time–frequency maps or multimodal grids to exploit cross-channel structure. Compared with feature-engineered pipelines, CNNs frequently improve accuracy by discovering discriminative motifs end-to-end, with common architectural elements including stacked convolutional blocks, nonlinear activations, and downsampling to control capacity. Typical limitations are data and compute demands and reduced interpretability relative to simpler models [65, 76].

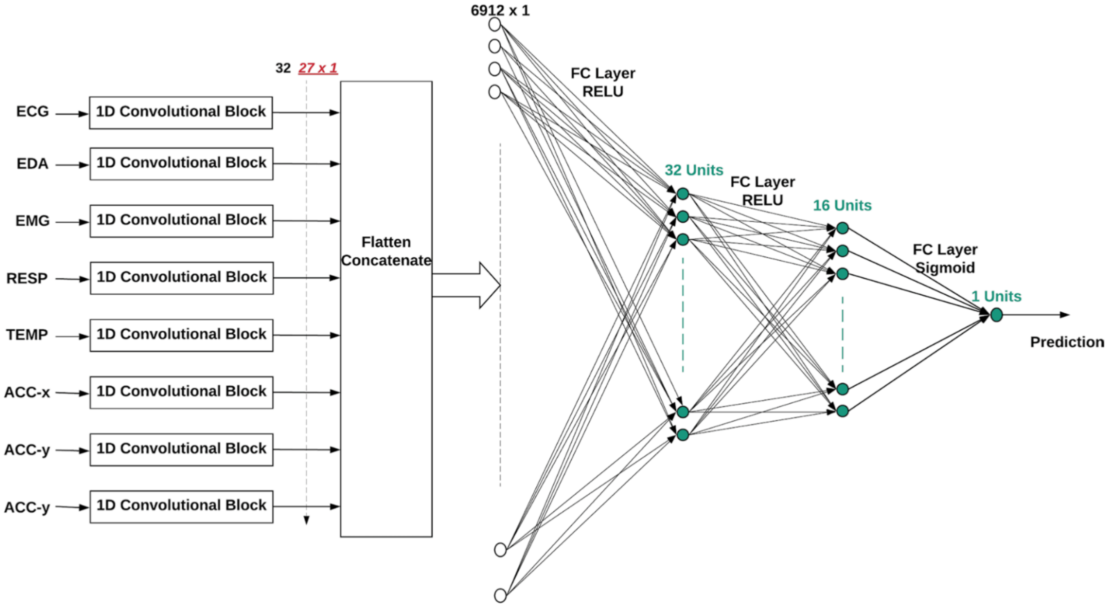


Figure 2.11: CNN with convolutional, pooling, and fully connected layers that extract and aggregate informative patterns from physiological signals for stress detection.

2.3.3 Performance Metrics for Classifiers

Performance metrics play a fundamental role in machine learning classification tasks, as they allow to quantify the performance of a model, compare it with alternative approaches, and assess its ability to generalize to unseen data [83, 73]. In the context of stress recognition from physiological signals, where classification models are often applied to distinguish between stressed and non-stressed conditions, it is crucial to rely on a set of well-defined evaluation criteria to ensure model robustness and clinical applicability.

The starting point for performance evaluation in binary classification is the *confusion matrix*, which summarizes the relationship between predicted and actual classes (Figure 2.12). Each element of the matrix corresponds to one of four possible outcomes: True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These quantities allow the computation of fundamental performance measures.

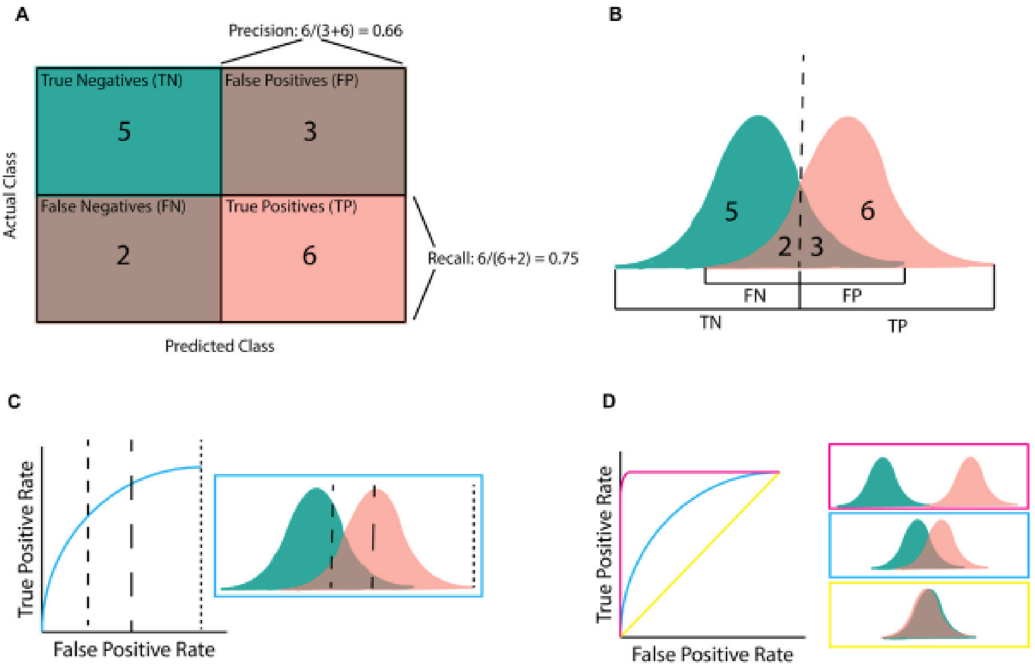


Figure 2.12: Illustration of classification metrics based on the confusion matrix, including true positive rate (TPR) and false positive rate (FPR) [73].

The most intuitive measure is the *accuracy*, which expresses the fraction of correctly classified instances over the total number of predictions:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{2.2}$$

Although widely used, accuracy can be misleading in the presence of imbalanced datasets, as it does not distinguish between types of errors and may overestimate performance when one class dominates the distribution [83].

To overcome this limitation, additional metrics are introduced. *Precision* measures the proportion of correctly identified positive predictions:

$$Precision = \frac{TP}{TP + FP}, \tag{2.3}$$

while *Recall* (or Sensitivity) quantifies the proportion of actual positives that are correctly classified:

$$Recall = \frac{TP}{TP + FN}. \tag{2.4}$$

Both metrics are essential in stress detection, since high precision ensures that predicted stress states are reliable, whereas high recall guarantees that most actual stress events are captured. The two can be combined in the *F1-score*, defined as the harmonic mean of precision and recall:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{2.5}$$

Another important metric is *Specificity*, also known as the true negative rate, which indicates the fraction of correctly classified non-stress events:

$$Specificity = \frac{TN}{TN + FP}. \tag{2.6}$$

When classifiers provide probabilistic outputs rather than hard labels, it is possible to evaluate performance across thresholds by means of the ROC curve. This curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) for all possible thresholds, thereby showing the trade-off between sensitivity and false alarms (Figure 2.13). The AUC is a widely adopted scalar summary, ranging from 0.5 (random performance) to 1.0 (perfect discrimination). A higher AUC indicates a greater ability of the model to distinguish between classes independently of the chosen decision threshold [83].

It should be noted, however, that ROC analysis may become less informative in cases of highly imbalanced datasets. In such situations, precision–recall (PR) curves are often preferred, as they directly illustrate the trade-off between correctly identifying positive cases and limiting false alarms [73].

2.4 Related Work

Early studies established that psychophysiological stress can be inferred from multiple peripheral signals recorded in ecologically valid settings. A seminal naturalistic driving experiment by Healey and Picard (2005) instrumented n=24 drivers

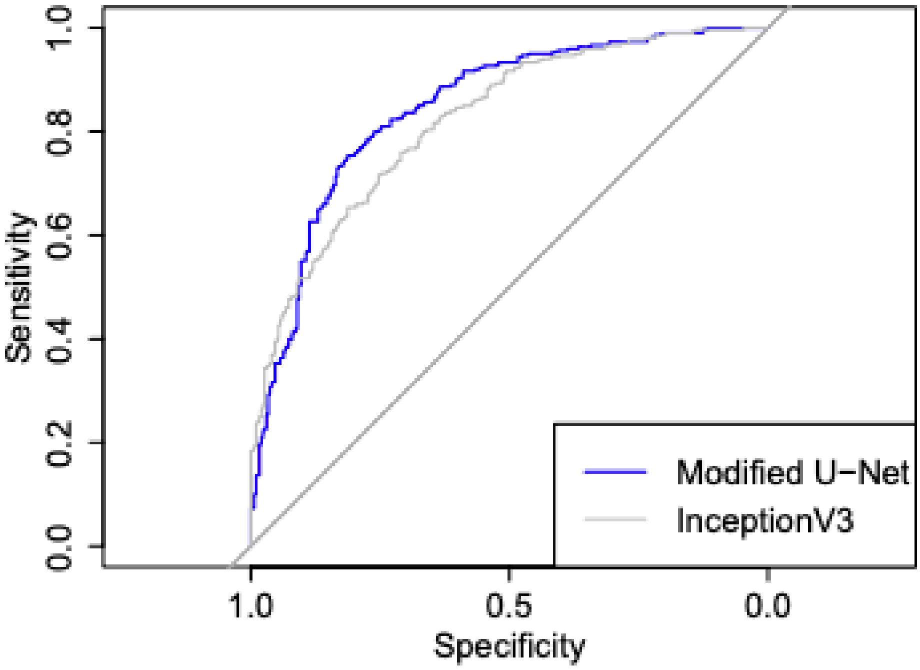


Figure 2.13: Example of ROC curves comparing two convolutional neural networks for binary classification. The AUC quantifies the discriminative power of the models [83].

with ECG, EMG, EDA, and respiration during ∼50 min on-road sessions; supervised classifiers trained on handcrafted descriptors (e.g., time-/frequency-domain HRV and EDA indices) discriminated low/medium/high driving stress, with multimodal fusion outperforming single-sensor models. The authors also documented core challenges—motion artefacts, context dependence, and label acquisition—that remain central to modern deployments [41]. In the early 2010s, affect-oriented corpora such as MAHNOB-HCI and DEAP broadened methodological foundations by combining peripheral physiology (ECG, EDA/GSR, respiration, skin temperature) with controlled audiovisual elicitation and self-reports; although designed for generic affect recognition, these datasets catalysed advances in feature engineering (e.g., HRV time–frequency measures, tonic/phasic EDA separation) and evaluation practices that later migrated to stress detection [97, 56]. SWELL-KW (2014) then introduced a knowledge-work paradigm closer to office reality: n=25 participants performed document-editing tasks under time pressure and email interruptions while ECG, EDA, posture, computer logging, and facial expressions were recorded; labels were provided through questionnaires and task design, offering a reproducible benchmark for work-related stressors [57].

With the maturation of wrist-worn devices, research shifted decisively toward ambulatory sensing and end-to-end pipelines evaluated beyond the lab. Hovsepian et al. (2015) proposed the *cStress* pipeline, specifying a full methodology—signal quality screening, filtering, feature computation and normalization, and supervised classification—validated both in controlled sessions (n=24) and in daily life over one week (n=20) using EMA as ground truth; the study reported high laboratory accuracy (around 90%) but substantially lower accuracy in the wild (around 70%), foregrounding the generalization gap from controlled to free-living conditions [46]. Complementarily, Gjoreski and colleagues demonstrated continuous stress detection with wrist wearables across lab and real-life contexts. In a two-stage design, a random-forest detector on short windows was followed by an SVM that ingested recent detector outputs together with contextual signals (activity, posture, phone usage); the addition of context markedly improved performance (reporting in-lab accuracy around 83% and up to ∼92% with context over longer horizons), highlighting that behaviour-aware features can reduce false positives outside the lab [35, 34].

The release of WESAD (2018) marked a watershed for reproducible benchmarking in stress detection. The dataset offers synchronized chest- and wrist-worn signals (ECG, EDA, BVP, respiration, ACC) across baseline, TSST social stress, and amusement, with questionnaire-based labels. WESAD rapidly became a default platform to compare preprocessing choices (e.g., HRV detrending/band-pass, EDA phasic decomposition), classical feature-driven models versus end-to-end deep architectures, and wrist-only versus chest+wrist configurations [91]. Alongside such canonical corpora, broad surveys systematized the method space and distilled empirical regularities. Can and colleagues (2019) reviewed laboratory and daily-life studies with smartphones and wearables, proposing a nomenclature to classify experimental designs (lab–lab, lab–daily, daily–daily) and synthesizing that HRV is a dominant information source; they also emphasized that accuracy typically ranges 70–90% in laboratory environments but often falls to 60–80% in unconstrained daily life where behavioural context and motion confound physiological cues [20]. A detailed systematic review by Vos et al. (2022) reinforced these findings, attributing the *in vivo* performance drop to small cohorts, heterogeneous stressors and protocols, and overfitting to dataset-specific artefacts [110]. Lazarou and Exarchos (2024) focused specifically on real-time pipelines on wearables, underscoring temporal alignment across channels, label noise (e.g., uncertain episode onsets), and latency/energy constraints for on-device inference as key obstacles to robust deployment [62].

Beyond in-domain evaluation, recent studies explicitly interrogate *generalizability* across populations and protocols. Benchekroun et al. (2023) trained HRV-based classifiers on one cohort and tested on another with different recording conditions and devices; a RF achieved cross-dataset macro-F1 of about 0.61, providing evidence of partial portability while clearly lagging behind within-dataset validation. The analysis further examined which HRV descriptors (time-domain variability vs. non-linear indices) are more stable across datasets [11]. Prajod et al. (2024) compared multiple public datasets (including WESAD, SWELL-KW, ForDigitStress, and VerBIO) and demonstrated that *stressor type* is a dominant determinant of transferability: models trained on datasets eliciting social stress generalize substantially better to similarly elicited datasets than to those driven by cognitive/arithmetic tasks; when stressor modalities are mismatched, F1 degrades markedly [79]. Finally, Ladakis et al. (2025) proposed an integrative training strategy that harmonizes features across open datasets before model fitting; they reported >80% accuracy on several tasks, exceeding 90% with pooled training, and around 70% when testing on unseen participants/datasets, suggesting that harmonization and data pooling can mitigate, but not eliminate, domain shift [61].

Taken together, the literature traces a coherent trajectory: from early multimodal evidence in ambulatory driving and controlled affect elicitation, through systems and datasets enabling reproducible comparisons in office-like and laboratory settings, to recent studies that scrutinize external validity with cross-dataset testing and harmonization. Across this trajectory, HRV and EDA remain central modalities; context signals (ACC, posture, usage) consistently improve real-world performance; and generalization beyond the source dataset emerges as the principal bottleneck, as repeatedly observed by surveys and explicit cross-dataset studies alike [20, 110, 62, 11, 79, 61].

2.5 Thesis Contribution

This thesis advances wearable-based stress detection along four tightly integrated axes, each designed to address known gaps in robustness and portability. *First*, a multi-dataset evaluation setting is established that deliberately spans both controlled laboratory protocols and more ecological, task-oriented recordings: four independent corpora are considered, two acquired under structured lab stressors and two gathered in realistically variable conditions. All four share the same wrist-worn sensing stack (electrodermal activity, BVP/HRV, skin temperature, and tri-axial accelerometry), allowing domain shift due to population and protocol to be isolated from hardware heterogeneity. *Second*, a uniform and reproducible preprocessing and curation workflow is applied *consistently* across all corpora—covering signal registry, resampling/synchronization, artifact-aware denoising, segmentation, label alignment, and subject-aware partitioning—so that observed differences originate from modeling choices rather than dataset-specific handling. *Third*, a head-to-head comparison is conducted between two dominant modeling paradigms under strictly matched evaluation: a *features-driven* pipeline (handcrafted physiological descriptors per modality coupled with classical ML) and a *data-driven* pipeline (end-to-end CNN architectures operating on minimally processed multi-channel time series). Each family is instantiated with strong baselines and tuned within its own design space to avoid single-model bias. *Fourth*, external validity is probed systematically via cross-dataset evaluation in two regimes: (i) *zero-shot* cross-testing, where models trained on a source corpus are evaluated as-is on a distinct target corpus; and (ii) *transfer learning*, where source-trained deep models are adapted to the target by freezing early layers and fine-tuning higher layers using limited target supervision. This design quantifies degradation under distribution shift, the recovery afforded by light adaptation, and the relative resilience of domain-knowledge features versus learned representations. Taken together, these choices deliver a modality-consistent, multi-dataset benchmark with harmonized preprocessing and explicit cross-dataset testing and transfer, offering reproducible evidence about when handcrafted features prevail, when learned representations better tolerate shift, and how much adaptation is needed to operate on unseen data.

To make these contributions operational and to guide the empirical study, the goals above are translated into targeted research questions and corresponding hypotheses that state expectations explicitly prior to evaluation:

- **RQ1:** Which modeling approach—features-driven (handcrafted features + classical ML) or data-driven (CNN over raw/cleaned signals)—achieves better *within-dataset* performance under Leave-One-Subject-Out (LOSO) evaluation?

- **RQ2:** Which approach generalizes more robustly *across datasets* (i.e., when applied to unseen datasets)?

- **RQ3:** Does transfer learning (freezing early layers, fine-tuning on a small amount of target data) improve cross-dataset performance compared to pure zero-shot cross-testing?
- **RQ4:** What are the key failure modes and limits in cross-dataset generalization (e.g., stressor mismatch, context/motion variability, annotation noise, class imbalance), and how do they differentially affect the two pipelines?

From these questions the following hypotheses are posed:

- **H1:** The features-driven approach will yield stronger in-domain performance (under LOSO) than the data-driven CNN in many datasets, due to explicit signal-specific priors and reduced sample complexity.
- **H2:** The data-driven (CNN) approach will suffer less degradation across datasets—i.e., will generalize more robustly—by learning flexible, transferable representations.
- **H3:** Transfer learning (fine-tuning) from a source dataset's best model to a target dataset will improve cross-dataset performance relative to direct zero-shot prediction.
- **H4:** Cross-dataset generalization will vary substantially with domain shift factors (stressor type, sensor placement, demographics), and any advantage of data-driven models will be modulated by the magnitude of this shift.

Chapter 3

Materials and Methods

A single experimental workflow was designed to systematically compare two paradigms for stress recognition from non-invasive, multimodal wrist-worn signals—namely a features-driven approach and a data-driven approach—while enforcing strict subject-wise separation to minimize information leakage. All experiments rely on four public datasets collected with the same wearable device (Empatica E4) and the same channels (electrodermal activity, blood volume pulse, skin temperature, and tri-axial accelerometry). After selecting the datasets to maximize protocol heterogeneity under sensor homogeneity and auditing their file structures, phases/timing, and native labels, raw EDA, BVP, TEMP, and ACC were ingested per subject (and session, when present), resampled onto a common 64 Hz grid, and accelerometry was aggregated into a single magnitude (ACC_MAG) to compactly index movement. Channel-specific denoising and artifact handling followed (e.g., Butterworth and notch filtering where appropriate), with visual checks performed by comparing raw versus cleaned traces. Signals were then segmented into fixed 45 s windows with 75% overlap, and each window received a single label—either the dominant label within the window or the label at onset—yielding window-level examples. In the features-driven branch, every window was transformed into signal-specific, hand-crafted descriptors (for example, tonic/phasic EDA indices, HR/HRV statistics derived from BVP, temperature trends, and activity indices from ACC_MAG), and a multi-stage selection retained only informative and robust variables by removing features with excessive missingness, pruning highly collinear variables, checking distributional assumptions, and assessing class discriminability with explicit attention to subject interaction; the surviving variables defined a subject–window feature matrix with labels and identifiers. In the data-driven branch, multivariate windows (EDA, BVP, TEMP, ACC_MAG) were provided directly to one-dimensional convolutional neural networks so that representation learning was performed by the model without hand-crafted features. For each dataset and for both branches, models were trained and evaluated with leave-one-subject-out cross-validation while preventing leakage by fitting any parameter-learning operation (such as imputation values, scaling statistics, and thresholds or rules used during feature selection) using training subjects only within each fold and then applying them to the held-out subject. For every dataset and every ML/DL family considered, the best LOSO instance according to macro-F1 was retained and its artefacts persisted (selected feature set, scaler, and weights when applicable) to enable subsequent cross-dataset and transfer-learning comparisons across multiple contenders. Cross-dataset generalization and transfer were then examined as follows: in the features-driven branch, datasets were aligned to a common subset of features and the persisted best models (with their scalers) were cross-tested; for neural models such as MLPs, transfer was explored by freezing early layers and fine-tuning the classifier head. In the data-driven branch, the best CNNs were cross-tested and analogously fine-tuned by partially freezing layers. Finally, the two paradigms were compared in terms of generalization on unseen datasets and implications for real-world stress monitoring.

3.1 Data

The experimental phase relies on four publicly available multimodal datasets: WESAD, Campanella et al., AffectiveROAD, and VERBIO. The selection was guided by the need for comparability and for a rigorous assessment of cross-dataset generalization. All corpora were acquired with the Empatica E4 wristband, ensuring homogeneous hardware and signal characteristics across studies. Each dataset provides the same set of wrist-worn physiological channels—electrodermal activity (EDA), blood volume pulse (BVP), skin temperature (TEMP), and tri-axial accelerometry (ACC). At the same time, the datasets differ in experimental protocols and stress induction procedures, offering a diverse range of conditions for evaluating model robustness and generalization. All resources are publicly accessible, anonymized, and distributed under research-oriented licenses.

WESAD Dataset WESAD (Wearable Stress and Affect Detection) is a controlled laboratory dataset collected on 15 healthy graduate students (12 males, 3 females; mean age 27.5 ± 2.4 years) using a dual-device setup: a chest-worn *RespiBAN Professional* (hub sampling all channels at 700 Hz) and a wrist-worn *Empatica E4* placed on the non-dominant hand. The chest platform recorded ACC and RESP natively and, via analog ports, ECG (three-lead), EDA (electrodes on the rectus abdominis, chosen for its high sweat gland density), EMG (bilateral upper trapezius), and TEMP (sternum). The E4 streamed BVP (64 Hz), EDA (4 Hz), TEMP (4 Hz), and 3-axis ACC (32 Hz). Devices were synchronised at the beginning and end of the protocol through a double-tap gesture; data were stored locally to avoid wireless packet loss and subsequently offloaded for analysis. In this thesis we exclusively used the wrist (E4) signals to ensure hardware homogeneity across all datasets [91].

Participants were instructed to abstain from caffeine and tobacco for at least one hour prior to the session and from strenuous exercise on the same day. After sensor instrumentation and initial synchronisation, the study aimed to elicit three affective states—*baseline (neutral)*, *stress*, and *amusement*—with interleaved *meditation* phases to restore arousal. The two stimulus blocks (stress vs. amusement) were counterbalanced across subjects (Version A and Version B) to mitigate order effects; in addition, within each block, approximately half of the subjects performed while standing and the other half while sitting to induce postural variance. The complete protocol lasted approximately two hours and included the following phases: (a) Baseline: a 20-minute neutral reading task at a desk; (b) Amusement condition: eleven humorous video clips interleaved by five-second neutral intervals, for a total duration of 392 seconds; (c) Stress condition (TSST): a modified TSST including a 3-minute silent preparation, a 5-minute impromptu speech on personal strengths and weaknesses delivered in front of a three-person panel (introduced as HR specialists), and a 5-minute serial subtraction task (counting backwards by 17 from 2023, restarting upon mistakes); (d) Meditation: a guided breathing exercise of 7 minutes conducted after each stimulation block; (e) Recovery: a final 10-minute rest following the TSST and debriefing.

During the experiment, self-reports were collected at five time points (indicated in red in Figure 3.1), including PANAS, STAI items, and SAM Scale. Additionally, after the TSST, participants completed the Short Stress State Questionnaire (SSSQ) to characterize stress type [91].

The official WESAD dataset provides condition labels sampled at 700 Hz from the chest device timeline. To obtain labels aligned with wrist data, we downsampled these labels to 64 Hz (BVP sampling rate) and synchronised them with the E4 signals. All E4 modalities (BVP 64 Hz, EDA/TEMP 4 Hz, ACC 32 Hz) were resampled to a uniform 64 Hz grid to allow windowing and multimodal fusion. Consistently with the binary stress-detection goal of this thesis, we retained only baseline and stress segments, excluding amusement and meditation periods from both training and evaluation [91].

Campanella et al. Dataset The Campanella et al. corpus [18] was acquired in a controlled laboratory setting on 29 healthy participants using a single wrist-worn *Empatica E4* device. The bracelet, worn on the non-dominant wrist, provided continuous electrodermal activity (EDA, 4 Hz), blood volume pulse (BVP, 64 Hz),

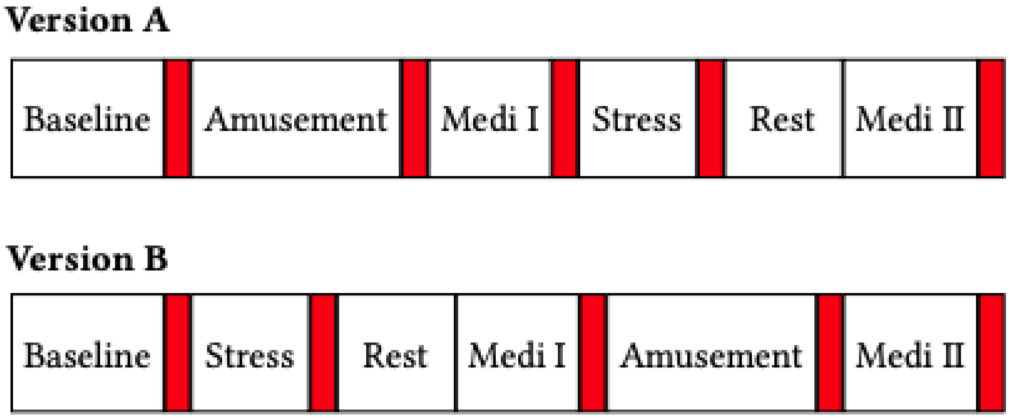


Figure 3.1: WESAD experimental protocol (Version A/B). The study included baseline, amusement, TSST-based stress, and meditation periods. Red markers indicate time points of self-reports [91].

skin temperature (TEMP, 4 Hz), and tri-axial accelerometry (ACC, 32 Hz). Signals were streamed via Bluetooth to a companion smartphone application and subsequently uploaded to the *Empatica E4 Connect* cloud workspace for storage and export. In this thesis, only the E4 channels were employed to maintain hardware homogeneity across datasets and to align with the wrist-centric focus adopted throughout.

The experimental design reproduced a demanding work-like environment by interleaving rest periods with heterogeneous stressors spanning cognitive, social, and combined modalities. The full session lasted approximately 37 minutes and comprised six tasks separated by short recoveries, as summarised in Figure 3.2. After an initial 3-minute rest to stabilise physiology, participants assembled a Lego structure for 10 minutes relying solely on the box images (no instructions), thereby inducing problem-solving under time pressure and manual dexterity demands. A brief 2-minute recovery followed, after which the same assembly was repeated for 5 minutes with access to instructions, reducing but not removing cognitive load. Following another 2-minute recovery, a 3-minute dual-task was introduced: assembly with larger Lego pieces while performing backward counting from 180 to 0, thereby combining manual activity and arithmetic stress. After a further 2-minute recovery, a purely cognitive stressor inspired by the MIST was administered with no fixed time limit, consisting of repeated subtraction of 13 from 511 to sustain mental workload. A final 2-minute recovery preceded a 1-minute social stressor in which participants delivered a brief oral self-presentation of their CV to an examiner, concluding with a short cool-down.

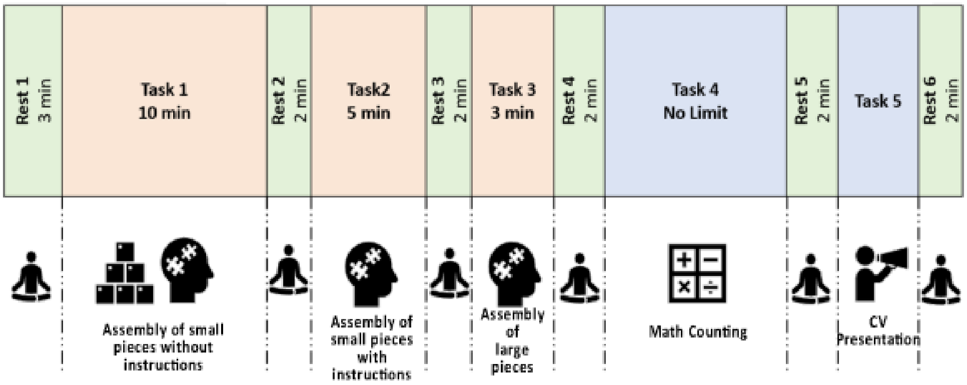


Figure 3.2: Experimental protocol of the Campanella et al. dataset. Alternating rest and heterogeneous stressors (cognitive, combined manual+cognitive, and social) are administered within a 37-minute laboratory session [18].

Label assignment followed the authors' schedule. The initial 27 minutes covering Tasks 1–3 with their interleaved recoveries were annotated as alternating baseline and stress in accordance with the momentary task demands. From minute 28 up to five minutes before the end of the session, recordings were labelled as sustained stress corresponding to the prolonged cognitive arithmetic phase. The final five minutes (social speech followed by recovery) were annotated with a fixed binary pattern reflecting the brief stressor embedded within a short cool-down. Within this thesis, the original binary taxonomy was preserved: baseline encompassed all rest intervals and low-demand conditions, whereas stress comprised the cognitive, social, and combined stress-inducing tasks. For integration into the common analysis pipeline, all wrist modalities (BVP 64 Hz; EDA/TEMP 4 Hz; ACC 32 Hz) were resampled onto a uniform 64 Hz grid, labels were synchronised to the E4 timeline, and only baseline vs. stress segments were retained to match the binary detection objective.

AffectiveROAD Dataset AffectiveROAD [27, 17] was designed to study the relationship between driving context and driver stress under naturalistic road conditions. Data were collected in real traffic in the Greater Tunis area along a 31 km route lasting approximately 85 minutes (including resting periods), alternating city centres, highways, suburban zones, and parking areas. The corpus integrates wearable physiological recordings, contextual road information, and subjective annotations of perceived stress. Each driver wore an *Empatica E4* wristband that continuously measured EDA (4 Hz), BVP (64 Hz), skin temperature (4 Hz), and tri-axial accelerometry (32 Hz). A frontal dashboard camera captured the roadway at 25 fps and an internal camera recorded the driver's face and behaviour; GPS and environmental data were logged to provide context such as location, time, and traffic density. A trained experimenter sat in the rear seat with a laptop showing the front camera feed and produced continuous annotations of perceived stress during the drive. All device streams were synchronised offline and stored in a unified dataset [17].

Recording sessions were conducted under typical daylight and dry weather conditions. Drivers followed the predefined 31 km route comprising segments intended to elicit different cognitive loads and stress levels. Very low-stress manoeuvring occurred in parking and Z-Zone areas; highway sections involved steady speeds (about 80–100 km/h) with limited interaction and corresponded to medium stress; urban segments featured dense traffic, frequent stops, and complex interactions with pedestrians and intersections, inducing higher stress; transitions such as U-turns and lane changes linked the segments; a second city portion replicated urban complexity while accumulating fatigue and sustaining elevated stress. Conversations with the experimenter were maintained to simulate everyday driving. The public release includes 13 complete sessions (9 unique participants, with some repeated runs), totalling approximately 676 minutes of synchronised video and biosignals [17]. An example route overview is shown in Figure 3.3.

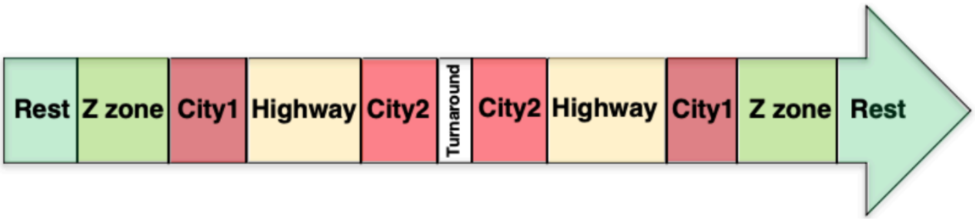


Figure 3.3: Example of the AffectiveROAD driving route in the Greater Tunis area. The path alternates between low-stress (parking, highway) and high-stress (urban) segments [27].

The stress ground truth, termed the human-driver-provided stress signal, was constructed as a continuous trace in [0,1] from the rear-seat experimenter's real-time slider annotation, reflecting both scene complexity and perceived driver tension. After each session, the driver reviewed synchronised internal and external videos and was allowed to validate or edit the trace; the final signal thus represents a consensus between observer and driver. Following [17], this continuous score was originally discretised into three classes (low: $0.0 \leq s < 0.4$, medium: $0.4 \leq s < 0.75$, high: $s \geq 0.75$). In this thesis the problem was adapted to a binary setting by thresholding at $s = 0.75$, assigning samples with $s > 0.75$ to stress and all others to baseline. Because the annotation frequency was variable and typically below 1 Hz, nearest-neighbour replication was used to upsample the stress trace to 64 Hz and align it with the physiological sampling grid; this preserved temporal coherence without introducing interpolation artefacts and ensured comparability with the other datasets. The subset analysed includes the first recording session for each of the 13 available runs to mitigate habituation effects and retains realistic variation in roadway conditions, traffic complexity, and driver behaviour, making AffectiveROAD one of the most ecologically valid multimodal resources for driver stress analysis to date.

VERBIO Dataset VERBIO is a multimodal corpus targeting stress detection during public speaking under both real and virtual reality exposure. Recordings integrate *Empatica E4* wrist signals—BVP (64 Hz), EDA (4 Hz), skin temperature (4 Hz), and tri-axial accelerometry (32 Hz)—together with chest ECG from an *Actiwave Cardio* device (512 Hz) and 16 kHz audio of the speech signal; VR stimuli were delivered via an Oculus Rift using the *Virtual Orator* platform [59, 114]. The user study spanned approximately five months and comprised ten sessions per participant organised across four days: PRE (day 1, real audience), TEST01–TEST08 (days 2–3, eight VR sessions), and POST (day 4, real audience). The cohort initially included 55 university students (18–30 years; near-balanced gender); 36 completed the TEST series and 29 completed POST. Each session followed a fixed three-phase structure with RELAX (5 min, a soothing nature video to establish baseline), PREP (10 min, silent preparation on a randomly assigned news article), and PPT (5 min, oral presentation). PRE and POST were conducted in a conference room with an audience of about five persons instructed to keep a neutral demeanour, whereas TEST used VR audiences. Across the eight TEST sessions, each speaker was randomly assigned eight of twelve VR configurations that factorially varied room type (board room, classroom, small theatre, seminar room), audience size (12, 25, 54, 90), and audience reaction (negative, neutral, positive), with low-level ambient classroom noise presented over headphones to increase ecological validity [114].

The public release provides, for every session and phase (RELAX, PREP, PPT), E4 time series (ACC, BVP, EDA, HR, IBI, TEMP) and Actiwave ECG/HR files, with PPT files including absolute timestamps synchronised to the audio stream; the folder structure is standardised by session (PRE, TEST01–TEST08, POST) and participant, with consistent CSV schemas to facilitate downstream processing [59]. Ground truth is distributed not only as self-reports but also as continuous stress annotations: four independent raters (R1, R2, R4, R5) scored perceived stress on a Likert scale 1–5 during the PPT phase; scores were mapped to $[0, 0.25, 0.5, 0.75, 1]$ and aggregated into a fused trace at 1 Hz, time-aligned with audio.

In this thesis only wrist (E4) channels were used to ensure hardware homogeneity across datasets. All streams were aligned on a common 64 Hz timeline by resampling the physiological channels and by nearest-neighbour replication of the fused annotation trace. Label semantics were harmonised to a binary scheme consistent with the cross-dataset stress detection objective. Specifically, RELAX segments were assigned to the baseline class, while PPT segments were assigned to stress if the fused continuous rating s satisfied $s \geq 0.20$, otherwise baseline. The 0.20 threshold was selected after sensitivity checks to balance classes while reflecting low-to-moderate perceived stress. This configuration preserves the original session structure (PRE and POST with real audiences; TEST with controlled VR scenarios) and supports integration into the uniform preprocessing and evaluation pipeline adopted throughout this work [59, 114].

3.2 Preprocessing

All datasets were harmonized under a unified ingestion and synchronization framework centered on the multimodal data streams recorded by the *Empatica E4* device, including EDA, BVP, SKT, and the three-axis Accelerometer (ACC$_x$, ACC$_y$, ACC$_z$). Raw data were organized *per subject* (and *per session* when available), with all signal streams mapped into a canonical tabular structure featuring standardized headers and explicit identifiers such as `subject`, `session`, and `phase`. Dataset-specific labels—either natively provided or reconstructed according to experimental protocols—were imported and linked to the corresponding temporal grid.

To enable joint multimodal analysis, all channels were resampled and synchronized on a common temporal grid at 64Hz, corresponding to the highest native sampling rate among the E4 channels (BVP). This design choice ensured that no high-frequency information was lost, particularly from the BVP signal, which contains fine-grained cardiovascular oscillations in the range of 0.5–10 Hz. Performing a downsampling to match lower-rate signals (e.g., EDA and TEMP at 4 Hz, ACC at 32 Hz) would have led to aliasing and loss of critical detail in these high-frequency components. Conversely, the *upsampling* of slower signals to 64 Hz preserves all original information while facilitating temporal alignment across modalities without degrading any channel.

Upsampling was achieved through interpolation schemes appropriate to the signal dynamics: a zero-order hold (ZOH) or nearest-neighbor replication for slowly varying channels (EDA and TEMP), and linear interpolation for motion data (ACC). Prior to synchronization, low-pass anti-alias filtering was applied where necessary to avoid spectral distortion. Labels sampled at non-uniform or low frequencies (e.g., self-assessed annotations or experiment phase markers) were upsampled through forward-fill or discrete interpolation to ensure the presence of a valid class label for each timestamp, supporting downstream windowing and feature extraction.

Given the tri-axial nature of accelerometric data, the three spatial components were aggregated into a single magnitude channel to reduce redundancy and emphasize the global motion intensity. The resulting signal, referred to as `ACC_MAG`, was computed as:

$$ACC_{\text{MAG}} = \sqrt{ACC_x^2 + ACC_y^2 + ACC_z^2}$$

This scalar representation serves as a compact proxy for the overall movement level, simplifying the subsequent analysis of motion-related artefacts and ensuring comparability across datasets. The unified registry produced at this stage consists of a fully synchronized, sample-synchronous corpus—each row representing a unique timestamp at 64Hz—with consistent feature naming, subject linkage, optional session metadata, and corresponding stress-state labels.

Table 3.1: Post-ingestion summary of the unified multimodal registries aligned to a common 64 Hz grid.

Dataset	Rows	Columns	Subjects	Label source
WESAD	1,765,022	6	15	Task-based
VERBIO	2,159,612	8	17	Rater-based
CAMPANELLA	4,149,992	7	29	Task-based
AffectiveROAD	5,693,652	6	13	Rater-based

Denoising and Artifact Handling The denoising and artifact handling stage constitutes a fundamental step of the preprocessing pipeline, aiming to enhance the quality and reliability of the physiological signals prior to feature extraction. Physiological data acquired through wearable sensors are often affected by motion artifacts, power-line interference, and baseline drifts, which can significantly distort the waveform morphology and compromise the extraction of meaningful physiological information. Therefore, signal cleaning is essential to ensure that downstream analyses—such as feature computation and stress classification—are based on physiologically valid and interpretable data.

In this work, the denoising process was applied uniformly to all datasets and to all modalities (*BVP*, *EDA*, *TEMP*, and *ACC*), ensuring methodological consistency and reproducibility across subjects and experimental protocols. The main objectives were to remove high-frequency noise components, suppress power-line interference, and mitigate artifacts due to abrupt movements or sensor instabilities.

To this end, a hybrid approach combining classical digital signal processing and domain-specific cleaning routines was adopted. Traditional filtering techniques—implemented through the `SciPy` library—were used to design Butterworth low-pass and band-pass filters with cutoff frequencies tailored to each signal's spectral content. In parallel, the `NeuroKit2` toolkit was employed for its advanced, physiologically-informed denoising algorithms. NeuroKit2 provides modular and optimized functions for cleaning and preprocessing biosignals such as PPG (Blood Volume Pulse), EDA, ECG, and EMG. Its filtering routines combine adaptive detrending, zero-phase Butterworth filtering, and artifact detection strategies, ensuring waveform preservation while attenuating noise and spurious fluctuations. The resulting cleaned signals are physiologically coherent, allowing the extraction of robust time- and frequency-domain features for subsequent analysis.

Blood Volume Pulse (BVP) Signal The BVP signal, recorded at a native sampling rate of 64Hz, represents one of the most noise-sensitive modalities due to its susceptibility to motion and pressure artifacts at the wrist. The cleaning procedure was performed using the dedicated `NeuroKit2` PPG processing module, which applies an adaptive band-pass filter designed to preserve the characteristic pulsatile components of the signal—typically within the frequency range of 0.5–8 Hz—while removing both low-frequency drifts and high-frequency noise. The filter operates in zero-phase mode to avoid phase distortion and maintain the temporal integrity of systolic peaks.

In addition to filtering, the algorithm performs automatic artifact detection based on peak morphology and inter-beat interval regularity. Outliers and implausible peaks are identified and corrected using local interpolation, ensuring the preservation of a physiologically consistent waveform. This process enables accurate reconstruction of clean pulsatile dynamics and allows reliable computation of HR and HRV parameters in subsequent analytical stages.

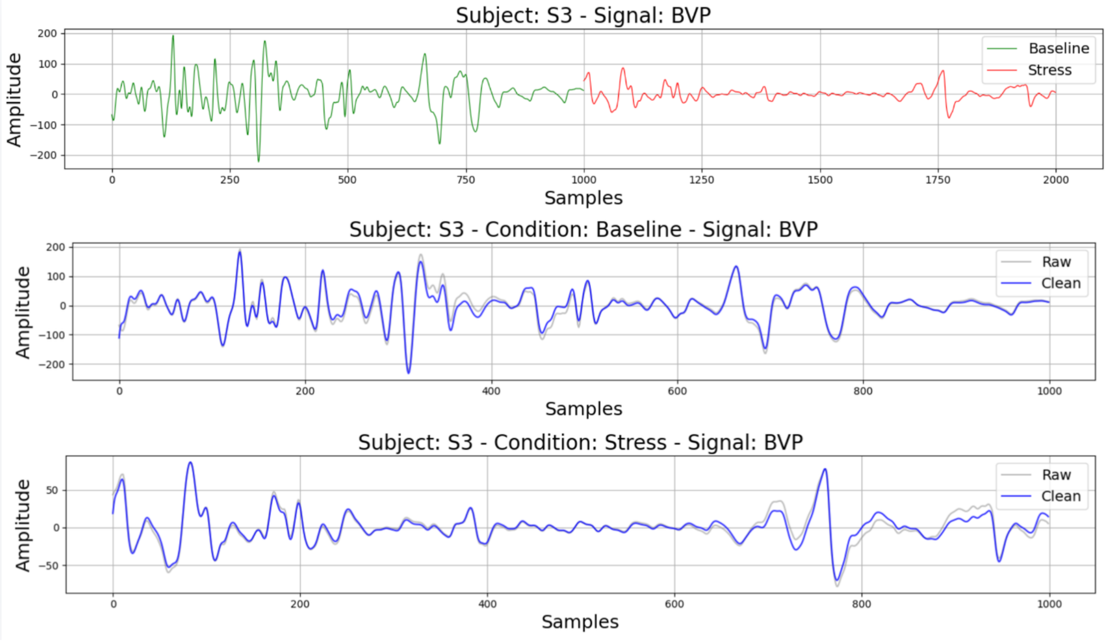


Figure 3.4: Comparison between raw and cleaned BVP signal for a representative subject.

Electrodermal Activity (EDA) Signal The EDA signal, acquired at a native sampling rate of 4Hz and resampled to 64Hz for synchronization with the other modalities, reflects variations in skin conductance caused by sympathetic nervous system activation. EDA is a particularly informative marker of stress, but it is also highly sensitive to artifacts resulting from movement, temperature fluctuations, or abrupt sensor contact changes. These artifacts can distort the slow-varying tonic baseline and introduce spurious peaks in the phasic component, leading to incorrect estimation of physiological arousal.

To ensure reliable feature extraction, the EDA signals were preprocessed using the `nk.eda_process()` function from the `NeuroKit2` library [69], which provides a complete and physiologically grounded pipeline for EDA denoising, decomposition, and feature derivation. The cleaning routine involves several sequential steps designed to isolate the true skin conductance response (SCR) activity while suppressing noise and motion-induced fluctuations.

Initially, the raw signal undergoes detrending to remove slow drifts in baseline conductance, followed by low-pass Butterworth filtering—typically with a cutoff around 5 Hz—to eliminate high-frequency components unrelated to electrodermal dynamics. The filter operates in zero-phase mode to preserve waveform morphology and avoid phase shifts. After filtering, `NeuroKit2` applies an adaptive smoothing algorithm and identifies outliers based on signal amplitude and derivative thresholds, ensuring that abrupt discontinuities or saturation artifacts are corrected.

Once the signal is cleaned, it is decomposed into two physiological components: the *tonic* component (Skin Conductance Level, SCL) representing the slow baseline conductance trend, and the *phasic* component (Skin Conductance Response, SCR) capturing rapid, transient changes linked to sympathetic activation. This decomposition enables the extraction of meaningful temporal and amplitude-based features (e.g., number, amplitude, and rise time of SCR peaks) that directly relate to stress-induced autonomic responses.

Accelerometry (ACC_MAG) Signal The aggregated accelerometry magnitude signal (`ACC_MAG`) was filtered via a two-stage Butterworth band-pass scheme to suppress low-frequency drift and high-frequency noise. Specifically, a high-pass Butterworth filter with cutoff at 0.5Hz was first applied to remove very slow baseline wander and gravitational offset—phenomena that are not informative for movement detection in our stress-monitoring context. Then, a low-pass Butterworth filter with cutoff at 20Hz was used to attenuate higher-frequency noise components (e.g. sensor electronics noise or high-frequency vibration artifacts) that are unlikely to contain relevant motion content.

The choice of these cutoff frequencies is supported by literature: for example, some wearable sensor pipelines adopt a high-pass filter at 0.5 Hz to mitigate drift effects [67], while studies on accelerometer-based activity measurement show that upper cutoffs around 20 Hz help suppress noise while preserving meaningful motion signals [30]. The double filtering approach was implemented with zero-phase

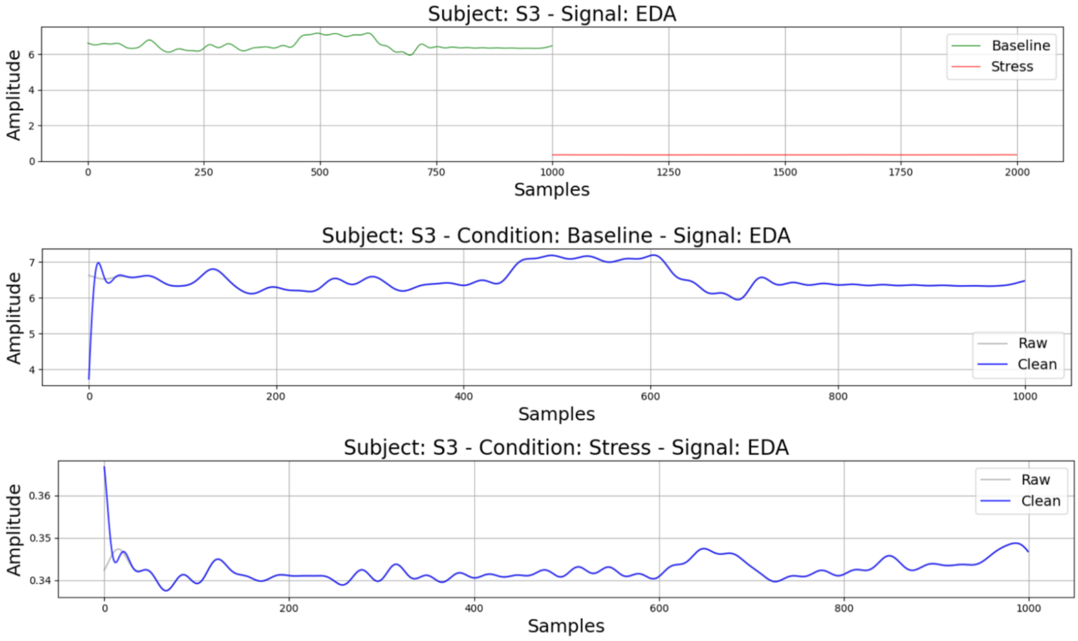


Figure 3.5: Comparison between raw and cleaned EDA signal for a representative subject.

forward–backward filtering, thereby avoiding phase distortion and preserving temporal alignment of peaks. The output is a cleaned accelerometry magnitude signal that emphasizes the band of interest for human movement detection.

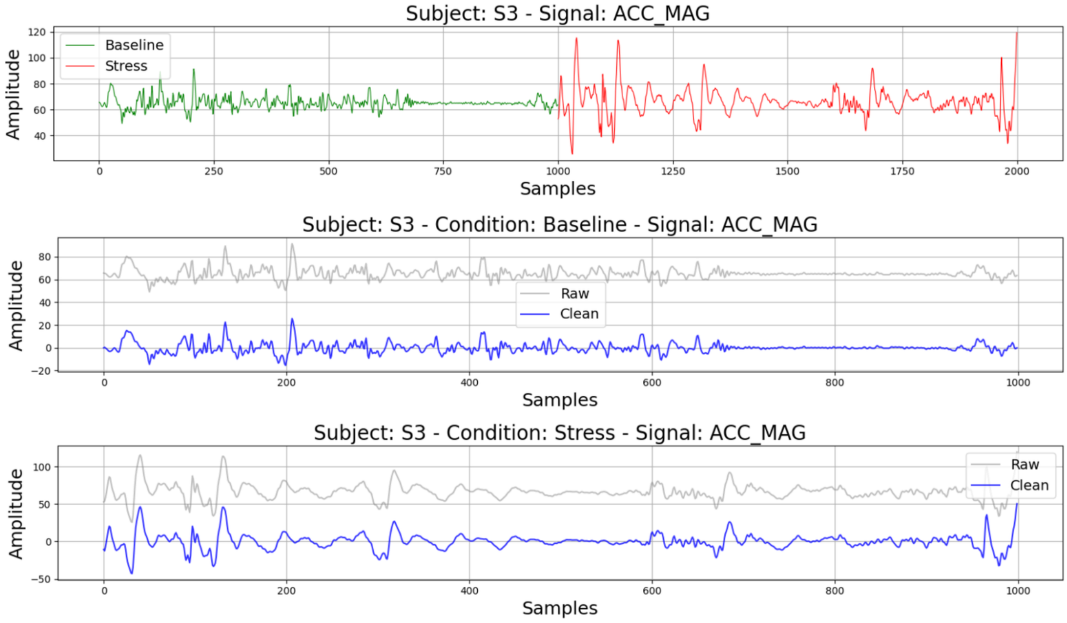


Figure 3.6: Comparison between raw and cleaned ACC signal for a representative subject.

Skin Temperature (TEMP) Signal The SKT signal—resampled to 64Hz—was denoised by applying a single-stage Butterworth low-pass filter with cutoff at 0.5Hz. This filtering step aims to suppress high-frequency perturbations (e.g. sensor noise, abrupt spikes) while preserving the slowly varying thermoregulatory dynamics. The filtering was performed via zero-phase forward–backward processing (e.g. using `filtfilt`) to avoid introducing phase shifts that could misalign the temperature trend relative to other modalities.

The cutoff at 0.5Hz was selected based on the physiological and thermal inertia of skin temperature measurements: human skin acts as a natural low-pass filter for thermal waves, smoothing fast fluctuations and limiting significant dynamics to very low frequencies (typically below 0.5 Hz) [87]. Because temperature variations at the wrist evolve slowly (on the order of seconds to minutes), applying a high-pass filter would risk removing relevant baseline drift or slow trends linked to ambient or physiological thermal changes. Hence, no high-pass stage was employed.

This minimalist filtering approach ensures that the cleaned temperature signal retains meaningful slow variations for subsequent feature extraction (e.g. gradients or temporal trends) while discarding high-frequency noise that could corrupt derivative-based analyses.

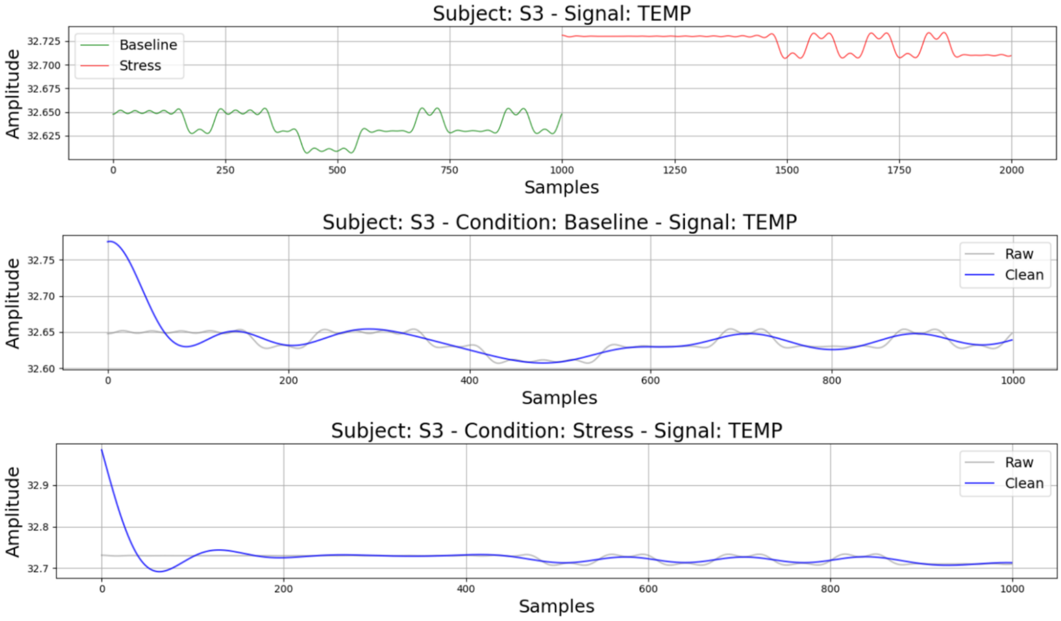


Figure 3.7: Comparison between raw and cleaned TEMP signal for a representative subject.

Windowing and Window-Level Label Assignment Following the denoising stage, all signals were segmented into fixed-length, partially overlapping windows to enable localized feature extraction and consistent labeling. This step allows transforming continuous physiological time series into discrete instances suitable for supervised learning while maintaining temporal continuity across samples.

Each signal was divided into windows of 45s with a 75 overlap. This configuration was selected after empirical testing of multiple combinations of window sizes and overlaps, as it provided the best trade-off between feature stability and sensitivity to short-term stress-related variations. Shorter windows introduced excessive variability, whereas longer ones tended to smooth out transient physiological responses.

For each window, a label was assigned using the *dominant-label rule*, which attributes the window to the most frequent class within it. To ensure label reliability, only windows with a dominant class proportion of at least 70% were retained, while ambiguous segments were discarded. This procedure minimizes labeling noise and enhances the robustness of subsequent model training.

The same windowing and labeling strategy was applied uniformly across all datasets, resulting in a consistent window-level dataset where each instance corresponds to a 45s segment associated with its dominant label, subject identifier, and experimental phase.

Feature Extraction Once all signals were cleaned and temporally segmented, the next stage of the pipeline consisted in the extraction of quantitative features from each window. This process transforms the time-domain signals into a compact set of numerical descriptors that capture relevant physiological dynamics associated with stress responses. The same extraction procedure was applied consistently across all datasets to ensure comparability and reproducibility.

Scope. This step applies *only* to the feature–driven (FD) pipeline. In the data–driven (DD) branch, no handcrafted descriptors are computed: after windowing and label assignment, raw windows (one channel per sensor, shaped as $(\text{n\_windows}, \text{n\_samples}, 1)$) are passed directly to the CNN for configuration, training, and tuning.

Each feature vector corresponds to a 45s window (with 75 overlap) and includes both time-domain and, where applicable, frequency-domain descriptors. The specific features extracted from each signal depend on its physiological nature and information content. For the Blood Volume Pulse (BVP) and Electrodermal Activity (EDA) signals, advanced physiological metrics were derived through the `NeuroKit2` toolbox, while for accelerometry (ACC) and skin temperature (TEMP) signals, statistical features were computed using standard numerical operations (`NumPy` and `SciPy`). The following subsections detail the extraction strategy for each signal modality.

Blood Volume Pulse (BVP) *FD only.* For the BVP signal, features were extracted from the cleaned waveform using `NeuroKit2` and `SciPy` routines. Peak detection relied on `NeuroKit2`'s `ppg_peaks()` function to identify individual pulse cycles, from which HR and HRV indices were computed via `nk.hrv()`. The HRV set comprised standard time-domain metrics such as SDNN, RMSSD, and Percentage of NN intervals differing by more than 50 ms (pNN50), together with frequency-domain parameters including LF, HF, and the LF/HF ratio. In addition, complementary time-domain descriptors were derived with `SciPy`: the mean peak amplitude within the window, representing average pulse strength; the mean rise time, defined as the time between pulse onset and peak; and the average inter-beat interval (IBI), corresponding to the mean period between consecutive BVP peaks. These features jointly characterize cardiovascular reactivity and the temporal structure of the BVP signal, offering complementary perspectives on autonomic dynamics.

Electrodermal Activity (EDA) *FD only.* For the EDA signal, feature extraction was performed with `NeuroKit2` using `nk.eda_peaks()` and `nk.eda_intervalrelated()`, which provide a comprehensive set of tonic and phasic descriptors. Within each 45 s window, the algorithm detects skin conductance responses (SCRs) and computes interval-based features including the number of SCRs, the mean SCR amplitude and latency, the SCL as an estimate of the tonic component, and the AUC of phasic responses. These measures capture sympathetic arousal by quantifying both the frequency and the magnitude of electrodermal fluctuations.

Accelerometry (ACC) *FD only.* For the accelerometry channel, represented by the magnitude of the three orthogonal axes (ACC_MAG), descriptive statistics were computed over each window using `NumPy` and `SciPy`. The feature set includes mean, standard deviation, minimum, maximum, range, median, interquartile range (IQR), skewness, and kurtosis, thereby quantifying the intensity and variability of wrist movements and enabling an indirect assessment of activity level and potential motion artifacts. The same computation scheme was applied uniformly across subjects and datasets to ensure a consistent numerical representation of movement dynamics.

Skin Temperature (TEMP) *FD only.* For the skin temperature channel, a statistical approach analogous to that used for accelerometry was adopted. From each window, mean, standard deviation, minimum, and maximum temperature values were extracted, together with the local slope estimated by linear regression as an indicator of short-term thermal change. Given the slow-varying nature of skin temperature, these descriptors capture both the absolute thermal level and local trends that may reflect thermoregulatory responses under stress or rest. As with the other signals, the extraction strategy was applied identically across all datasets to preserve cross-dataset consistency.

Feature Selection The feature selection process aims to retain informative, reliable, and non-redundant descriptors that contribute effectively to the discrimination between stress and baseline conditions, thereby improving model robustness and interpretability while reducing overfitting risk and computational burden. The same pipeline was applied uniformly across the four datasets, although the final subsets differ because of dataset-specific variability and signal characteristics. The procedure proceeds as follows: features with more than 50% missing values are removed and the remaining missing entries are imputed using the median; collinearity is reduced by excluding highly correlated variables according to a threshold of $|\rho| > 0.8$; distributional properties are examined through graphical inspection (histograms, boxplots, QQ-plots) and formally assessed with the Shapiro–Wilk test for normality; the subject effect and its interaction with the class label are evaluated via three-way Analysis of Variance (ANOVA); finally, discriminative power is quantified using the Mann–Whitney U test to compare stress versus baseline on a subject-wise basis. At the end of this process, a distinct set of selected features is obtained for each dataset, representing the most discriminative and statistically robust subset of descriptors for subsequent classification analysis.

Step 1: Handling of Missing Values The first step of the feature selection pipeline focused on managing incomplete or missing data across the extracted feature matrices. Each signal-specific feature table was first cleaned and harmonized by replacing non-finite values with `NaN` and computing the percentage of missing entries for each feature.

All features exhibiting a proportion of missing values greater than 50% were removed, as such high levels of missingness compromise the statistical reliability of subsequent analyses. For the remaining features, missing entries were replaced through median imputation, chosen for its robustness to outliers and non-normal distributions. Median-based imputation preserves the central tendency of the data without being excessively influenced by extreme values, which is particularly relevant in physiological datasets where distributions often deviate from Gaussianity.

This preprocessing ensured that the following correlation and hypothesis-testing stages operated on complete, statistically coherent feature matrices, minimizing bias induced by incomplete observations.

Step 2: Collinearity Analysis and Removal After addressing missing data, the next stage focused on the identification and removal of highly correlated features. Collinearity arises when two or more variables convey overlapping information, which can distort statistical inference, inflate variance in model coefficients, and reduce the interpretability of the resulting models. In the context of physiological signal analysis, where several descriptors may capture similar aspects of signal dynamics (e.g., variability, amplitude, or dispersion), addressing collinearity is essential to retain only unique and non-redundant predictors.

The procedure consisted of computing the pairwise Pearson correlation matrix across all numerical features within each signal-specific dataset. For each pair of features with an absolute correlation coefficient exceeding a threshold of $|\rho| > 0.8$, one of the two variables was automatically discarded according to a deterministic criterion: the first feature encountered in the matrix order was retained, while the subsequent feature showing correlation above the threshold with any of the previously retained ones was removed. This approach ensures a one-pass, rule-based pruning without the need for manual inspection or subjective decisions.

The choice of the threshold 0.8 was based on common practices in multivariate statistical analysis, balancing the need to reduce redundancy while preserving sufficient feature diversity. This value is widely accepted in biosignal processing literature as a suitable cutoff for detecting strong linear associations without overpruning the feature set.

As a result of this step, each signal-specific feature matrix was reduced to a subset of variables exhibiting low pairwise correlations, ensuring statistical independence among retained features and improving model stability in the subsequent classification phase.

Step 3: Distributional Analysis and Subject–Label Interaction Following the removal of redundant features, a statistical and graphical examination of feature distributions was conducted to characterize their shape, assess normality, and evaluate the potential influence of the subject on the stress label. This step was essential to determine the most appropriate statistical framework for subsequent analyses and to identify features whose variability might depend on subject-specific factors rather than experimental condition.

Distributional analysis. For each feature, the empirical distribution was visually inspected using histograms with superimposed kernel density estimates, boxplots, and quantile–quantile (QQ) plots. These visualizations provided an intuitive understanding of symmetry, skewness, and the presence of outliers within each feature. An illustrative example of two representative features is shown in Figure 3.8, highlighting the difference between an approximately Gaussian feature and another exhibiting strong deviation from normality.

Normality testing. In addition to graphical inspection, the *Shapiro–Wilk test* was applied to each feature to formally assess the assumption of normality. The null hypothesis assumed that the feature followed a Gaussian distribution; features with p-values below the significance level $\alpha = 0.05$ were considered non-normally distributed. The majority of features violated the normality assumption, confirming the non-Gaussian nature of most physiological descriptors—particularly those derived from EDA and BVP signals, which tend to exhibit skewed distributions and heavy tails.

Subject–label interaction. Based on these findings, an *Analysis of Variance (ANOVA)* was performed for each feature to evaluate the independent effects of the label (stress vs. baseline), the subject, and their interaction. The model can be expressed as:

$$\text{Feature} \sim C(\text{Label}) + C(\text{Subject}) + C(\text{Label}) : C(\text{Subject})$$

This approach allowed quantifying the proportion of variance explained by experimental condition, inter-subject differences, and their interaction. Overall, this analysis confirmed that a considerable number of features displayed significant

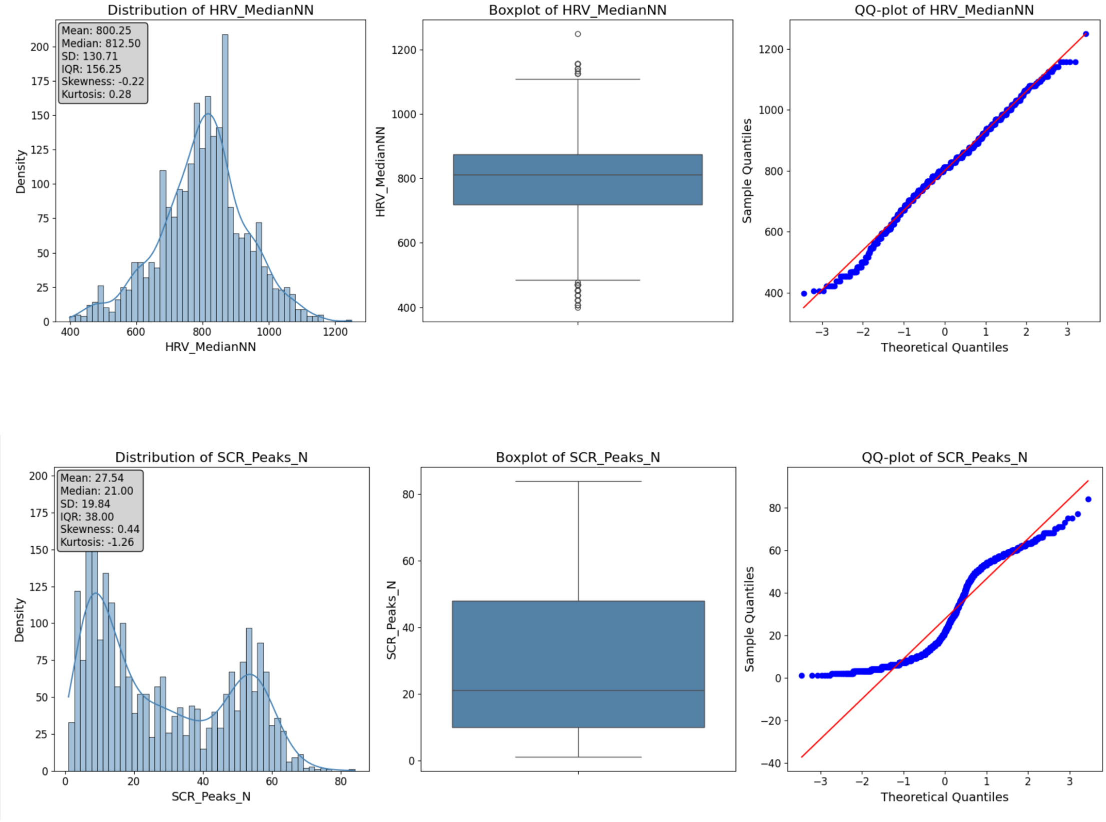


Figure 3.8: Example of distributional plots for two representative features: histogram with density curve (left) and QQ-plot (right). The plots illustrate typical deviations from normality observed in physiological features.

variability across subjects and non-Gaussian distributions, indicating the need for subject-wise non-parametric testing in the final discriminative evaluation stage.

Step 4: Discriminative Testing (Mann–Whitney U Test) The final stage of the feature selection process aimed to identify features that exhibited statistically significant differences between the two experimental conditions—*stress* (label 1) and *no stress* (label 0). This analysis was performed through a combination of graphical exploration and statistical hypothesis testing to evaluate the discriminative power of each feature.

For every feature, boxplots were generated to visualize its distribution under the two experimental conditions. This graphical comparison allowed for an intuitive understanding of whether the central tendency and dispersion differed between stress and baseline windows. An illustrative example of two representative features is reported in Figure 3.9, showing the variability between the two conditions across multiple subjects.

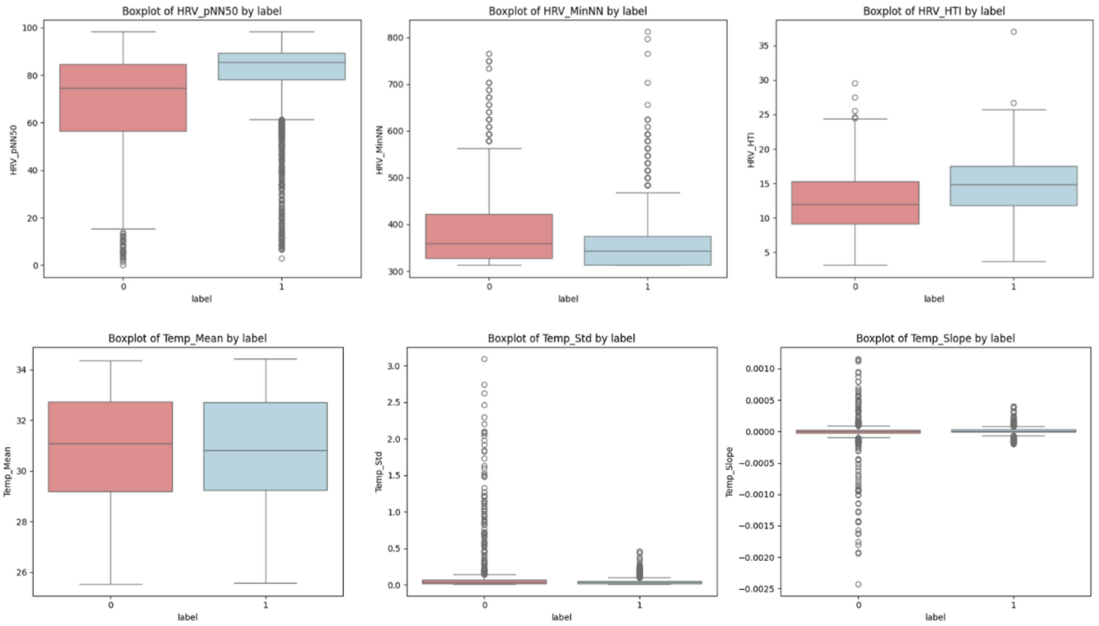


Figure 3.9: Example of boxplots for six representative features across the two experimental conditions: baseline (label 0) and stress (label 1).

Following the graphical analysis, the *Mann–Whitney U test* (also known as the Wilcoxon rank-sum test) was applied to quantify the statistical difference between the two conditions. This non-parametric test was chosen based on the results of the previous step, which revealed that the majority of features did not follow a normal distribution. Unlike parametric tests such as the Student's *t*-test, the

Mann–Whitney U test does not assume normality and is therefore more appropriate for the type of data considered in this study.

Each feature was tested independently for every subject, comparing the samples corresponding to stress and baseline windows. This subject-wise testing strategy was motivated by the results of the ANOVA analysis (Step 3), which demonstrated that the interaction between subject and label was statistically significant in most cases. Conducting the analysis at the individual level thus prevented subject-specific variability from confounding the overall results and allowed identifying features that consistently discriminated between conditions across participants.

For each subject, the null hypothesis stated that the distributions of the feature under the two conditions were equal; features with p-values below the significance level $\alpha = 0.05$ were considered discriminative for that subject. At the group level, a retention criterion was applied: each feature was discarded if it failed to reach statistical significance in more than a predefined number of subjects. The threshold for this criterion varied across datasets, being proportional to the number of participants in each experimental protocol—thus maintaining comparable statistical rigor despite differing sample sizes.

The outcome of this step was a refined set of features that showed consistent and statistically significant differences between stress and baseline conditions across the majority of subjects. These features represent the most discriminative subset within each dataset and were retained to form the final input space for classification.

Final Integration and Summary of Selected Features At the end of the feature selection pipeline, each dataset produced four distinct signal-specific feature matrices—one for each modality (*BVP*, *EDA*, *ACC*, and *TEMP*). Each matrix contained only the subset of features that passed all selection criteria, i.e., those that were complete, non-collinear, and statistically discriminative between stress and baseline conditions.

To build the final dataset used as input for the subsequent machine learning and deep learning models, these four matrices were aligned and merged into a single multimodal feature table. Prior to concatenation, the individual matrices were truncated to the same number of windows (corresponding to the shortest signal among the four modalities for that dataset) and re-indexed to ensure temporal alignment. Metadata columns such as `label`, `subject`, and `protocol_phase` were retained only once, avoiding redundancy after the merge.

A summary of the number and type of surviving features for each signal and dataset is reported in Table 3.2, highlighting the subset of descriptors deemed statistically relevant after the complete selection process.

Table 3.2: Surviving features after the selection pipeline, grouped by dataset and signal.

(a) BVP (HRV/PPG)

Dataset	**Features**
WESAD	MeanNN, SDNN, MedianNN, MadNN, SDRMSSD, Prc20NN, pNN50, MinNN, HTI, TINN, MFDFA_alpha1_Max, MFDFA_alpha1_Fluctuation, SampEn, FuzzyEn, MSEn, CD, HFD, KFD, LZC, Amplitude, Duration
Campanella	MadNN, Prc20NN, pNN50, HTI, MFDFA_alpha1_Peak, FuzzyEn, Amplitude, Duration
VerBIO	SDNN, MedianNN, MadNN, Prc20NN, pNN50, HTI, FuzzyEn, KFD, Amplitude
AffectiveROAD	MeanNN, SDNN, SDRMSSD, MinNN, VHF, S, PIP, PI, SampEn, FuzzyEn, MSEn, KFD, LZC, Amplitude

(b) EDA (SCR)

Dataset	**Features**
WESAD	Peaks_N, Peaks_Amplitude_Mean
Campanella	Peaks_N, Peaks_Amplitude_Mean
VerBIO	Peaks_N, Peaks_Amplitude_Mean
AffectiveROAD	Peaks_N, Peaks_Amplitude_Mean

(c) ACC_MAG

Dataset	**Features**
WESAD	Mean, Std, IQR, Skew, Kurtosis
Campanella	Mean, Std, Max, IQR, Skew
VerBIO	Mean, Std, Median, Skew, Kurtosis
AffectiveROAD	Mean, Std, Max, Skew, Kurtosis

(d) TEMP

Dataset	**Features**
WESAD	Mean, Std, Slope
Campanella	—
VerBIO	Slope
AffectiveROAD	Std, Slope

Learning Pipeline After preprocessing (windowing, feature extraction, and feature selection), the resulting tabular dataset—where rows represent subject–window instances and columns the retained features from all signals—is used to train binary stress versus no-stress classifiers. Generalisation across subjects is assessed with a LOSO protocol: in each fold, all windows from one subject form the test set while the remaining subjects constitute the training set; any preprocessing and model selection are fitted exclusively on the training split and then applied unchanged to the held-out subject, and predictions across folds are concatenated for global summaries. Six models are evaluated—LR, RF, GB, XGBoost, Support Vector Machine with RBF kernel, and a shallow Multi-Layer Perceptron—with reporting that includes confusion matrix, ROC/AUC when applicable, a full classification report, and model-specific feature importance. All transformations are performed within each LOSO fold to prevent leakage. In the features-driven branch, tabular features undergo a min–max scaling to [0,1], with the scaler fitted on training data and reused on the held-out subject; in the data-driven branch, raw windowed time series are reshaped to $(T,1)$, scaled independently per signal by a training-fold min–max transform, and transformed identically for the held-out subject, without imposing distributional assumptions such as zero-mean standardisation. Class imbalance is controlled without resampling by means of weighting computed per fold on the training labels only: learners that accept class weights use `class_weight=balanced`; XGBoost compensates the prior via `scale_pos_weight` set to $N_{\text{neg}}/N_{\text{pos}}$ for the current training split; algorithms lacking native weighting are trained on the original distribution. For convolutional models in the data-driven branch, imbalance is handled through per-fold class weights passed to the optimiser and by adopting a focal-loss objective to reduce the dominance of easy negatives. Decision thresholds for discrete predictions are selected for each held-out subject by maximising Youden's J on the ROC derived from that fold's scores.

3.3 Classification algorithms

Logistic Regression A penalised LR with ℓ_2 regularisation (default C=1.0; solver as library default) is employed to produce calibrated probabilities for the binary stress vs. no-stress task. Within each LOSO fold, the model is trained on min–max scaled features using `class_weight="balanced"` to compensate class prevalence and an increased iteration budget (`max_iter`=1000) to ensure convergence; probabilities from `predict_proba` enable ROC/AUC computation. Post hoc interpretation relies on the signed coefficients learned within each fold: under per-fold scaling to [0,1], larger absolute values indicate stronger monotonic association with the log-odds of the positive class, while the sign encodes directionality.

Random Forest A RF classifier is configured with `n_estimators`=100, `criterion`=`gini`, `max_depth`=None, `min_samples_split`=2, `min_samples_leaf`=1, and `max_features`=$\sqrt{p}$ to promote tree decorrelation. Class imbalance is compensated within each LOSO training fold via `class_weight`=balanced. Probabilistic outputs from `predict_proba` are used for ROC/AUC, and model interpretability relies on impurity-based feature importances (`feature_importances_`) computed per fold.

Gradient Boosting A GB classifier is configured with `n_estimators`=100, `learning_rate`=0.1, and base learners of depth `max_depth`=3. Other parameters are kept at library defaults. Class distribution is left unchanged (no native class weights), and probabilistic outputs are used for ROC/AUC. This shallow, slow-learning setup balances expressiveness and overfitting control across LOSO folds.

XGBoost An eXtreme Gradient Boosting (XGB) classifier is used with `n_estimators`=100, `learning_rate`=0.1, `max_depth`=3, `subsample`=0.8, `colsample_bytree`=0.8, `use_label_encoder`=False, `eval_metric`=`logloss`, and `verbosity`=0. Training is carried out within each LOSO fold on min–max scaled inputs; the original class prevalence is retained (no explicit class weighting). Probabilistic outputs from `predict_proba` are employed to compute ROC/AUC, while optimisation during training follows the specified log-loss metric.

Support Vector Machine A Support Vector Classifier with radial basis kernel is configured as `kernel`=`rbf`, `C`=1.0, `gamma`=`scale`, `class_weight`=`balanced`, and `probability`=True. Training is performed within each LOSO fold on min–max scaled features; probabilistic outputs from `predict_proba` are used to compute ROC/AUC. No additional tuning is applied so as to maintain comparability across datasets and avoid overfitting on small per-fold training sets.

Multi-Layer Perceptron A feed-forward MLP is trained on the tabular feature set within each LOSO fold using min–max scaled inputs. The network comprises Dense layers of sizes 256–128–64–32–16, with `BatchNormalization` after the first Dense, `LeakyReLU` activations (negative slope $\alpha = 0.1$) throughout, and `Dropout` regularisation applied with rates 0.4, 0.3, 0.2, and 0.1 on successive hidden blocks. The output layer is a single neuron with `sigmoid` activation for binary classification. The model is compiled with the `adam` optimiser and `binary_crossentropy` loss, reporting accuracy during training. Class weights are not applied; probabilities from the sigmoid output are used to compute ROC/AUC on the held-out subject. Training/validation data (including any internal validation split) are drawn exclusively from the training partition of each LOSO fold.

Convolutional Neural Network The data-driven branch employs a compact 1D multi-input CNN trained directly on windowed raw signals. Each physiological modality (EDA, BVP, TEMP, ACC axes) is processed by an independent stream that performs hierarchical temporal feature extraction; the learned representations are subsequently fused and fed to a shallow classifier.

In each per-signal stream, temporal patterns are extracted through three successive convolutional blocks that progressively shrink the receptive field while increasing representational capacity. Block 1 targets coarse dynamics using a Conv1D layer with 32 filters and a wider kernel, followed by batch normalisation, a ReLU nonlinearity, max temporal pooling for downsampling, and SpatialDropout1D with rate 0.2. Block 2 focuses on intermediate patterns via a Conv1D with 64 filters and a medium kernel, again followed by batch normalisation, ReLU, max pooling, and SpatialDropout1D with rate 0.3. Block 3 captures fine-grained temporal structure through a narrower-kernel Conv1D with 128 filters, batch normalisation, ReLU, max pooling, and SpatialDropout1D with rate 0.3. All convolutional layers use He initialisation and ℓ_2 weight decay with $\lambda = 10^{-3}$. To obtain a fixed-length descriptor from variable-length feature maps, each stream ends with global average pooling.

Late fusion is performed by concatenating the pooled descriptors from all streams into a single representation. The classifier head consists of two dense blocks: the first applies a fully connected layer with 64 units, batch normalisation, ReLU activation, and dropout 0.4 with ℓ_2 regularisation ($\lambda = 10^{-3}$); the second mirrors this design with 32 units and dropout 0.3. The output layer is a single-unit dense layer with sigmoid activation that returns the probability of the positive (stress) class.

Training is formulated as a class-imbalanced optimisation problem. The objective is focal loss with focusing parameter $\gamma = 2.0$ and class-balancing factor $\alpha = 0.25$, emphasising hard and underrepresented positives. Optimisation uses Adam with an initial learning rate 10^{-3}. Per-fold class weights computed from the training labels are passed to the loss to further counter residual imbalance. A conservative training schedule is adopted for comparability across datasets: early stopping on validation loss with patience = 5 and best-weight restoration, ReduceLROnPlateau on validation loss with factor 0.7, patience = 3, and a floor of 10^{-6}, plus model checkpointing on the best validation loss; batch size and maximum epochs are kept moderate and fixed across experiments, and best weights are reloaded at the end of each fold.

Preprocessing is performed within each LOSO fold to prevent leakage. For every signal, min–max scaling is fitted on the training partition and applied unchanged to the held-out subject. At inference, probabilistic scores on the test subject are converted to discrete labels by selecting the ROC operating point that maximises (TPR − FPR) on the validation split within the fold, yielding a data-driven threshold rather than a fixed 0.5 cutoff.

Reproducibility is ensured by persisting, for each fold, the best-performing model checkpoint together with the per-signal scaling transformers fitted on the training split. This guarantees deterministic, leakage-free preprocessing at inference time and supports portability to external datasets in cross-test and transfer-learning experiments. In line with the study's focus on end-to-end performance under distribution shift, no feature-importance visualisations are reported for the CNN and the data-driven branch is treated as non-interpretable.

Evaluation Methodology Models are evaluated under a Leave-One-Subject-Out (LOSO) protocol to assess generalisation to unseen individuals. For each fold, one subject is held out for testing and the remaining subjects constitute the training set; all data transformations and any hyperparameter selection are fitted exclusively on the training partition. After inference on the held-out subject, predictions and scores are stored, and the process is repeated until every subject has served as test. Fold-wise predictions are then concatenated to enable global assessments.

For each held-out subject, a standard summary is produced reporting accuracy, precision, recall, and F1-score. These per-subject figures are then aggregated across folds, yielding LOSO means and standard deviations to quantify between-subject variability. Complementing the per-subject view, a global classification report is computed on the concatenated predictions, providing class-wise precision, recall, and F1, as well as overall accuracy, macro averages, and weighted averages. A 2×2 confusion matrix on the aggregated predictions offers a compact view of error types. To characterise threshold-independent behaviour, ROC curves and the corresponding AUC are derived from the concatenated fold scores (probabilities or decision scores, depending on the model). A per-subject F1 visualisation is additionally generated, displaying each subject's F1 alongside the LOSO mean as reference to highlight dispersion and potential outliers.

When applicable, model-native feature importance is reported to aid interpretation: linear models are summarised via absolute coefficient magnitudes, while tree ensembles and gradient-boosted trees are summarised via impurity- or gain-based importances. For models without intrinsic importance (such as RBF-SVMs or MLPs), permutation importance may be computed on validation data drawn strictly from the training fold to avoid leakage; fold-wise importances are summarised (e.g., via the median) to obtain a stable ranking. All computations adopt the same fold partitions used for training to ensure methodological consistency and comparability across models and datasets.

3.4 Cross–Dataset and Transfer Learning

This section details the experimental methodology adopted to assess the *out-of-domain* generalisation of stress classifiers across multiple datasets and to evaluate whether *transfer learning* mitigates the performance drop observed under distribution shift. The protocol extends the within-dataset Leave–One–Subject–Out (LOSO) scheme previously described by explicitly testing models on *unseen datasets* (pure zero-shot) and by performing a lightweight fine-tuning on the target domain where specified. The rationale and study design follow the methodological framework outlined in the Methods chapter (uniform preprocessing, LOSO splitting, and consistent metrics), with the present section focusing exclusively on cross-dataset and transfer procedures. Let $\mathcal{D} = \{\mathcal{D}^{(1)}, \dots, \mathcal{D}^{(4)}\}$ denote the datasets considered, each comprising multimodal wearable signals (EDA, BVP, TEMP, ACC) processed under a common windowing scheme. For a given *source* dataset $\mathcal{D}^{(s)}$ and a distinct *target* dataset $\mathcal{D}^{(t)} \neq \mathcal{D}^{(s)}$, we denote by $f^{(s)}$ the model selected on $\mathcal{D}^{(s)}$ under LOSO (best fold by macro F1), together with the associated feature/scale transformers defined in the in-domain pipeline. LOSO is used to ensure subject-level independence during model selection, a standard choice in affect recognition from wearables.

Features–Driven Pipeline Cross-dataset inference with classical learners is enabled by expressing all datasets in a shared semantic feature space. The reference is the feature subset produced by the WESAD preprocessing pipeline, adopted as benchmark for completeness and consistency; non-WESAD datasets are re-extracted by retaining, per modality, exactly the WESAD-defined features, while WESAD preserves its native set. This harmonisation guarantees identical input dimensionality and meaning across datasets and constitutes a prerequisite for zero-shot transfer with tabular models. For each dataset and each algorithm (LR, RF, GB, XGBoost, RBF–SVM, feed-forward MLP), LOSO training is performed and the best model by macro F1 is retained together with the fitted scaler or normaliser, with probabilistic outputs preserved to support subsequent ROC/AUC analyses. In the pure cross-test setting, given $(\mathcal{D}^{(s)}, \mathcal{D}^{(t)})$, the source model $f^{(s)}$ is applied to $\mathcal{D}^{(t)}$ without any retraining; input windows from $\mathcal{D}^{(t)}$ are transformed using the source preprocessing (scalers and, where applicable, per-feature normalisation) to preserve the training distribution, and predictions are evaluated with the metrics defined in the Methods chapter—macro F1 as primary, with accuracy, precision, recall, and ROC–AUC as secondary—yielding a zero-shot estimate of cross-domain generalisability. Transfer learning for the features-driven MLP adopts a light fine-tuning protocol on the target domain to adapt the decision surface while preserving the source representation: all hidden layers are frozen and only the final classification layer remains trainable; inputs from $\mathcal{D}^{(t)}$ are transformed with the source scaler; optimisation uses Adam with gradient clipping (`clipnorm`=1.0), binary cross-entropy loss, a low learning rate $\eta = 10^{-4}$, mini-batches of size 16, and at most 10 epochs; early stopping monitors validation loss with patience = 5 and `restore_best_weights`=True, with a minimum improvement $\Delta = 10^{-3}$; a learning-rate scheduler (`ReduceLROnPlateau`) halves the rate on stagnation (factor 0.5, patience = 3, floor 10^{-7}). To mitigate imbalance on the target training split, balanced class weights are applied when both classes are present, and LOSO partitioning on $\mathcal{D}^{(t)}$ is always respected; unless otherwise stated, a fixed decision threshold of 0.5 is used at inference to maintain comparability with zero-shot cross tests, and folds whose target test portion contains a single class are skipped to avoid degenerate estimates. Two diagnostic comparisons complement standard confusion matrices and classification reports: per-algorithm ΔF1, defined as the difference between in-domain LOSO F1 on $\mathcal{D}^{(s)}$ and zero-shot F1 on $\mathcal{D}^{(t)}$, which quantifies generalisation loss; and ROC-curve overlays across algorithms on the same target, which characterise relative discrimination in the cross-dataset regime.

Data–Driven Pipeline (CNN over Raw/Cleaned Signals) Within each dataset, convolutional networks are trained under a Leave-One-Subject-Out (LOSO) protocol using identical input construction based on multi-stream windows per modality and uniform preprocessing. For in-domain selection, the best-performing checkpoint is retained together with the exact per-stream scaling used at training time, ensuring deterministic and leakage-free inference on external data. Cross-dataset testing follows a zero-shot protocol in which, for a source–target pair $(\mathcal{D}^{(s)}, \mathcal{D}^{(t)})$, the source-selected CNN is evaluated on $\mathcal{D}^{(t)}$ without any retraining. Target inputs are built with the same preprocessing as at source: signals are resampled to the common grid and windowed with fixed duration (45 s) and overlap (75%); the per-stream scaling learned on $\mathcal{D}^{(s)}$ (e.g., min–max to [0,1]) is reused on $\mathcal{D}^{(t)}$ without refitting; channel order, tensor shape, and padding/truncation rules strictly match the source model specification. The source operating point is preserved (fixed threshold τ=0.5 unless otherwise stated) to maintain comparability across domains. Target folds containing a single class in the test portion are skipped to avoid degenerate scoring; for evaluated folds, the same metrics defined in the Methods chapter are computed (macro–F1 as primary; accuracy, precision, recall, and ROC–AUC as secondary), with an explicit class-collapse diagnostic (all predictions in a single class) recorded for analysis. Transfer learning is performed by fine-tuning the source checkpoint on $\mathcal{D}^{(t)}$ under a conservative schedule designed to preserve backbone representations and limit overfitting and collapse: early convolutional blocks are frozen while only the terminal block and the classification head remain trainable; the source per-stream scalers are reused on target windows;

optimisation uses Adam with gradient clipping (clipnorm = 1.0) and binary cross-entropy at a low learning rate $\eta = 10^{-4}$, mini-batches of size 16, and an epoch cap of 10; early stopping monitors validation loss with patience = 5, minimum improvement $\Delta = 10^{-3}$, and `restore_best_weights`=True; a `ReduceLROnPlateau` scheduler halves the rate on stagnation (factor 0.5, patience = 3, floor 10^{-7}). When both classes are present in the target training split, balanced class weights are applied during optimisation; otherwise, no weighting is used. Fine-tuning respects the LOSO partition of $\mathcal{D}^{(t)}$ (training on target–training windows only) and evaluation on the corresponding target test fold uses a fixed threshold τ=0.5; as in zero-shot, single-class test folds are skipped and class-collapse is checked. For each target dataset, all best models trained on the remaining datasets are applied both in zero-shot cross test and, where defined, in transfer-learning mode, producing per-fold metrics, aggregated macro–F1, global confusion matrices, and ROC overlays. The combined use of LOSO for in-domain selection and explicit cross-dataset validation aligns with best practices in wearable affective computing and directly addresses generalisation challenges induced by dataset shift, including protocol heterogeneity, stressor mismatch, and label noise.

Chapter 4

Results and Discussion

4.1 Single-Dataset Evaluation

Table 4.1 reports, for each dataset, the LOSO performance of all classical models and the in-domain CNN, after applying a subject-level class-balance screening at five percent prevalence on the held-out fold. Metrics are macro-F1 (primary), macro-averaged precision, and macro-averaged recall, all expressed as mean ± standard deviation across subjects. Figure 4.1 summarises the ROC profiles of the classical, feature-driven models across all datasets, while Figure 4.2 illustrates the corresponding in-domain CNN ROC curves.

Two high-level patterns emerge after balance screening. First, CAMPANELLA and WESAD exhibit consistently higher scores (best macro-F1 in the ∼0.82–0.87 range), whereas AffectiveROAD and VerBIO remain more challenging (best macro-F1 around 0.60–0.63). Second, nonlinear learners tend to dominate, with gradient-boosted ensembles and RBF-SVM frequently leading among classical models, and with the CNN competitive or superior in three datasets. These trends align with study design and labelling strategies. WESAD comprises laboratory phases with labels by experimental condition; CAMPANELLA follows a similar structure with alternating rest and stress tasks and a final oral presentation. AffectiveROAD is collected in the wild during real driving and labelled by road segment difficulty; VerBIO adopts expert ratings subsequently reviewed with participants. Despite the five percent threshold, AffectiveROAD and VerBIO retain substantial inter-subject heterogeneity and residual imbalance, which translates into larger LOSO variances.

AffectiveROAD shows the greatest heterogeneity. The strongest classical result is reached by the RBF-SVM (macro-F1 0.626 ± 0.083). Several subjects display extreme skew and were removed by the threshold, yet the remaining set still shows heterogeneous driving contexts and motion artefacts. Precision generally exceeds

Table 4.1: LOSO results per dataset and algorithm after subject-level class-balance screening ($\tau = 5\%$). Primary metric: macro-F1 (mean ± SD).

Dataset	Algorithm	F1-score	Precision	Recall
AffectiveROAD	Logistic Regression	0.6182 ± 0.1303	0.6566 ± 0.1635	0.6365 ± 0.1849
	Random Forest	0.6195 ± 0.0674	0.6384 ± 0.1800	0.6106 ± 0.1109
	Gradient Boosting	0.5765 ± 0.1263	0.6278 ± 0.1943	0.5771 ± 0.1698
	XGBoost	0.6044 ± 0.0897	0.6506 ± 0.1781	0.6106 ± 0.1440
	SVM (RBF)	0.6264 ± 0.0830	0.6849 ± 0.1500	0.6017 ± 0.1149
	MLP	0.6005 ± 0.1222	0.7033 ± 0.1449	0.5509 ± 0.1596
	CNN	**0.6934±0.1089**	0.7692 ± 0.1027	0.6575 ± 0.1781
CAMPANELLA	Logistic Regression	0.7473 ± 0.1201	0.8162 ± 0.1122	0.7291 ± 0.1906
	Random Forest	0.8058 ± 0.0486	0.7688 ± 0.0721	0.8616 ± 0.1055
	Gradient Boosting	0.8050 ± 0.0516	0.7756 ± 0.0819	0.8542 ± 0.1101
	XGBoost	0.8117 ± 0.0550	0.7808 ± 0.0782	0.8612 ± 0.1113
	SVM (RBF)	0.7782 ± 0.0691	0.8184 ± 0.1003	0.7639 ± 0.1316
	MLP	**0.8185±0.0387**	0.7586 ± 0.0710	0.8999 ± 0.0786
	CNN	0.7434 ± 0.1910	0.8663 ± 0.0802	0.7011 ± 0.2163
VerBIO	Logistic Regression	0.6104 ± 0.2464	0.5752 ± 0.3051	0.7730 ± 0.2330
	Random Forest	0.5304 ± 0.1997	0.6611 ± 0.3288	0.5322 ± 0.2146
	Gradient Boosting	0.5877 ± 0.2432	0.6612 ± 0.3352	0.6166 ± 0.2283
	XGBoost	0.5889 ± 0.2205	0.6472 ± 0.3222	0.6366 ± 0.2213
	SVM (RBF)	0.6271 ± 0.2483	0.5650 ± 0.2950	0.8152 ± 0.2327
	MLP	0.5883 ± 0.2553	0.6082 ± 0.3146	0.6550 ± 0.2589
	CNN	**0.7418±0.1791**	0.7665 ± 0.1455	0.7767 ± 0.2501
WESAD	Logistic Regression	0.8063 ± 0.1128	0.7706 ± 0.1327	0.8831 ± 0.1668
	Random Forest	0.8204 ± 0.1816	0.8719 ± 0.1436	0.8225 ± 0.2234
	Gradient Boosting	0.8474 ± 0.1182	0.8446 ± 0.1606	0.8850 ± 0.1513
	XGBoost	0.8655 ± 0.1136	0.8681 ± 0.1574	0.8918 ± 0.1332
	SVM (RBF)	0.8241 ± 0.1146	0.7809 ± 0.1568	0.9131 ± 0.1491
	MLP	0.8459 ± 0.1355	0.8670 ± 0.1128	0.8588 ± 0.1915
	CNN	**0.9575±0.0841**	0.9879 ± 0.0238	0.9388 ± 0.1259

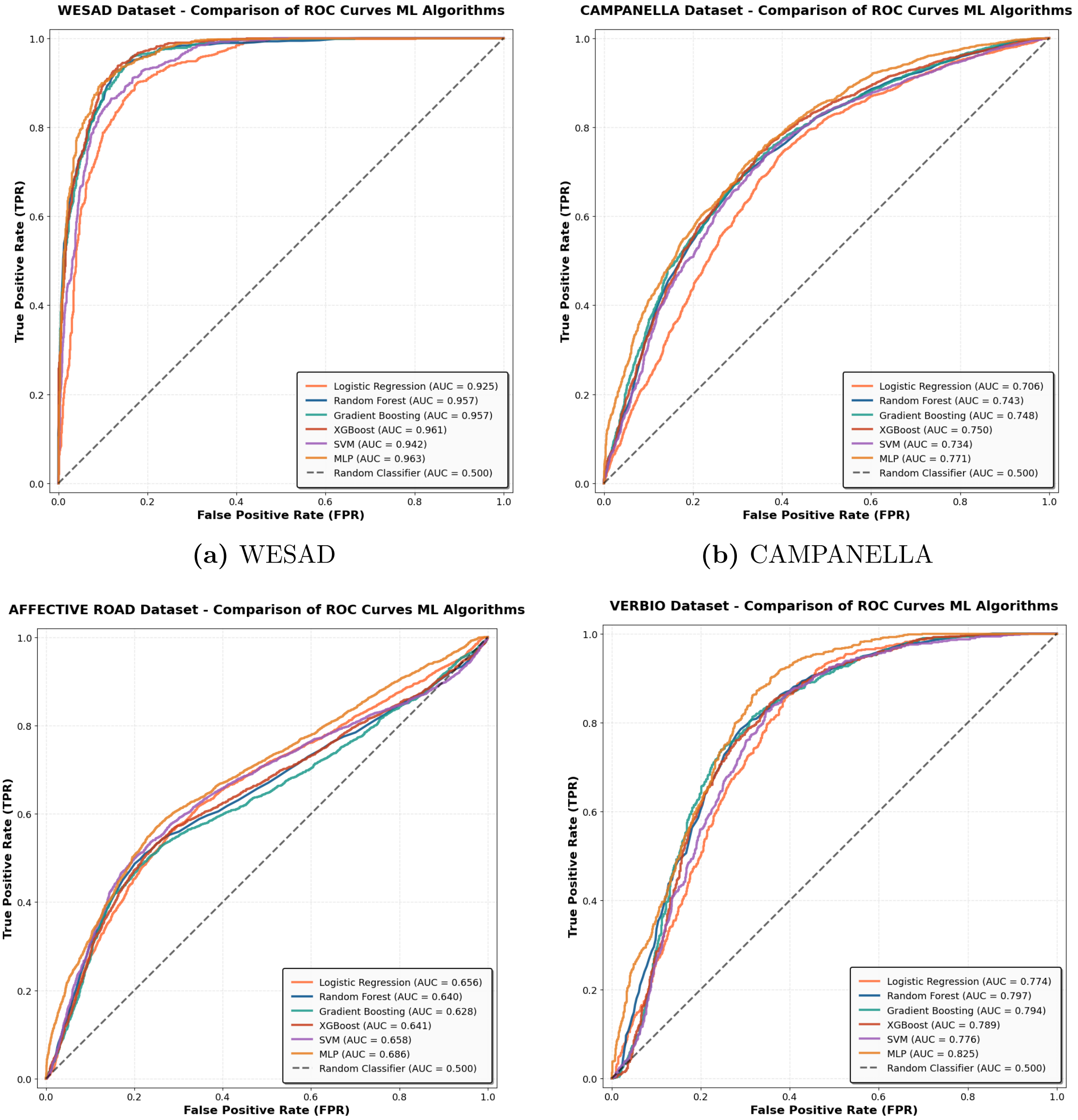
recall (for example, 0.685 vs. 0.602 for RBF-SVM), indicating conservative positive decisions. Movement-related descriptors (accelerometry magnitude statistics) and peripheral proxies (PPG amplitude) contribute prominently, consistent with the ecological, motion-rich setting. The CNN improves the operating point (macro-F1 about 0.69) but retains a precision–recall asymmetry.

CAMPANELLA exhibits comparatively balanced class proportions (stress typically above sixty percent) and stable responses. The Multilayer Perceptron (MLP) achieves the highest macro-F1 among classical learners (0.819 ± 0.039) with high recall (0.900 ± 0.079), suggesting that stress-inducing tasks elicit consistent physiological signatures across participants. Gradient-boosted ensembles are close behind (XGB/GB around 0.806–0.812 macro-F1), and the smaller standard deviations reflect inter-subject stability under a controlled protocol. The CNN operates at a conservative threshold (macro-F1 about 0.74 with high precision and lower recall), reflecting class skew and participant-specific dynamics.

VerBIO retains residual imbalance and rater-induced variability despite the five percent rule. The highest macro-F1 among classical models is obtained by the RBF-SVM (0.627 ± 0.248), with large dispersion attributable to residual subject heterogeneity and the semi-subjective rating process. Margin-based separation with regularisation appears more robust than tree ensembles, whose per-fold performance oscillates widely. The CNN improves the aggregate operating point (macro-F1 about 0.74) yet displays broader spread, consistent with heterogeneous sessions and thresholded continuous annotations.

WESAD includes only subjects satisfying the threshold (stress prevalence roughly thirty-five percent) and benefits from clean, phase-aligned labels in a laboratory setting. XGBoost attains the best macro-F1 among classical models (0.866 ± 0.114), followed by GB (0.847 ± 0.118), MLP (0.846 ± 0.136), and RBF-SVM (0.824 ± 0.115). The CNN surpasses all competitors (macro-F1 close to 0.96), with near-ceiling precision and high recall, indicating that temporal filters capture discriminative waveform structure beyond handcrafted summaries.

At the algorithm level, consistent patterns emerge across corpora and align with the operating conditions reflected in Table 4.1. Linear LR provides a stable lower bound with modest dispersion but rarely attains the top macro-F1, indicating limited capacity to capture nonlinear interactions among modalities. Tree-based ensembles exploit heterogeneous feature interactions more effectively: GB and XGBoost frequently rank among the best classical learners, with XGBoost leading in WESAD and remaining consistently strong in CAMPANELLA; these margins are accompanied by comparatively small standard deviations under controlled protocols, suggesting reliable decision boundaries when labels are phase-aligned and noise is limited. RF is competitive but somewhat more variable across folds, consistent with sensitivity to class prevalence and correlated descriptors. Margin-based SVM with an RBF kernel is particularly robust in noisier or heterogeneous domains such as AffectiveROAD and VerBIO, where maximising the margin mitigates label noise and residual imbalance; the precision–recall profile typically skews toward higher precision, reflecting conservative positive assignments under uncertainty. The shallow MLP is competitive in controlled settings, notably in CAMPANELLA where recall is high and macro-F1 reaches the classical-model maximum, but it degrades as imbalance and label ambiguity grow unless class weighting and calibration are carefully tuned. The CNN extends these trends by learning temporal representations directly from the raw streams: it dominates in WESAD and improves the operating point in VerBIO and AffectiveROAD, yet tends to settle on conservative thresholds that favour precision over recall when motion artefacts or diffuse supervision are present, as corroborated by the triptych figures. Taken together, ensembles and SVM provide strong and complementary baselines for feature-driven inputs—ensembles excelling with structured, clean labels and SVMs with heterogeneous, noisy regimes—while the CNN offers the largest gains where temporal regularities are reliable; when recall is critical under skewed distributions, classical models with explicit class handling can still be preferable or serve as calibration anchors for the end-to-end approach.



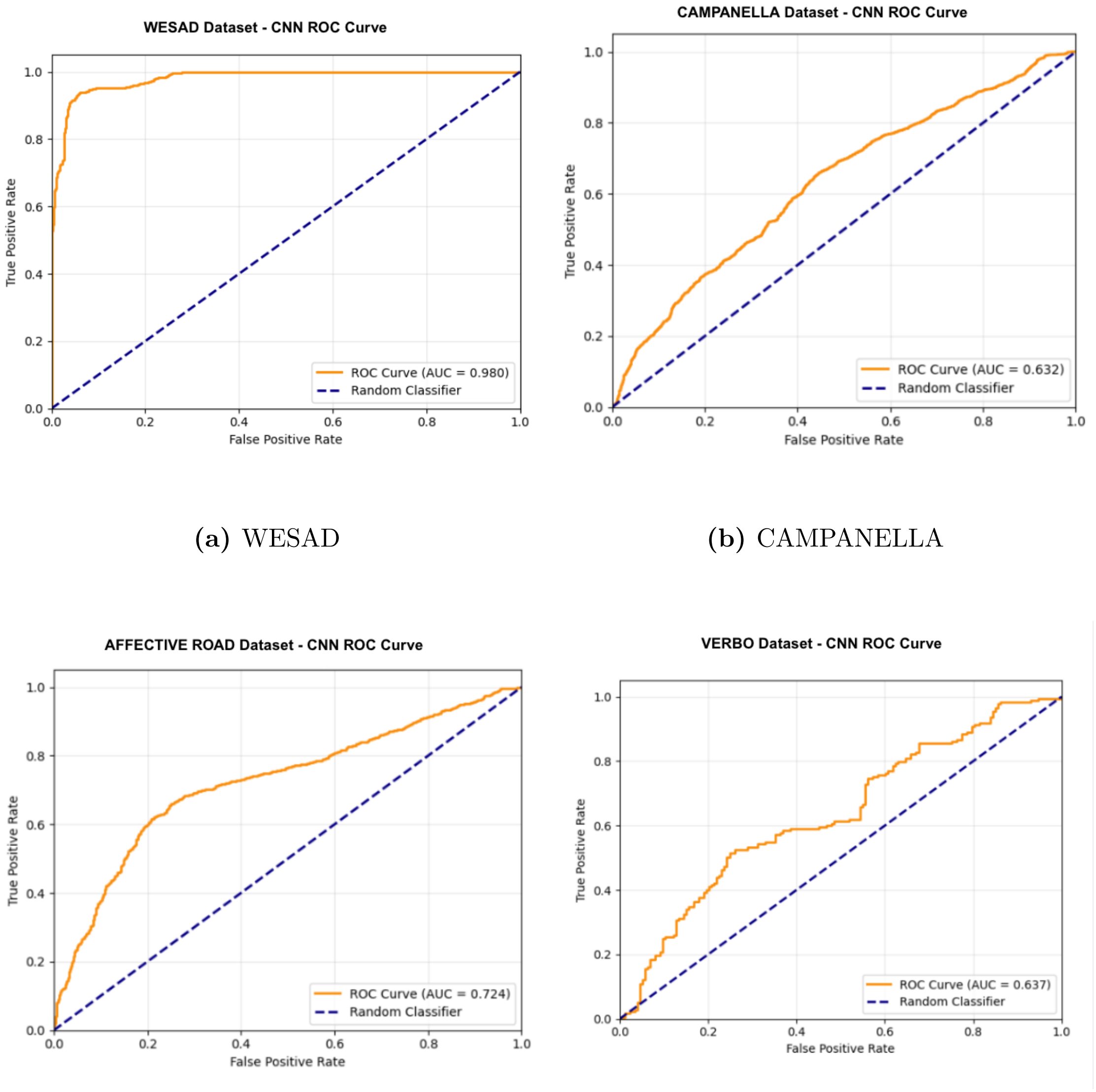
(a) WESAD

(b) CAMPANELLA

(c) AffectiveROAD

(d) VerBIO

Figure 4.1: ROC curves for classical feature-driven models across datasets (LOSO).



(a) WESAD

(b) CAMPANELLA

(c) AffectiveROAD

(d) VerBIO

Figure 4.2: ROC curves for the data-driven CNN across datasets (LOSO).

4.1.1 Feature importance: common markers and dataset-specific signals

Across datasets and models, a coherent set of physiological descriptors consistently emerges at the top of the importance rankings (see Tables 4.3–4.8). HRV measures dominate this core: classical dispersion indices (SDNN, pNN50), robust distributional summaries (MedianNN, Median Absolute Deviation of NN Intervals (MadNN), Interquartile Range of NN Intervals (IQRNN)), and non-linear complexity markers (Fuzzy Entropy (FuzzyEn), Sample Entropy (SampEn)) repeatedly surface across linear and non-linear learners, reflecting vagal withdrawal and sympathetic activation during stress. Electrodermal reactivity emerges through phasic indices such as the number and mean amplitude of skin conductance responses (SCR_Peaks_N, SCR_Peaks_Amplitude_Mean), corroborating arousal sensitivity beyond tonic level shifts. Photoplethysmography (PPG) morphology contributes with amplitude- and duration-related descriptors (PPG_Amplitude, PPG_Duration), consistent with stress-driven vasoconstriction and pulse contour changes. Taken together, these findings support a compact cross-dataset backbone composed of HRV, EDA, and PPG as the most informative triad for feature-driven stress recognition.

Context modulates which features lead within that backbone. In WESAD (controlled, task-driven labels), skin temperature summaries (Temp_Mean, Temp_Slope) enter the top ranks alongside HRV and SCR, benefiting from a stable thermal environment. In Campanella (laboratory protocol), HRV robustness measures and PPG_Amplitude rise consistently, aligning with clear cardiovascular responses. In VerBIO (semi-controlled, rater-assisted labels), SCR counts/amplitudes and robust HRV statistics rank highly, suggesting resilience to intra-subject variability. In the ecological AffectiveROAD, accelerometry magnitude statistics (AccMag_Std, AccMag_Kurtosis, AccMag_Max) gain salience, capturing movement- and context-related variance that co-occurs with operational stress. The same families of features recur across algorithms (LR, RF, GB, XGB, SVM, MLP), indicating physiological consistency rather than model-specific artefacts (cf. Tables 4.3–4.8). This convergence suggests practical guidelines for feature design: start from the HRV–EDA–PPG core, then specialise with temperature (controlled settings) or accelerometry (in-the-wild) to address protocol- and context-dependent variability.

Table 4.2 provides the acronyms referenced in the discussion, with concise definitions to improve readability of the subsequent per-algorithm importance tables.

Table 4.2: Acronyms and concise definitions of the features discussed in the importance analysis. NN intervals denote beat-to-beat intervals derived from BVP/PPG (IBI).

Acronym	**Signal**	**Definition (summary)**
SDNN	HRV (PPG)	Standard deviation of NN intervals within the analysis window.
RMSSD	HRV (PPG)	Root mean square of successive differences between adjacent NN intervals.
pNN50	HRV (PPG)	Percentage of successive NN pairs differing by $> 50\,\mathrm{ms}$.
MedianNN	HRV (PPG)	Median of NN intervals.
MadNN	HRV (PPG)	Median absolute deviation of NN intervals (robust dispersion).
IQRNN	HRV (PPG)	Interquartile range of NN intervals (robust spread).
LF	HRV (PPG)	Spectral power in the 0.04–0.15 Hz band (Welch/AR).
HF	HRV (PPG)	Spectral power in the 0.15–0.40 Hz band.
LF/HF	HRV (PPG)	Ratio between LF and HF power.
SampEn	HRV (PPG)	Sample entropy of the NN-interval series (pattern irregularity).
FuzzyEn	HRV (PPG)	Fuzzy entropy of the NN-interval series (noise-robust complexity).
PPG Amp	PPG	Mean systolic peak amplitude across beats in the window.
PPG Dur	PPG	Mean beat duration (onset-to-onset or foot-to-foot) across beats.
SCR Peaks N	EDA	Count of skin conductance responses detected in the window.
SCR Peaks Amp Mean	EDA	Mean amplitude of detected SCR peaks.
SCL	EDA	Mean tonic skin conductance level.
EDA AUC Phasic	EDA	Area under the curve of the phasic EDA component over the window.
SCR Latency Mean	EDA	Mean latency from onset to SCR peak (when estimable).
Temp Mean	TEMP	Mean skin temperature in the window.
Temp Slope	TEMP	Linear trend (slope) of skin temperature within the window.
AccMag Mean	ACC	Mean of accelerometer magnitude $\sqrt{x^2 + y^2 + z^2}$.
AccMag Std	ACC	Standard deviation of accelerometer magnitude.
AccMag Max	ACC	Maximum accelerometer magnitude observed in the window.
AccMag Kurtosis	ACC	Fourth standardized moment of accelerometer magnitude (tailedness).

Table 4.3: Top-5 features by importance per dataset (Logistic Regression).

Dataset	Top 5 features (highest → lowest)
WESAD	*HRV_MedianNN, HRV_Prc20NN, SCR_Peaks_N, HRV_MadNN, Temp_Mean*
Campanella	*HRV_FuzzyEn, PPG_Amplitude, HRV_MedianNN, HRV_MadNN, HRV_pNN50*
VerBIO	*HRV_MadNN, HRV_pNN50, HRV_MedianNN, HRV_FuzzyEn, SCR_Peaks_N*
AffectiveROAD	*HRV_SDNN, AccMag_Std, Temp_Std, AccMag_Kurtosis, AccMag_Max*

Table 4.4: Top-5 features by importance per dataset (Random Forest).

Dataset	Top 5 features (highest → lowest)
WESAD	*HRV_Prc20NN, SCR_Peaks_Amplitude_Mean, Temp_Mean, Temp_Slope, SCR_Peaks_N*
Campanella	*PPG_Amplitude, HRV_FuzzyEn, HRV_IQRNN, HRV_pNN50, SCR_Peaks_Amplitude_Mean*
VerBIO	*HRV_SDNN, HRV_MadNN, SCR_Peaks_N, Temp_Slope, SCR_Peaks_Amplitude_Mean*
AffectiveROAD	*AccMag_Kurtosis, AccMag_Std, PPG_Amplitude, HRV_MeanNN, SCR_Peaks_Amplitude_Mean*

Table 4.5: Top-5 features by importance per dataset (Gradient Boosting).

Dataset	Top 5 features (highest → lowest)
WESAD	*HRV_Prc20NN, Temp_Slope, SCR_Peaks_N, Temp_Mean, SCR_Peaks_Amplitude_Mean*
Campanella	*PPG_Amplitude, HRV_IQRNN, HRV_FuzzyEn, SCR_Peaks_N, HRV_pNN50*
VerBIO	*SCR_Peaks_N, HRV_SDNN, HRV_pNN50, Temp_Slope, HRV_MadNN*
AffectiveROAD	*AccMag_Kurtosis, AccMag_Std, PPG_Amplitude, HRV_MeanNN, AccMag_Max*

Table 4.6: Top-5 features by importance per dataset (XGBoost).

Dataset	**Top 5 features (highest → lowest)**
WESAD	*HRV_Prc20NN, SCR_Peaks_N, HRV_pNN50, SCR_Peaks_Amplitude_Mean, Temp_Mean*
Campanella	*HRV_IQRNN, HRV_FuzzyEn, PPG_Amplitude, HRV_MeanNN, SCR_Peaks_N*
VerBIO	*HRV_SDNN, SCR_Peaks_N, HRV_FuzzyEn, PPG_Duration, SCR_Peaks_Amplitude_Mean*
AffectiveROAD	*AccMag_Kurtosis, AccMag_Std, HRV_MinNN, AccMag_Max, PPG_Amplitude*

Table 4.7: Top-5 features by importance per dataset (SVM, RBF kernel).

Dataset	**Top 5 features (highest → lowest)**
WESAD	*SCR_Peaks_N, Temp_Mean, HRV_pNN50, HRV_CD, HRV_SampEn*
Campanella	*HRV_pNN50, PPG_Amplitude, HRV_MadNN, HRV_FuzzyEn, SCR_Peaks_N*
VerBIO	*HRV_pNN50, HRV_MadNN, HRV_FuzzyEn, HRV_MedianNN, SCR_Peaks_N*
AffectiveROAD	*AccMag_Max, HRV_PI, SCR_Peaks_N, PPG_Amplitude, HRV_MinNN*

Table 4.8: Top-5 features by importance per dataset (Multi-Layer Perceptron).

Dataset	**Top 5 features (highest → lowest)**
WESAD	*SCR_Peaks_N, HRV_MadNN, Temp_Mean, HRV_Prc20NN, HRV_pNN50*
Campanella	*HRV_pNN50, PPG_Amplitude, HRV_MedianNN, AccMag_IQR, PPG_Duration*
VerBIO	*HRV_MadNN, SCR_Peaks_N, HRV_pNN50, HRV_Prc20NN, HRV_MedianNN*
AffectiveROAD	*AccMag_Std, PPG_Amplitude, HRV_MinNN, SCR_Peaks_Amplitude_Mean, Temp_Std*

4.2 Cross-dataset and Transfer-learning Evaluation

To enable a direct, like-for-like comparison between approaches, cross-dataset (*zero-shot*) and transfer-learning results are consolidated below into two unified tables. Table 4.9 extends the original feature-driven cross-test matrix by adding a *CNN* row within each target block (thus replacing the separate CNN cross-test table), while Table 4.10 augments the transfer-learning matrix with an *Algorithm* column so that both *MLP* (feature-driven) and *CNN* (data-driven) transfer results are presented side-by-side. The light-green diagonal reports in-domain LOSO performance; within each target, the best off-diagonal score is highlighted in bold. All values are macro-F1 ± SD at the subject level where available. It is important to note that the in-domain results reported on the light-green diagonal of Tables 4.9 and 4.10 are not numerically identical to the LOSO baselines in Table 4.1. For the unified cross-dataset and transfer-learning analysis, all models were retrained on the common feature subset derived from the WESAD pipeline, as detailed in Section 3.4. Consequently, the diagonal cells in Tables 4.9–4.10 quantify in-domain performance under this harmonised feature space, whereas Table 4.1 reflects the best-performing, dataset-specific feature-selection pipelines. Differences between the two sets of scores therefore stem from the change in input representation rather than from discrepancies in evaluation protocol.

A coherent picture emerges when reading Tables 4.9–4.10 holistically. In-domain performance mirrors protocol control and label discreteness: WESAD stands out at the top for both approaches, followed by CAMPANELLA and VerBIO, with AffectiveROAD lower on average due to ecological variability. Zero-shot cross-test exposes sizeable domain gaps across the board; nevertheless, certain laboratory-to-laboratory directions are relatively resilient (e.g., WESAD↔CAMPANELLA with LR in the feature-driven branch), while CAMPANELLA appears the most transferable *source* towards both VerBIO (XGB) and AffectiveROAD (RF) within the feature-driven family. The data-driven CNN, now embedded in the same matrix, is competitive as a cross-source in some directions but shows marked off-diagonal drops whenever stress semantics and motion context differ, especially towards AffectiveROAD and VerBIO, consistently with label diffuseness and artefacts.

Transfer learning reduces these gaps substantially, particularly for the CNN where fine-tuning yields large absolute gains on AffectiveROAD and WESAD and consistently improves CAMPANELLA and VerBIO as targets. Under the shared handcrafted feature space, MLP fine-tuning provides smaller and less systematic improvements, with isolated benefits (e.g., CAMPANELLA→WESAD) but no universal lift; by contrast, CNN adaptation tends to reorder the best cross-sources

Table 4.9: Unified *cross-test* results (macro-F1 ± SD). Rows are *targets*; columns are *sources*. The light-green diagonal reports in-domain performance (train & test on the same dataset). Within each target, the best *off-diagonal* cell is in **bold**.

Target	Algorithm	WESAD	CAMPANELLA	VERBIO	AFFECTIVE ROAD
WESAD	**Logistic Regression**	0,8063 ± 0,1128	**0,7192 ± 0,1555**	0,0000 ± 0,0000	0,2467 ± 0,2669
	Random Forest	0,8204 ± 0,1816	0,5462 ± 0,0920	0,2717 ± 0,1957	0,1640 ± 0,1295
	Gradient Boosting	0,8474 ± 0,1182	0,5731 ± 0,2093	0,5583 ± 0,2576	0,2475 ± 0,2099
	XGBoost	0,8655 ± 0,1136	0,6303 ± 0,1418	0,5525 ± 0,2165	0,1968 ± 0,1941
	MLP	0,8616 ± 0,1296	0,5733 ± 0,0959	0,0000 ± 0,0000	0,3029 ± 0,2051
	CNN	0,9575 ± 0,0841	0,5370 ± 0,1893	0,5005 ± 0,3210	0,4589 ± 0,2369
CAMPANELLA	**Logistic Regression**	**0,7675 ± 0,0775**	0,7591 ± 0,1270	0,0000 ± 0,0000	0,5371 ± 0,2142
	Random Forest	0,5202 ± 0,1606	0,8127 ± 0,0460	0,2806 ± 0,1670	0,3689 ± 0,2019
	Gradient Boosting	0,5912 ± 0,1511	0,8118 ± 0,0547	0,5629 ± 0,1298	0,3657 ± 0,1337
	XGBoost	0,5201 ± 0,1669	0,8098 ± 0,0570	0,5896 ± 0,1197	0,2835 ± 0,1792
	MLP	0,6941 ± 0,1323	0,8276 ± 0,0477	0,0000 ± 0,0000	0,6043 ± 0,1632
	CNN	0,5763 ± 0,2552	0,7434 ± 0,1910	0,5686 ± 0,2694	0,5875 ± 0,2458
VERBIO	Logistic Regression	0,4766 ± 0,1534	0,4766 ± 0,1534	0,5895 ± 0,2605	0,4766 ± 0,1534
	Random Forest	0,4387 ± 0,2422	0,4854 ± 0,1606	0,5402 ± 0,2450	0,2799 ± 0,1659
	Gradient Boosting	0,4685 ± 0,2277	0,3256 ± 0,2465	0,5549 ± 0,2531	0,2405 ± 0,1621
	XGBoost	0,4588 ± 0,2036	**0,5852 ± 0,1843**	0,5370 ± 0,2614	0,3093 ± 0,2166
	MLP	0,4784 ± 0,1524	0,4766 ± 0,1534	0,5385 ± 0,2696	0,4780 ± 0,1529
	CNN	0,3610 ± 0,3238	0,4941 ± 0,2785	0,7418 ± 0,1791	0,4585 ± 0,3163
AFFECTIVE ROAD	Logistic Regression	0,5910 ± 0,1837	0,5477 ± 0,1758	0,0000 ± 0,0000	0,6201 ± 0,1296
	Random Forest	0,3270 ± 0,2906	**0,6678 ± 0,1208**	0,1540 ± 0,1275	0,5794 ± 0,0997
	Gradient Boosting	0,4129 ± 0,3108	0,5662 ± 0,1525	0,2149 ± 0,1784	0,5878 ± 0,1307
	XGBoost	0,3829 ± 0,3230	0,5841 ± 0,1517	0,2909 ± 0,1913	0,6033 ± 0,1113
	MLP	0,4760 ± 0,2699	0,6602 ± 0,1141	0,0000 ± 0,0000	0,6110 ± 0,1251
	CNN	0,2675 ± 0,3168	0,2787 ± 0,2237	0,0930 ± 0,1367	0,6934 ± 0,1089

Table 4.10: Unified *transfer learning* results (macro-F1 ± SD). Rows are *targets* with the *Algorithm* column distinguishing feature-driven MLP vs. data-driven CNN. The light-green diagonal shows in-domain performance; within each target, the best *cross-dataset* score is in **bold**.

Target	Algorithm	WESAD	CAMPANELLA	VERBIO	AFFECTIVE ROAD
WESAD	MLP	0,8616 ± 0,1296	0,5934 ± 0,0934	0,0000 ± 0,0000	0,3022 ± 0,2044
	CNN	0,9575 ± 0,0841	0,7854 ± 0,1357	**0,8030 ± 0,2035**	0,6780 ± 0,1558
CAMPANELLA	MLP	0,6742 ± 0,1425	0,8276 ± 0,0477	0,0000 ± 0,0000	0,5764 ± 0,1731
	CNN	0,7681 ± 0,1330	0,7434 ± 0,1910	0,7664 ± 0,1907	**0,7713 ± 0,1516**
VERBIO	MLP	0,4766 ± 0,1505	0,4766 ± 0,1534	0,5385 ± 0,2696	0,4785 ± 0,1530
	CNN	0,5942 ± 0,2517	0,6368 ± 0,2533	0,7418 ± 0,1791	**0,6458 ± 0,2321**
AFFECTIVE ROAD	MLP	0,5060 ± 0,2406	0,6295 ± 0,1351	0,0000 ± 0,0000	0,6110 ± 0,1251
	CNN	0,4970 ± 0,2754	**0,6550 ± 0,1373**	0,5088 ± 0,2992	0,6934 ± 0,1089

per target (e.g., VerBIO→WESAD and CAMPANELLA→AffectiveROAD after fine-tuning), indicating that representation learning profits more from modest target supervision than do shallow classifiers on fixed descriptors. From an algorithmic standpoint, tree ensembles (GB/XGB/RF) remain dependable baselines for feature-driven transfer—often yielding the best zero-shot cells per target—while the CNN attains the highest in-domain ceilings and the largest transfer-induced recoveries when distributions are far apart. Overall, the unified view confirms three practical implications: cross-dataset stress recognition suffers from pronounced domain shift; source–target pairing and model family need to be chosen with the target distribution in mind; and adaptation is especially beneficial for data-driven models, whereas feature-driven pipelines retain advantages in stability and interpretability under controlled protocols.

A deeper comparative reading of Tables 4.9 and 4.10 clarifies how representation choice and adaptation shape portability across corpora. In the zero-shot regime (Table 4.9) the feature-driven pipeline provides more reliable out-of-distribution behaviour, particularly along laboratory-to-laboratory directions and whenever class semantics and sensor context are closely aligned. CAMPANELLA emerges as the most transferable source within the feature space, repeatedly yielding the highest off-diagonal scores towards WESAD, VerBIO, and AffectiveROAD, while also benefiting from WESAD-trained linear models in the reverse direction. The CNN, when evaluated strictly out-of-domain without fine-tuning, shows larger drops on pairs that combine social-evaluative or rating-derived labels with motion-rich acquisition (for example towards VerBIO and AffectiveROAD), consistent with its greater sensitivity to distributional shift in temporal patterns and noise characteristics. Where the CNN remains competitive in zero-shot settings, the gap to the best feature-driven model narrows primarily when source and target share stressor style and operating conditions.

Fine-tuning decisively alters the picture (Table 4.10). The CNN leverages even modest target supervision to close and often invert the zero-shot gap, with marked jumps on distant domains such as CAMPANELLA → AffectiveROAD and VerBIO → WESAD, and consistent gains on CAMPANELLA and VerBIO as targets. The best cross-source per target frequently changes once adaptation is allowed, indicating that representation learning can reshape the alignment between domains more effectively than adjusting a shallow classifier on fixed descriptors. By contrast, MLP-based transfer on the shared handcrafted features produces smaller and less systematic improvements: when the feature subspace already encodes robust physiology, re-optimising the head has limited headroom, whereas the CNN can re-tune early temporal filters and fusion to the target's label formation and artefact profile.

Stability patterns mirror these dynamics. In zero-shot cross-test, feature-driven models generally exhibit lower variance across subjects on controlled targets and retain moderate dispersion on in-the-wild targets; the CNN displays higher spread off-diagonal, especially where label noise or context heterogeneity is prominent. With transfer learning, CNN variability contracts substantially on all targets, most visibly where zero-shot dispersion was largest, while feature-driven transfer shows modest variance reductions that track its smaller mean gains. These observations suggest a pragmatic division of labour: when no target labels are available, starting from the feature-driven space is safer for cross-dataset deployment; when a limited amount of target supervision can be afforded, adapting a CNN offers the largest return on investment and often reshapes the optimal source–target pairing. In practical terms, the choice between the two branches should be informed by the anticipated access to target labels and by the severity of the domain shift observed in Table 4.9; where shift is substantial and some adaptation is possible, Table 4.10 indicates that the data-driven route is preferable, whereas under strict zero-shot constraints the feature-driven route remains the more dependable option.

4.3 Comparison with Existing Studies

Several recent works have explored stress detection using wearable physiological sensors, often focusing on multimodal signal integration and machine learning approaches. Most of these studies rely on intra-dataset validation protocols, such as k-fold or leave-one-subject-out cross-validation, and only a limited number have addressed cross-dataset generalization or transfer learning.

The study by Benchekroun et al. [12] evaluated several classical machine learning models on the WESAD and AffectiveROAD datasets. Their approach involved the extraction of hand-crafted features from ECG, EDA, and respiration signals, followed by classification using SVM, RF, and LR. Although the intra-dataset performance on WESAD was high (F1-scores above 0.8), cross-dataset results were limited, especially when training on WESAD and testing on AffectiveROAD, highlighting poor generalization. Unlike that work, the present study adopts a unified experimental protocol across datasets and evaluates both cross-test and transfer learning performance.

Prajod et al. [80] proposed a ResNet-based CNN architecture trained on the WESAD dataset and tested on the AffectiveROAD dataset using spectrogram representations of the biosignals. The best cross-dataset result reported was an F1-score of 0.41, significantly lower than the 0.65 obtained in the present study using CNN with transfer learning. This highlights the advantage of domain adaptation and data-driven learning strategies adopted in this thesis.

Ladakis et al. [60] explored stress recognition using spectro-temporal representations of biosignals and a lightweight CNN. Their results showed good intra-dataset performance, but they did not perform cross-dataset testing or transfer learning evaluation, limiting their study's generalizability compared to the approach used in the present work.

Albaladejo-González et al. [3] introduced a large multimodal dataset (UBFC-Phys) and evaluated several neural models. While this work shares similarities in using deep learning, it focuses mostly on in-lab validation, with limited exploration of cross-domain generalization or transfer strategies, which are central in the present study.

Can et al. [20] conducted a review of stress detection using wearable sensors, emphasizing generalizability across datasets. However, their reported results were constrained to binary classification and generally lower than those achieved in the present thesis. For example, a cross-dataset F1-score of 0.7675 was achieved in this study using LR when training on WESAD and testing on CAMPANELLA, outperforming all values reported in that survey.

Wu et al. (2021) [111] explored deep learning models for stress assessment using photoplethysmogram (PPG) signals. They proposed a CNN-based method and reported F1-scores around 0.75 in intra-dataset evaluation but did not perform any cross-dataset validation or transfer learning. In contrast, the present study emphasizes model generalization across heterogeneous data sources, which is lacking in Wu et al.'s evaluation pipeline.

Attallah et al. (2025) [2] presented a comprehensive analysis of stress detection using HRV features and ensemble learning techniques. Although their work reported solid intra-dataset results, cross-dataset generalization was not thoroughly addressed. Furthermore, their approach was limited to a binary classification scheme, whereas the present work evaluates performance in both binary and multiclass setups using a unified cross-domain framework.

Unlike most of the above works, the current study includes a comprehensive transfer learning evaluation, adopting both feature-based (MLP) and data-driven (CNN) approaches. Results confirm the efficacy of CNN-based models, which achieved top in-domain performance (e.g., F1 = 0.9575 on WESAD) and strong cross-dataset generalization (e.g., F1 = 0.8030 from CAMPANELLA to VERBIO), even in the absence of target domain retraining.

Furthermore, the unified experimental protocol across four datasets (WESAD, AffectiveROAD, CAMPANELLA, and VERBIO) highlights both the challenges and the potential of stress detection models under real-world data distribution shifts. The inclusion of both cross-test and transfer learning evaluations, rarely considered together in prior studies, enables a robust assessment of model generalizability. The results underline that data-driven CNN approaches are particularly effective under such conditions.

Table 4.11: Overview of selected studies on wearable stress detection. Reported F1-scores are for intra-dataset evaluation, cross-dataset testing, and transfer learning where applicable.

Article	Stress Signal(s)	Stress Test	Datasets	Method	F1 Intra	F1 Cross
[12] (2023)	ECG, EDA, RESP	Task-based	WESAD, AffectiveROAD	Classical ML	>0.80	0.57
[80] (2024)	ECG, EDA, TEMP	Task-based	WESAD, AffectiveROAD	DL (CNN)	0.95	0.41
[60] (2025)	BVP, EDA, TEMP, ACC	Task-based	WESAD, AffectiveROAD, SWELL-KW	DL (CNN)	0.90	0.70
[3] (2023)	ECG, TEMP, EDA, PPG	Task-based	UBFC-Phys, SWELL-KW	Hybrid (ML + DL)	0.90	<0.60
[20] (2019)	ECG, EDA, RESP	Task-based	WESAD, SWELL-KW, AffectiveROAD	Classical ML	0.92	0.60
[62] (2024)	ECG, EDA, TEMP	Task-based	WESAD, AffectiveROAD	DL (RNN/LSTM)	0.91	0.55
[2] (2025)	EEG	VR stress protocol	Custom EEG dataset	EEG-based DL	0.93	–
This Study	BVP, EDA, TEMP, ACC	Mixed/ real-world	WESAD, AffectiveROAD, CAMPANELLA, VERBIO	Feature-based ML, CNN	0.96	0.77

4.4 Conclusion

This work compared two modelling paradigms for wearable stress recognition—one feature-driven, based on handcrafted descriptors and classical ML, and one data-driven, based on an end-to-end 1D CNN—across four heterogeneous datasets under three complementary protocols: within-dataset LOSO, zero-shot cross-dataset testing, and transfer learning with limited target supervision. The empirical picture that emerges is consistent and actionable. In-domain performance is high for both families but depends on corpus characteristics rather than model class alone: controlled laboratory signals with discrete labels favour very high ceilings for the CNN, while structured yet class-skewed laboratory protocols can still reward feature ensembles. When moving to zero-shot cross-dataset evaluation, the feature-driven pipeline provides the most reliable portability across targets; its handcrafted representations mitigate distribution shift when no target labels are available. Allowing modest adaptation reverses the balance: fine-tuning enables the CNN to close and often surpass the zero-shot advantage of feature-based models, with the largest gains observed when source and target differ substantially in protocol and context. These trends, consolidated in the unified cross-test and transfer-learning tables of in section 4.2, reflect the interplay between representation, supervision, and domain shift.

Several limitations qualify these findings. The corpus set covers four datasets and, although heterogeneous, does not exhaust the range of stressors, recording contexts and sensor stacks encountered in real-world deployments; generalization beyond these domains remains to be demonstrated. Label definitions differ across corpora, with variations between binary stress versus baseline, multi-class affective states, and continuous annotations; despite harmonization, residual label mismatch can confound cross-dataset outcomes. Sample sizes are modest and subject distributions unbalanced, which may inflate variance of LOSO and cross-test estimates; confidence intervals and formal significance testing across off-diagonal comparisons were not exhaustively reported. The evaluation emphasised macro-F1; calibration quality, decision-cost curves and temporal detection latency were not assessed and could alter conclusions for thresholded operation. Pre-processing and windowing choices were standardised for comparability rather than individually optimised per dataset or modality; suboptimal settings may disadvantage specific signals. The data-driven branch relied on a single 1D CNN family without self-supervised pretraining, foundation backbones or domain generalization objectives; alternative architectures and pretraining regimes might yield different trade-offs. Transfer learning involved limited target supervision and a fixed fine-tuning budget; sensitivity to adaptation schedule, layer freezing strategies and source selection was not fully explored. Motion context was only implicitly modelled through ACC channels; explicit activity recognition or context-aware conditioning might mitigate confounds between stress and movement. Interpretability was addressed indirectly through feature families and coefficient inspection for classical models; a comprehensive physiological attribution analysis for the CNN is absent. Finally, hardware variability, sensor placement differences and potential device-specific artefacts were not isolated through device-controlled ablations, leaving open questions about robustness to sensor drift and replacement.

The research questions can therefore be answered succinctly. Regarding in-domain LOSO performance, there is no universal winner and the top approach is dataset-specific; temporal filters in the CNN can exploit controlled stimuli and clean labels, whereas curated feature spaces remain competitive when signal dynamics are stable but class imbalance is pronounced. On zero-shot cross-dataset generalisation, the feature-driven approach is preferable, consistently achieving higher off-diagonal macro-F1 without any target supervision. Concerning transfer learning, fine-tuning brings substantial improvements for the CNN and typically elevates it above the best feature-driven zero-shot scores on most targets, while adaptation of shallow models on fixed features yields smaller, less systematic gains. With respect to failure modes, generalisation varies with stressor type, annotation strategy, class prevalence, and movement artefacts; wrist signals recorded in the wild exacerbate noise and label diffuseness, and source–target asymmetries are common, making the choice of source a material design decision.

The hypotheses align with this evidence. The expectation that feature-driven methods would dominate in-domain is only partially supported, as the CNN attains the best LOSO results in several settings. The conjecture that the CNN would degrade less under zero-shot transfer is not borne out; handcrafted features suffer less degradation when no adaptation is possible. The anticipated benefit of transfer learning is confirmed with nuance: it is large and systematic for the CNN, modest for feature-driven models. Finally, the idea that the effect of adaptation scales with the magnitude of the domain shift is supported by the largest fine-tuning gains on the most distant source–target pairs.

From a deployment perspective, the choice of pipeline should reflect the expected availability of target labels and the severity of shift. When no supervision can be collected on the target domain, the feature-driven route is the safer default for immediate portability. When even a small amount of target supervision is feasible, adapting a CNN becomes the preferred option, typically yielding the highest generalisation. In practical terms, systems should screen sources for compatibility with the intended target, retain a robust feature-driven baseline for zero-shot operation, and reserve a light fine-tuning budget for the CNN to consolidate performance once target data are acquired.

Looking ahead, three directions appear most promising. First, pretraining at scale with self-supervised objectives on large pools of unlabeled wearable signals can supply a general backbone whose representations transfer with minimal labels, narrowing the zero-shot gap while preserving the strong fine-tuning gains observed here. Second, domain generalisation and alignment techniques that enforce invariances across datasets and tolerate missing or swapped modalities can strengthen zero-shot robustness without requiring target labels. Third, test-time and source-free adaptation, together with lightweight personalisation, can track sensor drift and subject idiosyncrasies during deployment while respecting privacy constraints. Combining these ingredients with the unified evaluation protocol adopted in this thesis offers a concrete path to shrink the off-diagonal gap highlighted by the results and to deliver reliable, adaptive stress recognition in real-world scenarios.

Bibliography

[1] Mohammad Ahmed, Michael Grillo, Amirtahà Taebi, Mehmet Kaya, and Peshala Gamage. A comprehensive analysis of trapezius muscle emg activity in relation to stress and meditation. *BioMedInformatics*, 4:1047–1058, 04 2024.

[2] Teresa Albaladejo-González, Norberto Malpica, and Alfredo J Pérez. Evaluating machine learning models and transfer learning for stress detection using heart rate variability. *Journal of Ambient Intelligence and Humanized Computing*, 13:11285–11299, 2022.

[3] Mariano Albaladejo González, José A. Ruipérez-Valiente, and Felix Gomez Marmol. Evaluating different configurations of machine learning models and their transfer learning capabilities for stress detection using heart rate. *Journal of Ambient Intelligence and Humanized Computing*, 14, 08 2022.

[4] Andrew P. Allen, Paul J. Kennedy, Samantha Dockray, John F. Cryan, Timothy G. Dinan, and Gerard Clarke. The trier social stress test: Principles and practice. *Neurobiology of Stress*, 6:113–126, 2017.

[5] John Allen. Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, 28:R1–39, 04 2007.

[6] JF Alonso, S Romero, MR Ballester, RM Antonijoan, and MA Mañanas. Stress assessment based on eeg univariate features and functional connectivity measures. *Physiological Measurement*, 36(7):1351–1365, 2015.

[7] Eleni Andreou, Evangelos C. Alexopoulos, Christos Lionis, Liza Varvogli, Charalambos Gnardellis, George P. Chrousos, and Christina Darviri. Perceived stress scale: Reliability and validity study in greece. *International Journal of Environmental Research and Public Health*, 8(8):3287–3298, 2011.

[8] Anjana Bali and Amteshwar Singh Jaggi. Clinical experimental stress studies: methods and assessment. *Reviews in the Neurosciences*, 26(5):555–579, 2015.

[9] Ö. T. Başaran et al. Relieving the burden of intensive labeling for stress: challenges and approaches in ground truth establishment. *Frontiers in Psychology*, 14:1293513, 2023.

[10] Simon Becker, Johanna Schultz, Wiebke Rauch, Viktoria Kegel, Wolff Schlotz, and Clemens Kirschbaum. Activation of the hypothalamic–pituitary adrenal axis in response to a verbal fluency task. *PLOS ONE*, 15(4):e0227721, 2020.

[11] Mouna Benchekroun, Dan Istrate, Vincent Zalc, and Dominique Lenne. Cross dataset analysis for generalizability of hrv-based stress detection. *Sensors*, 2023.

[12] Yassine Benchekroun, Mohamed Abdelaziz, and Samir Messaoud. Stress detection using wearable physiological sensors: A comparative study. In *2021 IEEE International Conference on Smart Healthcare (ICSH)*, pages 1–6. IEEE, 2021.

[13] Alberto Betella and Paul F. M. J. Verschure. The affective slider: A digital self-assessment scale for the measurement of human emotions. *PLOS ONE*, 11(2):e0148037, 2016.

[14] Qifang Bi, Katherine E Goodman, Joshua Kaminsky, and Justin Lessler. What is machine learning? a primer for the epidemiologist. *American Journal of Epidemiology*, 188(12):2222–2239, 2019.

[15] Davide Bolpagni et al. Personalized stress detection using biosignals from wearable devices: A scoping review. *Sensors*, 24(10):3221, 2024.

[16] T.H. Bullock et al. Habituation of the stress response multiplex to repeated cold pressor exposure. *Frontiers in Physiology*, 13:752900, 2023.

[17] Cristina Bustos, Neska El Haouij, Albert Solé-Ribalta, Javier Borge-Holthoefer, Agata Lapedriza, and Rosalind Picard. Predicting driver self-reported stress by analyzing the road scene. *IEEE Transactions on Affective Computing*, 2021.

[18] Fabio Campanella and et al. A multimodal dataset for stress assessment and workload monitoring using wearable devices. *Sensors*, 2022.

[19] Sara Campanella, Ayham Altaleb, Alberto Belli, Paola Pierleoni, and Lorenzo Palma. A method for stress detection using empatica e4 bracelet and machine-learning techniques. *Sensors*, 23(7):3565, 2023.

[20] Yekta Said Can, Nazli Chalabianloo, Deniz Ekiz, and Cem Ersoy. Continuous stress detection using wearable sensors in real life: A review. *Sensors*, 19(19):4415, 2019.

[21] Eldhian Bimantaka Christianto and Beni Rio Hermanto. Analysis of psychological stress and muscle activity using electromyography and stress test. In *2024 International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, pages 760–765, Bandung, Indonesia, March 2024. IEEE.

[22] Bee Seok Chua, H. Shahrul Abd Hamid, Jasmine Adela Mutang, and Rosnah Ismail. Psychometric properties of the state-trait anxiety inventory (form y) among malaysian university students. *Sustainability*, 10(9), 2018.

[23] Olivia Craw, Michael A. Smith, and Mark A. Wetherell. Techniques for inducing stress: Problems and opportunities. In *Handbook of Stress: Stress in the Modern World*, pages 109–123. ABC-CLIO, 2020.

[24] Nadine Darwish, Jessica Yu, Lin Zheng, Rania Nashed, Kevin Yuen, Antonio Toma, Gabrielle Affleck, John Daun, Hymie Anisman, and Shawn Hayley. From lab to real-life: A three-stage validation of wearable technology for stress monitoring. *Frontiers in Behavioral Neuroscience*, 19:1512432, 2025.

[25] M. de Witte et al. Self-report stress measures to assess stress in adults. *Frontiers in Psychology*, 12, 2021.

[26] K. Dedovic, R. Renwick, N. K. Mahani, V. Engert, S. J. Lupien, and J. C. Pruessner. The montreal imaging stress task: using functional imaging to investigate the effects of perceiving and processing psychosocial stress in the human brain. *Journal of Psychiatry & Neuroscience*, 2005.

[27] Neska El Haouij, Jean-Marc Poggi, Samia Sevestre-Ghalila, Rania Ghozi, and Monji Jaïdane. Affectiveroad system and database to assess driver's attention. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, pages 800–803, 2018.

[28] Veronika Engert, Arcangelo Merla, Joshua A. Grant, Daniela Cardone, Anita Tusche, and Tania Singer. Exploring the use of thermal infrared imaging in human stress research. *PLOS ONE*, 9:1–11, 03 2014.

[29] Elissa S. Epel, Alexandra D. Crosswell, Stefanie E. Mayer, Aric A. Prather, George M. Slavich, Eli Puterman, and Wendy Berry Mendes. More than a feeling: A unified view of stress measurement for population science. *Frontiers in Neuroendocrinology*, 2018.

[30] Jonatan Fridolfsson. Effects of frequency filtering on intensity and noise in accelerometer-based physical activity measurements. 2019. discusses choice of filter bands in accelerometry.

[31] Shreyans Gandhi, Maryam Shojaei Baghini, and Soumyo Mukherji. Mental stress assessment - a comparison between hrv based and respiration based techniques. In *Computing in Cardiology Conference (CinC)*, pages 1029–1032. IEEE, 2015.

[32] Farshad Ghasemi, David Q. Beversdorf, and Keith C. Herman. Stress and stress responses: A narrative literature review from physiological mechanisms to intervention approaches. *Journal of Pacific Rim Psychology*, 2024.

[33] Giorgos Giannakakis, Dimitris Grigoriadis, Katerina Giannakaki, Olympia Simantiraki, Alexandros Roniotis, and Manolis Tsiknakis. Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*, 2022.

[34] Martin Gjoreski, Boštjan Čvetković, Hristijan Gjoreski, and Mitja Luštrek. Monitoring stress with a wrist device using context. *Journal of Biomedical Informatics*, 73:159–170, 2017.

[35] Martin Gjoreski, Hristijan Gjoreski, Mitja Luštrek, and Matjaž Gams. Continuous stress detection using a wrist device—in laboratory and real life. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct (UbiComp Adjunct)*, pages 1185–1193, 2016.

[36] Noor Halim, Khairul Sidek, and H. Mansor. Stress recognition using photoplethysmogram signal. *Indonesian Journal of Electrical Engineering and Computer Science*, 8:495–501, 11 2017.

[37] Fatimah Abdul Hamid, M. Naufal M. Saad, and Aamir Saeed Malik. Characterization stress reactions to stroop color-word test using spectral analysis. *Materials Today: Proceedings*, 16:1949–1958, 2019.

[38] K. M. Harris et al. The perceived stress scale as a measure of stress. *Frontiers in Psychology*, 14:10498818, 2023.

[39] Sandra G. Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 50, pages 904–908, 2006.

[40] Jennifer A. Healey and Rosalind W. Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):156–166, 2005.

[41] Jennifer A. Healey and Rosalind W. Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):156–166, 2005.

[42] Seongsil Heo, Sunyoung Kwon, and Jaekoo Lee. Stress detection with single ppg sensor by orchestrating multiple denoising and peak-detecting methods. *IEEE Access*, 9:47777–47788, 2021.

[43] Katherine A. Herborn, Jenny L. Graves, Paul Jerem, Neil P. Evans, Ruedi Nager, Dominic J. McCafferty, and Dorothy E.F. McKeegan. Skin temperature reveals the intensity of acute stress. *Physiology & Behavior*, 152:225–230, December 2015. Epub 2015 Oct 3.

[44] Blake Anthony Hickey, Taryn Chalmers, Phillip Newton, Chin-Teng Lin, David Sibbritt, Craig S. McLachlan, Roderick Clifton-Bligh, John Morley, and Sara Lal. Smart devices and wearable technologies to detect and monitor mental health conditions and stress: A systematic review. *Sensors*, 21(10):3461, 2021.

[45] Nis Hjortskov, Dag Rissén, Anne Katrine Blangsted, Niels Fallentin, Ulf Lundberg, and Karen Søgaard. The effect of mental stress on heart rate variability and blood pressure during computer work. *European Journal of Applied Physiology*, 92(1-2):84–89, June 2004. Epub 2004 Feb 27. PMID: 14991326.

[46] Karen Hovsepian, Mustafa al'Absi, Emre Ertin, Thomas Kamarck, Motohiro Nakajima, and Santosh Kumar. cstress: towards a gold standard for continuous stress assessment in the mobile environment. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, page 493–504, 2015.

[47] S. Immanuel, M. N. Teferra, M. Baumert, and N. Bidargaddi. Heart rate variability for evaluating psychological stress changes in healthy adults: A scoping review. *Neuropsychobiology*, 2023.

[48] Ladakis Ioannis and Ioanna Chouvarda. Overview of biosignal analysis methods for the assessment of stress. *Emerging Science Journal*, 5:233–244, 04 2021.

[49] Muhammad Iqbal and Zhu Yan. Supervised machine learning approaches: A survey. *ICTACT Journal on Soft Computing*, 5(3):946–952, 2015.

[50] Tanveer Iqbal, Asad Elahi, Paul Redon, Pablo Vazquez, William Wijns, and Asim Shahzad. A review of biophysiological and biochemical indicators of stress for connected and preventive healthcare. *Diagnostics*, 2021.

[51] Arthur R. Jensen and William D. Rohwer. The stroop color-word test: A review. *Acta Psychologica*, 25(1):36–93, 1966.

[52] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[53] E. E. Kaczor, S. Carreiro, J. Stapp, B. Chapman, and P. Indic. Objective measurement of physician stress in the emergency department using a wearable sensor. In *Proceedings of the Annual Hawaii International Conference on System Sciences*, pages 3729–3738, 2020.

[54] Hye-Geum Kim, Eun-Jin Cheon, Dai-Seg Bai, Young Hwan Lee, and Bon-Hoon Koo. Stress and heart rate variability: A meta-analysis and review of the literature. *Psychiatry Investigation*, 15(3):235–245, March 2018. Epub 2018 Feb 28. PMID: 29486547; PMCID: PMC5900369.

[55] Clemens Kirschbaum, Klaus M. Pirke, and Dirk H. Hellhammer. The 'trier social stress test'—a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, 28(1-2):76–81, 1993.

[56] Sander Koelstra, Christian Mühl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

[57] Saskia Koldijk, Maya Sappelli, Suzan Verberne, Mark A. Neerincx, and Wessel Kraaij. The swell knowledge work dataset for stress and user modeling research. In *Proceedings of the ACM International Conference on Multimodal Interaction (ICMI)*, pages 291–298, 2014.

[58] Thomas Kosch, Jakob Karolus, Johannes Zagermann, Harald Reiterer, Albrecht Schmidt, and Paweł W. Woźniak. A survey on measuring cognitive workload in human-computer interaction. *ACM Comput. Surv.*, 2023.

[59] Chaspari Lab. Verbio dataset documentation (readme). Supplementary documentation distributed with the public VERBIO dataset release, 2023. Includes folder structure, signal formats, and continuous fused annotation specification at 1 Hz.

[60] Emmanouil Ladakis, Georgios Athanasopoulos, and Christos Papadopoulos. A lightweight cnn for multimodal stress detection using biosignal spectrograms. In *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC)*, 2022.

[61] Ioannis Ladakis, Dimitris Fotopoulos, and Ioanna Chouvarda. Integrative analysis of open datasets for stress prediction. *Journal of Medical and Biological Engineering*, 2025.

[62] Evgenia Lazarou and Themis P. Exarchos. Predicting stress levels using physiological data: Real-time stress prediction models utilizing wearable devices. *PMC*, 2024. Open access review.

[63] Richard S Lazarus, Anita DeLongis, Susan Folkman, and Rand Gruen. Stress and adaptational outcomes: The problem of confounded measures. 1985.

[64] L.Bajardi. Multisignal approach for stress and workload analysis. Master's thesis, Politecnico di Torino, 2022. Accessed from internal dataset.

[65] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[66] Gaang Lee, Byungjoo Choi, Houtan Jebelli, Changbum Ryan Ahn, and SangHyun Lee. Noise reference signal-based denoising method for eda collected by multimodal biosensor wearable in the field. *Journal of Computing in Civil Engineering*, 34(6):04020044, 2020.

[67] Z. Liu et al. Progress in data acquisition of wearable sensors. *PMC*, 2022. "In the first stage, it uses a second-order Chebyshev high-pass filter with a cutoff frequency of 0.5 Hz".

[68] Rosan Luijcks, Hermie J. Hermens, Lonneke Bodar, Catherine J. Vossen, Jim van Os, and Richel Lousberg. Experimentally induced stress validated by emg activity. *PLOS ONE*, 9(4):e95215, 2014.

[69] Dominique Makowski, Tam Pham, Zen J. Lau, James C. Brammer, Fabien Lespinasse, Hai Pham, Christopher Schölzel, and S. H. Annabel Chen. Neurokit2: A python toolbox for neurophysiological signal processing, 2021. Accessed: 2025-10-08.

[70] M. Malik and A. J. Camm. Heart rate variability. *Clinical Cardiology*, 13(8):570–576, 1990.

[71] Lokesh Malviya, Sandip Mal, and Praveen Lalwani. Eeg data analysis for stress detection. In *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*, pages 148–152, 2021.

[72] Elizabeth Mattera and Brian Zaboski. Rethinking the subjective units of distress scale: Validity and clinical utility of the suds. *Clinics and Practice*, 15(7), 2025.

[73] Catriona Miller, Theo Portlock, Denis M. Nyaga, and Justin M. O'Sullivan. A review of model evaluation metrics for machine learning in genetics and genomics. *Frontiers in Bioinformatics*, 4:1457619, 2024.

[74] Nir Milstein and Ilanit Gordon. Validating measures of electrodermal activity and heart rate variability derived from the empatica e4 utilized in research settings that involve interactive dyadic states. *Frontiers in Behavioral Neuroscience*, 14:148, 2020.

[75] Simon Mugisha, Mohammed Wassajja, Gerald Makiika Bamutiire, John Bashabe, and Mohammed Dahiru Buhari. Review and analysis on digital filter design in digital signal processing. *KIU Journal of Science, Engineering and Technology*, 3(2):12–20, December 2024. Open access.

[76] Vanessa Nilsen. An introduction to machine learning, 2018. Technical Report.

[77] María D. Olmo and Rafael Domingo. Emg characterization and processing in production engineering. *Materials*, 13(24):5815, 2020.

[78] Karthikeyan Palanisamy, Murugappan Murugappan, and Sazali Yaacob. Descriptive analysis of skin temperature variability of sympathetic nervous system activity in stress. *Journal of Physical Therapy Science*, 24(12):1341–1344, 2012. PMID: 24567668.

[79] Pooja Prajod, Bhargavi Mahesh, and Elisabeth André. Stressor type matters! exploring factors influencing cross-dataset generalizability of physiological stress detection. *arXiv preprint arXiv:2405.09563*, 2024.

[80] Pranav Prajod and S Ramya. Stress detection using deep learning on physiological signals: A cnn approach. *Biomedical Signal Processing and Control*, 71:102783, 2022.

[81] Guo-Jun Qi and Jiebo Luo. Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4757–4776, 2022.

[82] M. B. I. Raez, M. S. Hussain, and F. Mohd-Yasin. Techniques of emg signal analysis: detection, processing, classification and applications. *Biological Procedures Online*, 8:11–35, 2006.

[83] Oona Rainio, Jarmo Teuho, and Riku Klén. Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, 14:6086, 2024.

[84] S. Raja, R. Sinha, and A. Chandra. A review on stress inducement stimuli: Methods and protocols. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, 6(7):6557–6564, 2017.

[85] Keerthi G Reddy, P A Vijaya, and S Suhasini. Ecg signal characterization and correlation to heart abnormalities. *International Research Journal of Engineering and Technology (IRJET)*, 4(5):1212–1216, 2017. ISSN: 2395-0056 (e), 2395-0072 (p).

[86] Tatyana Reinhardt, Christian Schmahl, Stefan Wüst, and Martin Bohus. Salivary cortisol, heart rate, electrodermal activity and subjective stress responses to the mannheim multicomponent stress test (mmst). *Psychiatry Research*, 198(1):106–111, 2012.

[87] A. Sagaidachnyi. Human skin as a low-pass filter for thermal waves. *IEEE Transactions on Biomedical Engineering (or appropriate journal)*, 2019. Modeling skin's thermal dynamics as a filter; experimental verification.

[88] R Sathya and Annamma Abraham. Comparison of supervised and unsupervised learning algorithms for pattern classification. *International Journal of Advanced Research in Artificial Intelligence*, 2(2):34–38, 2013.

[89] Chhavi Saxena, Avinash Sharma, Rahul Srivastav, and Hemant Kumar Gupta. Denoising of ecg signals using fir & iir filter: A performance analysis. *International Journal of Engineering & Technology*, 7(4.12):1–5, 2018. Open access under Creative Commons Attribution License.

[90] Lawrence M. Schleifer, Thomas W. Spalding, Scott E. Kerick, James R. Cram, Ronald Ley, and Bradley D. Hatfield. Mental stress and trapezius muscle activation under psychomotor challenge: a focus on emg gaps during computer work. *Psychophysiology*, 45(3):356–365, 2008.

[91] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 400–408, 2018.

[92] Dongyeol Seok, Sanghyun Lee, Minjae Kim, Jaeouk Cho, and Chul Kim. Motion artifact removal techniques for wearable eeg and ppg sensor systems. *Frontiers in Electronics*, 2:685513, 2021.

[93] Claudia Setz, Bert Arnrich, Johannes Schumm, Roberto La Marca, Gerhard Tröster, and Ulrike Ehlert. Discriminating stress from cognitive load using

a wearable eda device. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):410–417, 2010.

[94] Claudia Setz, Bert Arnrich, Julia Schumm, Roberto La Marca, Gerhard Tröster, and Ulrike Ehlert. Discriminating stress from cognitive load using a wearable eda device. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):410–417, 2010.

[95] Amy L. Shilton, Robin Laycock, Sheila G. Crewther, and David P. Crewther. The maastricht acute stress test (mast): Physiological and subjective responses in anticipation, and post-stress. *Frontiers in Psychology*, 8:567, 2017.

[96] Tom Smeets, Sandra Cornelisse, Conny W. E. M. Quaedflieg, Tanja Meyer, Marko Jelicic, and Harald Merckelbach. Introducing the maastricht acute stress test (mast): A quick and non-invasive approach to elicit robust autonomic and glucocorticoid stress responses. *Psychoneuroendocrinology*, 37(12):1998–2008, 2012.

[97] Mohammad Soleymani, Johan Lichtenauer, Thierry Pun, and Maja Pantic. Mahnob-hci: A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012.

[98] Christopher Spiewak, Md Rasedul Islam, Md Assad-Uz-Zaman, and Mohammad Rahman. A comprehensive study on emg feature extraction and classifiers. *Open Access Journal of Biomedical Engineering and its Applications*, 1, 02 2018.

[99] D.P. Subha, P.K. Joseph, U.R. Acharya, and C.M. Lim. Eeg signal analysis: a survey. *Journal of Medical Systems*, 34(2):195–212, April 2010.

[100] F. Suni Lopez, N. Condori-Fernandez, and A. Catala. Towards real-time automatic stress detection for office workplaces. In *Information Management and Big Data: 5th International Conference, SIMBig 2018, Lima, Peru, September 3–5, 2018, Revised Selected Papers*, pages 273–288. Springer, 2019.

[101] Fatma M. Talaat and Rana M. El-Balka. Stress monitoring using wearable sensors: Iot techniques in medical field. *Neural Computing and Applications*, 35:18571–18584, 2023.

[102] Tomohiro Tanosoto, Taro Arima, Akio Tomonaga, Noboru Ohata, and Peter Svensson. A paced auditory serial addition task evokes stress and differential effects on masseter-muscle activity and haemodynamics. *European Journal of Oral Sciences*, 120(4):363–367, 2012.

[103] Tomohiro Tanosoto, Karina H. Bendixen, Taro Arima, John Hansen, Astrid J. Terkelsen, and Peter Svensson. Effects of the paced auditory serial addition task (pasat) with different rates on autonomic nervous system responses and self-reported levels of stress. *Journal of Oral Rehabilitation*, 42(5):378–385, 2015.

[104] Georgios Taskasaplidis, Dimitris A. Fotiadis, and Panagiotis D. Bamidis. Review of stress detection methods using wearable sensors. *IEEE Access*, 12:38219–38239, 2024.

[105] Abhishek Tiwari, Shrikanth Narayanan, and Tiago H Falk. Breathing rate complexity features for “in-the-wild” stress and anxiety measurement. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, 2019.

[106] W. Tramel, B. Schram, E. Canetti, and R. Orr. An examination of subjective and objective measures of stress in tactical populations: A scoping review. *Healthcare*, 11(18):2515, 2023.

[107] Vincent T. van Hees, Lukas Gorzelniak, Emilio C. Dean León, Michael Eder, Matias Pias, Saeed Taherian, Ulf Ekelund, Frida Renström, Paul W. Franks, Alexander Horsch, and Søren Brage. Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity. *PLOS ONE*, 8(4):e61691, 2013. Open access; includes high-pass and band-pass Butterworth filtering specifications for raw accelerometry.

[108] Christiaan H. Vinkers, Rachel Penning, Dirk H. Hellhammer, Joris C. Verster, Jan H. G. M. Klaessens, Berend Olivier, and Cor J. Kalkman. The effect of stress on core and peripheral body temperature in humans. *Stress*, 16(5):520–530, September 2013. Epub 2013 Jul 9.

[109] Carl L. von Baeyer, Tuula Piira, Christine T. Chambers, Manuela Trapanotto, and Lonnie K. Zeltzer. Guidelines for the cold pressor task as an experimental pain and stress induction technique. *The Journal of Pain*, 6(10):681–690, 2005.

[110] Gideon Vos, Kelly Trinh, Zoltan Sarnyai, and Mostafa Rahimi Azghadi. Generalizable machine learning for stress monitoring from wearable devices: A systematic literature review. *arXiv preprint arXiv:2209.15137*, 2022.

[111] Zhi-Hao Wang and Yu-Chan Wu. A novel rapid assessment of mental stress by using ppg signals based on deep learning. *IEEE Sensors Journal*, 22:21232–21239, 11 2022.

[112] David Watson, Lee Anna Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: The panas scales. *Journal of Personality and Social Psychology*, 54(6):1063–1070, 1988.

[113] Lan Xia, Weifeng Li, Wenya Zhang, and Xiaoshuang Shi. A physiological signal-based method for early mental-stress detection. *Biomedical Signal Processing and Control*, 46:18–28, 2018.

[114] Megha Yadav, Md Nazmus Sakib, Ehsanul Haque Nirjhar, Kexin Feng, Amir H. Behzadan, and Theodora Chaspari. Exploring individual differences of public speaking anxiety in real-life and virtual presentations. *IEEE Transactions on Affective Computing*, 13(3):1168–1182, 2022.

[115] Vahid Zakeri, Alireza Akhbardeh, Nasim Alamdari, Reza Fazel-Rezai, Mikko Paukkunen, and Kouhyar Tavakolian. Analyzing seismocardiogram cycles to identify the respiratory phases. *IEEE Transactions on Biomedical Engineering*, 63(12):2532–2540, 2016.

[116] Li Zhu, Panagiotis Spachos, Patrick C. Ng, Yiyu Yu, Yimin Wang, Konstantinos N. Plataniotis, and Dimitrios Hatzinakos. Stress detection through wrist-based electrodermal activity monitoring and machine learning. *IEEE Journal of Biomedical and Health Informatics*, 27(5):2155–2165, 2023. Epub 2023 May 4.

[117] Lili Zhu, Pai Chet Ng, Yuanhao Yu, Yang Wang, Petros Spachos, Dimitrios Hatzinakos, and Konstantinos N. Plataniotis. Feasibility study of stress detection with machine learning through eda from wearable devices. In *ICC*, pages 4800–4805, 2022.

Acknowledgements

I would first and foremost like to express my deepest and most heartfelt gratitude to my parents, whose unwavering support has been the foundation of every step of this journey and of countless others before it. Their presence has never faltered, neither in moments of enthusiasm nor in times of uncertainty, and knowing I could always count on them has been one of the greatest strengths I have carried throughout these years.

To my mother, I owe more than words can truly convey: for her boundless patience, her warmth, and her extraordinary ability to attend to every detail, even those I tended to overlook. Her constant encouragement, her reassuring words, and her remarkable capacity to understand me even in silence have been an endless source of comfort and motivation.

To my father, I am profoundly grateful for the example he has set through his values and actions, integrity, dedication, and perseverance. His way of facing challenges with clarity and determination has guided me through every important decision, reminding me of the value of commitment and responsibility. His unwavering faith in my abilities has encouraged me to believe in myself even in the most difficult moments.

Without their love, their sacrifices, and their unconditional support, reaching this milestone would not have been possible.

I warmly thank my brother for always being by my side, sharing both worries and moments of lightness, for his sincerity, his humour, and the calm perspective he has so often brought to challenging situations. I am equally grateful to his partner for her kindness and support, for her thoughtful presence throughout these years, and for the genuine interest she has always shown in my work and my wellbeing.

I would also like to extend my heartfelt thanks to the rest of my family, my uncles and aunts, my cousins, and especially my grandmothers and grandfathers, for their constant affection and the sincere interest they have always taken in my studies. Their encouragement, warmth, and the profound sense of belonging they have given me have accompanied me throughout this journey in ways I deeply cherish.

A very special thank you goes to my girlfriend, whose support over these months has meant more than she knows. Her closeness, understanding, and encouragement, even in the short time we have shared, have helped me face the most demanding moments of this journey with greater serenity and motivation. I am truly grateful for the patience, kindness, and genuine interest she has shown in both my work and in me.

My heartfelt thanks also go to my friends, who have walked alongside me through this path with friendship, humour, and countless moments of respite from study, and to my colleagues at work, for their understanding, support, and the stimulating discussions that have enriched both my professional and academic growth.

Finally, I wish to express my sincere gratitude to my thesis supervisor for his guidance, availability, and constructive feedback throughout the development of this work. His scientific rigour and insightful suggestions have been instrumental in shaping this thesis and in helping me grow both as a researcher and as a professional.