Politecnico di Torino

Politecnico di Torino

Master's Degree in Biomedical Engineering (LM-21)

A.y. 2026/2027

Graduation Session March 2026

Multimodal Classification of Interactive and Perceptive Stressors using Time-Series Foundation Models

Supervisors:

Prof.ssa Gabriella Olmo

Ing. Giulia Masi

Candidate:

Francesco Tavella

Abstract

Stress is a pervasive psychological state that can profoundly impact people's lives and have a negative impact on their well-being, from both a mental and physical health aspect. Long exposure to stress can cause severe implications, not only for the single subject but also for society in terms of social and healthcare costs. Research in this field has been growing considerably over the last years, helped by the proliferation of wearable technologies which has provided the opportunity to monitor physiological signals continuously, establishing a key pillar of Human-Computer Interaction (HCI) field or in Affective Computing sector. However, most studies have focused on the binary detection of stress vs. relaxed state. Instead, stress can originate from stimuli of a very different nature, and trigger a variety of physiological responses, which are not the same for all people.

That is why this study focuses on differentiating the nuance between two different stressor stimuli: interactive stimuli (for cognitive load) and perceptive stimuli (emotional arousal). Furthermore, physiological signals exhibit high inter-subject variability when exposed to stressor stimuli, leading to severe performance degradation when models are tested on unseen individuals (i.e., Domain Shift). To address these challenges, this work proposes a novel multimodal framework for subject-independent stressor discrimination. A pre-trained Time-Series Foundation Model is leveraged to extract robust embeddings from three physiological signals: Electrodermal Activity (EDA), Photoplethysmogram (PPG), and Electrocardiogram (ECG). To explicitly address the domain shift, a signal-adaptive adversarial neural network (DANN) is introduced and its performances rigorously compared with a basic MOMENT architecture without explicit adaptation. The proposed frameworks are trained and tested on the CLAS dataset (a well-known free public dataset in the field of stress detection) using a subject-independent cross validation scheme. The Multimodal Late Fusion framework achieves competitive performance, with a mean Accuracy of 0.7260 (+/- 0.1455) across all folds. Moreover, the results highlight that ECG signal emerges as the most discriminative single modality for stressor discrimination but the Late Fusion of the three signals yields a significant improvement over the single ECG model ($p<0.05$). Statistical analysis reveals that no significant difference between DANN-model and baseline Multimodal Late Fusion model ($p> 0.05$): while the DANN model improves the stability by lowering the variance, the average metrics remain comparable to the baseline model. These results offer an interesting insight: large-scale pre-trained models such as MOMENT inherently encode robust, subject-independent representations, effectively mitigating the need for complex adversarial adaptation and positioning them as promising candidates for applications in real-life scenarios.

Keywords: Stress Discrimination, Time-Series Foundation Model, Multimodal Classification

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In today's fast-paced society, stress has become a silent and omnipresent traveling companion. Although in moderate doses it can be a vital natural stimulus for dealing with everyday challenges, its chronicization has become a major threat to public health and individual psychophysical well-being. Until only recently, understanding and objectively measuring this state was a task relegated to complex clinical examinations or questionnaires, heavily influenced by subjective perception. However, nowadays, we are witnessing a revolution driven by the widespread diffusion of wearable devices. Increasingly discreet devices, such as smartwatches and smart bands, are now able to listen to the invisible language of our bodies, with a continuous stream of data of multiple biological signals. This availability of information, in combination with the raise of deep learning models and Artificial Intelligence (AI), has given a strong impetus to decode the human emotional and cognitive state. It is at this crossroads between biology, psychology, and automatic learning that this thesis work fits in. The aim is not simply to detect the presence or absence of stress, but to go further exploring how the latest AI models can learn to distinguish the subtle nuances of our physiological responses when faced with stimuli of a different nature, laying the foundations for future technologies that are increasingly empathic, proactive, and personalized.

1.1 Stress Monitoring in the Age of Wearables

1.1.1 Stress: definition and impact on the society

In common language, the term stress is frequently associated with a purely negative connotation, understood as a synonym for anxiety or psychological malaise. However, from a neurophysiological point of view, stress is not an emotion, but a complex and fundamental biological adaptation reaction. As defined pioneeringly

by Hans Selye in his seminal text "The Stress of Life" (1956)[1], stress represents the organism's "non-specific response to any demand made upon it". This reaction, known as General Adaptation Syndrome (GAS), which is a general physiological reaction to a wide range of stressors, is an essential mechanism for human survival, as it prepares the individual to face threats or challenges (fight-or-flight response). When people are exposed to stressful events, physiological responses are triggered under the control of the Autonomic Nervous System (ANS): the ANS loses its state of equilibrium, called homeostasis, and the parasympathetic branch becomes inhibited, while the Sympathetic Nervous System (SNS) becomes hyperactive, triggering immediate peripheral reactions such as vasoconstriction, increased heart rate, and bronchodilation. The term may be used in reference to external (way of living, relationship problems, financial problems) or internal (personality structure, way of thinking) affairs triggering negative emotions (worry, fear) and associated physiological (i.e., bodily) changes. The common notion testifies that the experience of stress is related both to the perception and subjective evaluation of an event, as well as to the perception of the bodily changes triggered by it.

A physiological stress response refers to the bodily changes elicited by environmental events or conditions, known as stressors. This response comprises physiological processes responsible for: (i) processing the potential stressor and organizing an adaptive response, (ii) mobilizing the myoskeletal system in order to prepare and execute motor actions, and (iii) preparing the body to withstand injuries and increased metabolic demands[2]. It is crucial to distinguish, as also highlighted in the recent literature on wearable devices[3], between a positive form of stress, called 'eustress', which stimulates creativity and problem-solving, and a negative, uncontrolled form, called 'distress', as illustrated in Figure 1.1. The When the stress response becomes chronic or out of proportion to the stimulus, it ceases to be a protective mechanism and becomes a pathological risk factor.

The impact of distress on public health is alarming today. According to the World Health Organization, stress-related disorders constitute a growing public health concern. Prolonged exposure to cortisol levels and constant activation of the sympathetic nervous system are correlated with serious physical disorders, including hypertension, cardiovascular diseases, and a weakened immune system, as well as mental disorders such as depression and anxiety [4]. In addition to the individual cost in terms of healthy, the impact becomes critical when associated with people with special occupations such as police, or doctors, where they are exposed to daily and highly stressful situations in the working environment.

Furthermore, in contemporary society, this physiological burden is severely amplified by the pervasive phenomenon of 'Digital Stress': relentless hyperconnectivity, information overload, and the incessant demand for cognitive multitasking are some of the main non-physical stressors. This phenomenon is particularly evident in high-pressure work environments, as demonstrated by Hossein et al. in their

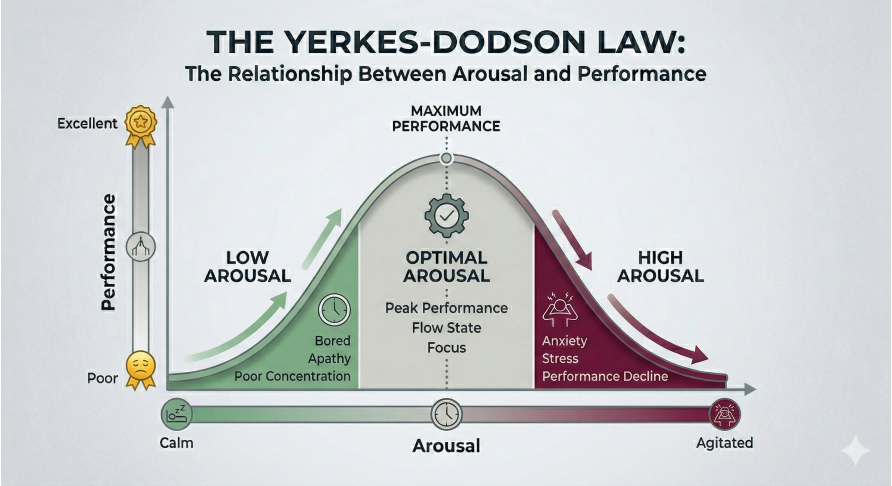


Figure 1.1: Illustration of the Yerkes–Dodson law reporting the association between arousal levels and human performance (Conceptual figure; adapted from the Yerkes–Dodson principle).

study of nurses' conditions during the COVID-19 pandemic[5], where continuous monitoring of biosignals revealed how psychological and environmental factors converge in generating stress states, which need to be dealt with immediately.

1.1.2 The technological evolution of stress monitoring

Until the past decade, the measurement of physiological responses associated with stress was almost an exclusive research domain of both laboratories and clinics. This approach guaranties good accuracy and high signal-to-noise ratio but presents several limitations that were insurmountable for long-term analysis. The main criticism lay in the lack of ecological validity of the measurements: the subject is forced to wear cumbersome cables, electrodes, and devices located in an aseptic and artificial environment (such as the clinic) altering the emotional state of the subject and making it very difficult to isolate and capture the true stress effects. However, rapid advances in wearable technologies, mobile sensors, and Internet of Things (IoT) infrastructures have radically transformed stress monitoring paradigms. Moreover, the improvement in battery efficiency, wireless data transmission capabilities, and the miniaturization of biosensors have enabled continuous and unobtrusive acquisition of physiological signals in everyday scenarios. Devices such as wrist-worn photoplethysmography sensors, compact ECG patches, and wearable electrodermal activity monitors now allow large-scale in-the-wild data collection with unprecedented temporal resolution. An example of a wearable device is provided in Figure 1.2, showing the Empatica E4 wristband.

So, with the increasingly widespread use of wearable sensing devices, the monitoring of physiological data is no longer limited to laboratory studies, but can be performed continuously in naturalistic contexts. This change has contributed to the emergence of publicly available datasets that capture physiological responses across a broader range of user, contexts, and stimuli. An example is the work by Hosseini et al.[5], in which biosignals were acquired directly in the workplace. They monitored and collected specific physiological variables, such as electrodermal activity, Heart Rate, and skin temperature of the nurse in a hospital during the COVID-19 pandemic.

Figure 1.2: Empatica E4 wristband, a wearable device to collect physiological signals.

These datasets have been instrumental in advancing data-driven approaches for stress detection, particularly through modern Machine Learning techniques. Nevertheless, signals acquired in real-world scenarios are subjected to higher noise, due to motion artifacts and variability in the placement of the sensors, and so

are more heterogeneous than laboratory data. Consequently, robust models with strong inter-subject generalization are essential for in-the-wild stress detection applications. Figure 1.3 below summarizes the technological evolution for stress monitoring.

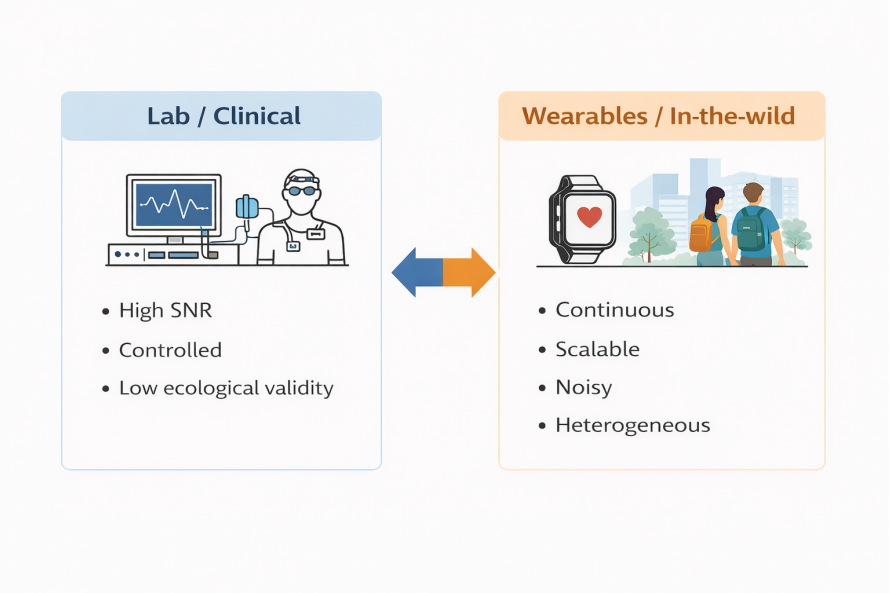


Figure 1.3: Conceptual comparison between laboratory/clinical and wearable in-the-wild monitoring.

1.1.3 Affective Computing and its matter for stress research

Affective Computing is a multidisciplinary research field whose goal is to enable computational systems to recognize, interpret, and respond to human affective and cognitive states. While early work in this area was strongly motivated by the idea that human-computer interaction could be improved if technology became sensitive to users' emotions, modern Affective Computing has evolved into a broad methodological framework that integrates psychology, physiology, machine learning, signal processing, and interaction design. A key aspect of Affective Computing is the assumption that latent internal states, such as stress, anxiety, engagement, cognitive load, or emotional arousal, are manifested through measurable signals. These signals can be behavioral (e.g., facial expressions, speech prosody, interaction patterns) or physiological (e.g., EDA, ECG/PPG, respiration) and can be combined to build models that can infer the user's state. A canonical view of Affective Computing was introduced by Rosalind Picard in her seminal book Affective Computing (MIT Press, 1997). Although the original vision focused on emotion-sensitive machines, the field has since expanded to include stress and cognitive load sensing as main

targets, as these states strongly modulate decision-making, attention, performance, and well-being. In other words, stress is not only a clinical construct, but also a human factors variable: it influences how people interact with technology, how effectively they perform tasks and their propensity to make mistakes in challenging environments.

From a methodological point of view, modern Affective Computing is increasingly based on data-driven approaches. A widely cited survey by Calvo and D'Mello describes affect detection as an interdisciplinary problem and discusses how models must take into account both theoretical assumptions and practical constraints, especially when physiological data are used to infer internal states[6]. This perspective is particularly relevant to stress research: physiological measures are interesting because they provide objective proxies for autonomic activation, but the mapping between physiological patterns and psychological constructs is not biunivocal. Stress-related physiological changes depend not only on the stressor, but also on context, individual appraisal, coping style, and baseline differences. Therefore, Affective Computing does not just "measure emotions" but builds computational inference systems that must be validated and interpreted carefully.

Over the last decade, Affective Computing has been increasingly influenced by the proliferation of wearable devices and mobile sensing. The availability of continuous streams of biosignals has enabled the development of models that operate beyond controlled experiments in the laboratory, with the ambition to support real-world applications such as mental well-being monitoring, personalized interventions, adaptive learning, and safety-critical assistive systems. This shift from offline assessment to continuous inference directly motivates the need for robust machine learning frameworks for physiological time series.

1.1.4 The role of HCI in real-time stress detection

Human-computer interaction (HCI) provides the conceptual and methodological basis for translating sensing and inference technologies into usable, reliable and meaningful systems for people. While Affective Computing defines the computational goal of inferring users' latent states, HCI addresses how such inferences should be integrated into interactive systems and how they affect user experience, performance, trust and ethical acceptability. In practical terms, HCI shifts the focus from 'can we detect stress?' to 'how should a system behave when stress is detected?' and 'what are the consequences of acting on that inference?'. A useful lens for understanding this integration is the idea of physiological computation, in which real-time psychophysiological measurements are used as input to an adaptive system. Fairclough's seminal review formalizes the main challenges and design issues of physiological computation, including: the complexity of psychophysiological

inference, validation of inferred states, how user states should be represented computationally, how systems should intervene (explicitly or implicitly) and the ethical implications of physiological monitoring[7]. This work is often cited as a milestone in what is sometimes called "PhysioHCI": the branch of HCI that uses physiological signals to build interactive systems that respond to the user's internal state. In real-time physiological sensing, the interaction design problem is inseparable from the technical pipeline. Real-time settings impose constraints that do not exist in offline laboratory analysis: algorithms must operate continuously on streaming data, with low latency and limited computation. At the same time, signals acquired by wearable devices in everyday environments are noisy and heterogeneous. Motion artifacts, sensor displacement, variable skin contact, and differences between devices can distort signals and compromise the reliability of inferences. For this reason, HCI research emphasizes that physiological systems should be evaluated not only for classification accuracy, but also for robustness, stability over time, calibration requirements, and user burden (e.g. comfort, ease of use and intrusiveness). Another fundamental HCI concept relevant in this context is that adaptive systems should "close the loop" in a responsible manner. In stress-sensitive interfaces, the system could adapt the difficulty of tasks, reduce interruptions, change the intensity of feedback, or require micro-pause when the user is under a high load. However, these adaptations must be designed carefully: incorrect inferences may lead to inappropriate interventions that damage performance or increase frustration.

Therefore, HCI introduces evaluation criteria that go beyond predictive metrics, such as perceived usefulness, transparency, controllability, and user trust. Furthermore, physiological sensing raises strong ethical considerations that HCI explicitly treats as first-class design requirements. Physiological signals can be highly sensitive and can reveal intimate aspects of health, mental state or behavioral patterns. Consequently, real-time stress monitoring systems must address issues such as privacy, informed consent, data minimization, and potential for abuse, especially in work or institutional settings. Importantly, ethical concerns are not an "add-on": they determine which detection modes are acceptable and which implementation scenarios are realistic. Finally, from a machine learning perspective, the implementation conditions relevant to HCI amplify the generalization problem. Real systems have to operate on unseen users during training, in different contexts and acquisition conditions. This directly links physiological HCI to the issue of domain shift and motivates approaches that learn robust, subject-independent representations. In this thesis, this requirement is central: stressor discrimination is assessed in a strictly subject-independent context precisely because this context is most in line with real-world HCI applications.

1.1.5 State of the Art in Stress Detection

Over the last decade, physiological stress recognition has become a central topic in Affective Computing and wearable sensing due to the availability of biosignals such as ECG, PPG, and EDA combined with the growing interest in real-world mental well-being monitoring. Despite these advances, the dominant formulation in the literature still treats stress detection primarily as a binary classification task, typically "stress vs. non-stress" or "stress vs. relaxation". A representative example is the WESAD benchmark, one of the most widely used datasets for stress and affect detection using wearable devices, where stress discrimination is commonly operationalized as a binary or multi-class coarse problem built around controlled conditions[8]. Recent critical reviews further confirm this trend: many Machine Larning (ML) pipelines focus on detecting stress exposure (i.e. whether a person is in a stressful condition) rather than modeling the heterogeneous causes and mechanisms of stress, especially in ecologically valid contexts. For example, Mentis et al. discuss how AI/ML methods have performed well in controlled studies, but point out that translating these solutions into practical monitoring remains difficult, in part due to oversimplified definitions of the problem and inconsistent operation of "stress" between studies [9].

Although binary paradigms are convenient for benchmarking, they also impose conceptual and practical limitations:

- Conceptual limitation: binary labels implicitly treat stress as a single, uniform state, ignoring that different stressors may induce distinct physiological patterns.

- Application limitation: real-world applications often require understanding why stress occurs (e.g. cognitive overload vs. emotional arousal), as interventions and adaptation strategies depend on the type of stressor. A system that merely reports "stress detected" provides limited operational value in areas such as learning technologies, safety-critical monitoring, or personalized mental health support.

In short, binary detection is a useful starting point, but risks reducing the multiple underlying mechanisms to a single label, reducing interpretability, and limiting downstream decision-making.

1.1.6 Stress as a Heterogeneous Response to Different Stressors

Stress is not a monolithic concept: it derives from heterogeneous stressors and is influenced by context, assessment, and coping strategies. Even when two situations

are subjectively perceived as stressful, they can activate partially different physiological pathways and temporal dynamics. This becomes particularly relevant in physiology-based inference, as biosignals reflect not only the "stress intensity", but also the type of autonomic activation and response pattern over time. Affective science often models affective responses in terms of dimensions such as arousal and valence, formalized in influential works such as Russell's circumplex model[10] of affect and shown in Figure 1.4. This theoretical framework supports the idea that "emotional stress" is often characterized by changes in arousal (and valence), which can be provoked by perceptual stimuli such as images or videos designed to elicit affective responses.

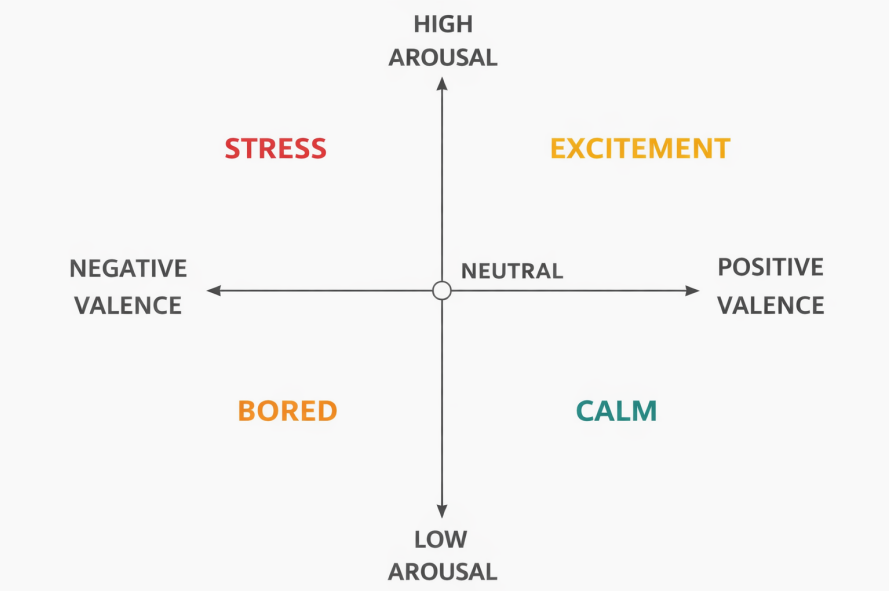


Figure 1.4: Circumplex representation of affect in the arousal–valence plane. Affective stimuli can be positioned across four quadrants (high/low arousal × positive/negative valence), providing a conceptual basis for designing perceptive tasks that elicit distinct emotional responses. (Conceptual figure based on Russell's circumplex model of affect)

In contrast, cognitive stress is commonly related to mental workload, sustained attention, and performance under time pressure (e.g. arithmetic, logical tasks, Stroop-type paradigms). These stressors typically require active involvement and demanding control and are often associated with strong task-related reactivity in cardiovascular measures. Importantly, the psychophysiological literature distinguishes between stressors not only on the basis of the stimulus modality, but also on the basis of the coping demands they impose. A widely discussed distinction is that between:

- Active coping stressors: situations in which the individual can act in response to the demand (e.g., problem solving, mental arithmetic, tasks under time pressure);

- Passive coping stressors: situations in which the individual primarily endures or observes stimuli without direct control (e.g., exposure to emotionally evocative content, unpleasant perceptual stimuli).

Evidence suggests that active versus passive coping may produce different cardiovascular response patterns and mechanisms. For example, Zanstra & Johnston examine how active handling tasks were associated with a more "myocardial" response pattern, whereas passive handling may be more strongly related to vascular response patterns[11]. This perspective directly supports the argument to go beyond a single stress label: if distinct stressors can induce systematically different physiological patterns, then modeling stress as a single binary variable may obscure the meaningful structure of the data.

1.1.7 Stressor Discrimination as an Actionable Objective for Real-World Systems

Given the limitations of binary stress detection, an increasingly relevant objective is the discrimination of stressors: instead of predicting the presence of stress, the model aims at inferring which type of stressor is driving the physiological response. This is particularly useful in real-world contexts in which different stressors require different responses or interventions from the system. In this thesis, the discrimination of stressors is formulated as a supervised classification problem between two categories:

- Interactive stressors (cognitive load; active coping): tasks that require active participation, performance under time pressure and prolonged cognitive control;

- Perceptual stressors (emotional arousal; passive coping): stimuli that elicit affective responses primarily through perception (e.g., affective images or video clips), without requiring active problem solving.

This formulation is well aligned with common affective benchmarks and experimental designs. For example, the DEAP dataset, widely used for emotion analysis using physiological signals, structures affect in terms of self-reported arousal and valence while recording peripheral physiological signals[12].

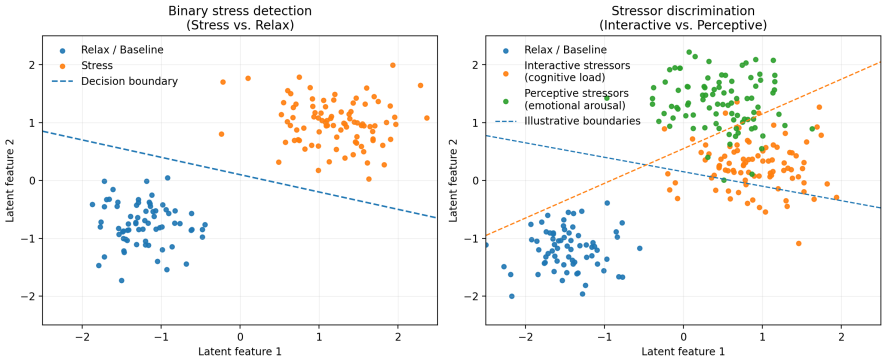


Figure 1.5: From binary stress detection to stressor discrimination: a schematic latent-space illustration showing how a binary "stress vs relax" boundary (left) can be insufficient when stress responses arise from different stressor categories (right)

Similarly, the availability of open-access multimodal datasets for stress and emotion has been recognized as a key factor for progress in this field; Ometov et al. provide a focused review on open-access datasets, modalities and challenges in stress and emotion recognition[13]. Why this is important in real-life applications The discrimination of stressors is not only conceptually richer, it is also more applicable to applied systems:

- Adaptive learning and training systems: cognitive overload (interactive stress) may suggest reducing the difficulty of tasks, increasing the time spent on explanations or scheduling micro-pause; emotional arousal (perceptual stress) may require different strategies (e.g. suggestions for emotion regulation, environment adjustments).

- Safety-critical HCI (e.g., driving, monitoring of operators): cognitive overload may imply impending performance deterioration and increased likelihood of error, whereas emotional arousal may be related to risk-taking, distraction or panic-like reactions, each requiring distinct support mechanisms.

- Technologies for mental health and well-being: distinguishing whether physiological stress is caused by cognitive demands or affective triggers can improve the personalization and interpretability of interventions and reduce the ambiguity of "stress detected" notifications.

In short, discrimination of stressors reformulates physiological monitoring from an approximate detection task to a mechanism-aware inference problem, better suited to the way real systems must support users in dynamic environments.

1.2 Technical Challenge: Inter-Subject Variability and Domain Shift

1.2.1 The Physiological Inter-Subject Variability

A fundamental obstacle in the interpretation of physiological stress is the marked inter-individual variability of biosignals. Even under identical experimental conditions, individuals show different baseline levels and magnitudes of reactivity due to factors such as age, gender, fitness, autonomic balance, circadian rhythms, medication, lifestyle, sensor placement and idiosyncratic stress assessment. This variability is not a marginal effect: it can dominate the statistical structure of the data and can cause models to learn "who the person is" rather than "what state they are in". In the field of wearable affective computing, the impact of between-subject variability has been explicitly highlighted as one of the main reasons why generalization remains difficult. A comprehensive investigation of affect recognition based on physiological signals discusses how between-subject variance substantially degrades performance in the case of subject-independent splits, pointing out that model accuracy often decreases when tested on unknown users and that the between-subject variance can be very large [14]. This phenomenon is particularly relevant for the main signals used in stress research:

- EDA (electrodermal activity): the tonic level is highly subject dependent (e.g. baseline skin conductance varies greatly) and phasic responses depend on both physiological traits and stimulus sensitivity;
- ECG: the dynamics and morphology of the heart rhythm have strong biometric characteristics; resting heart rate and heart rate variability vary greatly between individuals and stress-induced changes may manifest differently depending on autonomic regulation;
- PPG: its morphology is influenced by peripheral vascular tone and the quality of sensor contact; motion artifacts and local physiological factors (temperature, vasoconstriction) often amplify user-to-user variability.

Consequently, a model trained on one group of subjects may fail on a new user not because the stress physiology is absent, but because the model's decision limit becomes overly tuned to the base models and acquisition characteristics of the training population. This problem becomes even more severe in real-world contexts, where wearable device data are noisier and the recording context is less controlled; recent reviews on stress prediction using wearable devices point out that the shift to real-time monitoring "under real-world conditions" increases heterogeneity and complicates robust inference[15].

1.2.2 Domain Shift in Subject Independent Scenarios

In machine learning terms, inter-subject variability induces a form of domain shift: the statistical distribution of physiological features differs between the training set (subjects seen during training) and the test set (new users). When evaluation protocols are subject-dependent (e.g., random splits across segments), train and test samples share the same individuals, often inflating performance estimates. By contrast, subject-independent evaluation (e.g., Leave-One-Subject-Out, or strict subject-level train/validation/test splits) better reflects real deployment conditions, where the system must operate on unseen users.

This gap between subject-dependent and subject-independent performance has been repeatedly observed across physiological affect and stress tasks. Benchmark datasets such as WESAD—widely used for stress and affect recognition—are often used to evaluate models, yet the results can change dramatically depending on whether the split is subject-wise or sample-wise[8].

In applied contexts, the "domain" may change not only due to the subject identity but also due to sensor type, placement, activity level, and environmental conditions. Consequently, domain shift appears at multiple levels:

- Subject shift: new physiological baselines and personal reactivity patterns;
- Sensor shift: different devices or placements alter signal morphology and noise characteristics;
- Context shift: movement, posture, temperature, and daily routines affect signal quality and physiology.

A practical implication is that standard supervised learning tends to produce models that work well in-distribution but degrade sharply out-of-distribution. In adjacent physiological applications (e.g., driver monitoring), cross-subject generalization is also recognized as challenging and has motivated transfer learning and cross-subject strategies[16].

1.2.3 Possible Mitigation Strategies and their Limitations

To address inter-subject variability and domain shift, stress detection research has historically relied on three broad families of strategies:

1. Feature engineering and normalization. Classic approaches extract handcrafted features (e.g., HRV indices from ECG/PPG, statistical and event-related descriptors from EDA such as SCL/SCR measures) and apply normalization techniques (z-scoring, baseline correction, subject-wise rescaling). These methods can partially reduce between-subject differences, but they often require careful tuning, are sensitive to signal quality, and may not transfer reliably

across devices and contexts without recalibration. As wearable datasets and computational resources have grown, the literature has increasingly moved away from rigid handcrafted pipelines toward end-to-end deep learning, where representation learning replaces manual feature design. This methodological shift has been explicitly investigated in physiological affect/stress modeling, with works asking whether deep neural architectures can effectively reduce or even replace feature engineering without sacrificing performance[17]. Nonetheless, even end-to-end models remain vulnerable to cross-subject distribution shifts when trained on limited, biased, or noisy datasets.

2. Personalization and calibration. Some systems incorporate subject-specific calibration steps to learn individual baselines. Although effective, this approach reduces scalability and conflicts with the goal of plug-and-play stress monitoring.

3. Domain adaptation. Domain adaptation seeks to learn representations that transfer across domains (e.g., across subjects), often by aligning feature distributions. Among the most influential approaches, Domain-Adversarial Neural Networks (DANNs) propose learning features that are discriminative for the main task but invariant with respect to domain identity, using adversarial learning with a Gradient Reversal Layer[18].

While these approaches can improve robustness, they also have practical limitations in stress detection. Physiological datasets are often relatively small compared to typical deep learning regimes, labels can be noisy or coarse, and adversarial training can be unstable or sensitive to hyperparameters. Moreover, when the domain is "subject identity", the discriminator may easily exploit biometric cues in signals such as ECG and PPG, making the adversarial game difficult to balance.

This motivates the key question behind the methodological design of this thesis: can modern representation learning—particularly time-series foundation models—reduce the need for explicit adaptation by producing embeddings that are intrinsically robust to subject shift? The next section introduces this paradigm shift and explains why large-scale pre-trained models may offer an alternative route to subject-independent physiological inference.

1.3 Time-Series Foundation Model: a Novel Approach

1.3.1 From Handcrafted Features to Representation Learning

Early research on physiological stress detection relied heavily on handcrafted features extracted from ECG, PPG, and EDA, such as time-domain and frequency-domain HRV metrics, statistical descriptors, and event-related features, followed by traditional classifiers (e.g., SVM, Random Forests). This pipeline has two historical advantages: interpretability (the features have physiological meaning) and feasibility on small datasets. However, it also has important limitations. Manual features are often sensitive to preprocessing choices and signal quality and may fail to capture complex temporal dependencies or subtle patterns distributed over time. Deep learning approaches were introduced to reduce dependence on manual feature engineering by learning representations directly from raw or lightly processed signals. However, training deep models from scratch is often limited by the typical characteristics of physiological datasets: limited number of subjects, domain heterogeneity, and label noise or rough annotations. As a result, deep models may overfit to dataset-specific artifacts and exhibit fragile generalization in subject-independent contexts. This tension, between the expressive power of deep learning and the scarcity/variability of biosignal dataset data, has motivated the search for methods capable of learning robust representations without requiring large task-specific labeled corpora. A broader perspective on recent trends in deep time series modeling and the shift from conventional pipelines to representation learning is discussed in modern reviews of deep time series models[19].

1.3.2 Time-Series Foundation Models

Foundation Models (FMs) represent a paradigm shift in machine learning: instead of learning a model from scratch for each task, a large model is pretrained on massive, diverse data to learn general-purpose representations, and then adapted (or even used directly) for downstream tasks. While this idea has long been established in NLP and computer vision, it has only recently gained strong momentum in time-series research.

In time-series analysis, the motivation for foundation models is particularly compelling. Time-series data across domains (healthcare, engineering, finance, IoT) share common structural properties—trends, seasonality, autocorrelation, abrupt changes, periodicity, and noise characteristics—even if the semantics differ. Pretraining can enable a model to internalize general time-series "primitives," which can then transfer to downstream tasks such as classification, forecasting, anomaly

detection, and representation learning.

A comprehensive tutorial and survey specifically focused on foundation models for time-series analysis outlines the main design choices: model architectures (often Transformer-based), pretraining objectives (e.g., masked modeling), adaptation mechanisms (fine-tuning, prompting, linear probing), and the crucial role of large-scale time-series corpora[20].

From the standpoint of physiological inference, the key promise is that pretrained time-series models may produce embeddings that are more robust to nuisance variability—including sensor noise, acquisition differences, and subject-specific baselines—because the model has learned representations across many distributions during pretraining. This property is particularly attractive when the downstream task is evaluated in strict subject-independent conditions, where distribution shift is the norm rather than the exception.

1.3.3 MOMENT: a Foundation Model for Feature Extraction

In this emerging landscape, MOMENT is a family of open-ended time series base models designed for generic time series analysis. MOMENT is pre-trained using a masked modeling objective on a large and diverse set of time series datasets (the "Time-Series Pile"), enabling the model to learn general temporal representations that can be transferred across tasks and domains[21]. A key practical advantage of base models such as MOMENT is that they can be used in multiple regimes: Fine-tuning, where the pre-trained model is adapted end-to-end to the target dataset, typically improving task performance but increasing computational cost and the risk of overfitting on small datasets. Frozen feature extraction, where the pre-trained model is used to compute embeddings and only a lightweight classifier is trained on top (often called linear probing or shallow adaptation). For physiological stress modeling, the frozen feature extraction approach is particularly interesting. First, it reduces computational complexity and makes the pipeline more feasible for experimentation and eventual real-time implementation. Second, it allows for scientific verification of a central hypothesis: whether pre-trained representations are already robust enough to be generalized to unseen subjects. In other words, if frozen embeddings extracted from a base model support stable subject-independent classification, this suggests that pretraining has captured the invariances that traditional subject-specific models struggle to learn. In this thesis, MOMENT is specifically adopted in this "frozen embedding" regime as the representation engine for EDA, PPG, and ECG. This design is intentionally aligned with the subject-independent goal: instead of fitting high-capacity temporal encoders directly on CLAS (which risks learning the idiosyncrasies of the subject and dataset), the model leverages a pre-trained representation space and focuses downstream learning

on discriminating stressor types.

1.3.4 Multimodal Approach and Late Fusion

In real-world wearable monitoring, relying on a single physiological channel is often risky. Each sensor captures only a partial view of autonomic activation and is affected by mode-specific artifacts:

- EDA is highly sensitive to sympathetic arousal but slow and influenced by skin properties and electrode contact;
- PPG is easy to acquire but strongly affected by motion artifacts and peripheral circulation changes;
- ECG is physiologically rich but can be more intrusive and also contains strong biometric signatures.

The broader literature on multimodal machine learning formalizes this logic and classifies fusion strategies into early (feature-level), late (decision-level), and hybrid fusion, emphasizing that different levels of fusion involve a trade-off between expressiveness, robustness, synchronization requirements, and handling of missing modalities[22]. In biomedical contexts, reviews on multimodal deep learning[23] highlight that late fusion can be beneficial when modalities have different noise profiles, different sampling characteristics, or inconsistent availability, as each modality-specific model can learn its own optimal decision boundary and the fusion stage can reduce the weight of unreliable channels. This consideration is particularly relevant for physiological stress inference. Several studies in stress-related fields adopt late fusion at the decision or weighted level to improve reliability over single-modality baselines, especially under realistic detection conditions (e.g., multimodal stress detection experiments combining physiological streams and other signals[24]. For these reasons, the framework of this thesis adopts a multimodal late fusion strategy rather than early fusion. Conceptually, this design fits physiological reality: each signal carries a different part of the stress response and suffers from different artifacts. Methodologically, late fusion reduces the risk that a noisy modality will contaminate the shared representation too early. From a practical standpoint, it supports robustness in subject-independent scenarios, allowing the system to rely more heavily on the most informative modality (often ECG), while still leveraging complementary information from EDA and PPG when available. Finally, multimodality naturally ties in with the core model paradigm: by using the same pre-trained representation engine (MOMENT) across different physiological channels, the framework builds a consistent embedding space for each modality while preserving the flexibility to integrate them at the decision level, maximizing robustness without requiring heavy end-to-end multimodal training.

1.4 Objectives and Contributions of this Thesis

This thesis investigates the subject-independent discrimination of stressor types using multimodal wearable physiology. Moving beyond the common binary formulation of "stress vs. relax," the work focuses on differentiating the stress responses elicited by interactive stressors (cognitive load, active coping) and perceptive stressors (emotional arousal, passive coping). To address the generalization gap caused by inter-subject variability, the proposed approach leverages modern representation learning through a pre-trained time-series foundation model and systematically evaluates whether explicit domain alignment is necessary.

The main objectives of this thesis are:

O1 — Stressor discrimination instead of binary detection: Formulate and evaluate a supervised classification task that distinguishes interactive vs. perceptive stressors from physiological signals.

O2 - The use of a time-series Foundation Model in a frozen modality: This represents an unexplored approach in physiological signal analysis. The time-series foundation model is employed as a feature extractor for physiological signals. The main objective here is to assess whether such a strategy can yield meaningful results and potentially open new directions for research in this field.

O3 — Subject-independent generalization as a primary requirement: Assess model performance under strict subject-independent protocols, ensuring that the test subject is never observed during training and that validation is performed on separate held-out subjects to prevent leakage.

O4 — Multimodal integration for robustness: Quantify the contribution of combining EDA, PPG, and ECG through multimodal fusion, and test whether fusion improves reliability compared to single-modality classifiers.

O5 — Testing the need for explicit domain alignment: Evaluate whether adversarial domain adaptation (DANN) applied to the extracted embeddings improves subject-independent performance and stability compared to a baseline without explicit adaptation.

Chapter 2

Materials and Methods

In this chapter, we present the experimental setup and the proposed computational framework for the binary classification of cognitive and emotional states.

2.1 The CLAS Dataset

The research is conducted utilizing the *Cognitive Load, Affect and Stress* (CLAS) dataset, which is a publicly available repository developed by Markova et al.[25] to support research on the automated assessment of certain states of mind and emotional conditions of a person. Furthermore, the CLAS dataset can also be used to support general studies research, such as attention assessment, cognitive load, emotion recognition, and stress detection. The CLAS dataset consists of time-synchronized recordings of physiological signals, such as Electrocardiography (ECG), Plethysmography (PPG), Electrodermal Activity (EDA), and accelerometer data. Additionally, also metadata were recorded, such as the stimuli tags and the responses given by the participants. Those signals were recorded during different tasks, purposely designed for eliciting specific cognitive and emotive responses and evaluating different aspects of the momentary cognitive load, the degree of attention and concentration, or the cognitive capacity of the person.

The test consists of two different tasks: *perceptive tasks* and *interactive tasks*. In the perceptive tasks, the subjects are exposed to a series of selected stimuli, with the aim of eliciting some specific emotions in the four quadrants of the arousal-valence space, as shown in figure 1.4. The images are taken from a specific dataset, the International Affective Picture System (IAPS), which is basically an online database of images, very diffused in the psychological research, that have been validated as consistently eliciting a specific emotional response in viewers. Instead, the video clips are taken from the DEAP database platform, and also these have been selected to represent the four quadrants of the arousal-valence space. During those tasks

the subjects are simply asked to watch video clips and images on a screen, without any active participation. For what concerns the interactive tasks, the focus is on the evaluation of the level of concentration, the cognitive capacity and certain personality traits, which are related to the ability of a person to quickly solve logical and mathematical problems under strict time constraint and psychological pressure. The interactive tasks consists on three different tests:

- Stroop test;
- Math test;
- Logic test.

The logical and mathematical problems have a low complexity and can be easily solved by persons with average Intelligence Quotient (IQ) scores and basic level of math skills when time is unrestricted. Here, during those tests, they allow a short time period for response (only a few seconds) to increase the cognitive load on the subjects. Markova et al. evaluated the degree of concentration indirectly based on the success rate during the Stoop test and the cognitive effort based on the success rates obtained in the Math test and Logic test. The neural stimuli at the beginning of each session is considered eliciting a low cognitive load. The plot of the success rates of each subjects finds out to be represented with four compact groups of scores.

2.1.1 Protocol description

Before the start of the start of the recording procedures all the volunteers compiled a questionnaire collecting general information of their current health status, sleeping habits, the intake of drugs, cigarettes, alcohol, caffeine or any other stimulator of the mind and the body. The anonymized questionnaires are collected in Figure 2.1. The population of this dataset signals collection is formed by 62 healthy volunteers. Among these are 17 women and 42 men. Most of the participants are in their twenties, with exception for one person in his thirties, one in her late forties, and one man who is 50 years old.

As described in Figure 2.2, the CLAS dataset is based on a very well-defined and unique protocol for all participants. The video clips amount to 16, organized in 4 groups of 4 videos each. Each block is supposed to elicit a specific emotion, one in each quadrant of the arousal-valence space as previously described. The duration of each video clip is about 60 seconds. In between different blocks a neutral video stimuli of the duration of 30 seconds is played with the aim to restore a relaxed condition (low arousal and neutral valence). Also the images are separated in 4 groups composed by four images each, and each image is displayed for 20 seconds. Again, after each block of images, a neutral image is displayed

age	20-27											
gender	male						female					
count	44						16					
leading hand	*right*			*left*			*right*			*left*		
	37			7			15			1		
	normal	*corrected*	*uncorrected*	*normal*	*corrected*	*uncorrected*	*normal*	*corrected*	*uncorrected*	*normal*	*corrected*	*uncorrected*
vision	25	11	1	5	1	1	9	6	0	1	0	0
	never	*rarely*	*often*	*never*	*rarely*	*often*	*never*	*rarely*	*often*	*never*	*rarely*	*often*
alcohol	4	28	5	1	3	3	1	13	1	0	1	0
coffee	4	12	21	2	1	4	2	2	11	0	0	1
black/green tea	11	25	1	2	5	0	4	8	3	0	1	0
cigarettes	17	6	14	2	1	4	5	1	9	0	0	1
medicine	11	25	1	1	5	1	1	13	1	0	1	0
drugs	32	5	0	5	2	0	15	0	0	1	0	0
	low	*medium*	*high*	*low*	*medium*	*high*	*low*	*medium*	*high*	*low*	*medium*	*high*
efficiency	0	19	18	0	3	4	1	9	5	0	1	0

Figure 2.1: Summary of participant characteristics. This information is directly extracted from the documentation present in the publicly CLAS folder.

for 30 seconds. For what concerns the cognitive stimuli, it consists of various multiple-choice assignments, followed by the display of the correct answer. The details of each test are listed below:

- The Math test is composed of a sequence of 24 math problems. Each assignment is presented for a very short time (4 seconds) and then the person has 2 seconds to select the correct answer. Afterwards the correct answer is shown for 1 second;

- The Stoop test consists of 30 assignments. The single problem is shown for 3 seconds, then the candidate has 2 seconds to select the correct answer and lastly the correct answer is displayed for one second;

- The Logic test consists of 20 assignments. The time for showing each problem was 10 seconds, the time for answering is 4 seconds and 1 second for displaying the correct answer.

In between the different interactive tasks, a neutral 30-second-length audio-visual stimuli is given, in order to relax the subject and restore his emotional state back to neutral. The sequence of the various tasks and the duration of each stimuli is detailed in Figure 2.2. The protocol used is identical for all the 62 participants. The tasks and timing are identical, they just inverted the order of the image stimuli and video clips stimuli occurred after participant number 12; participants from 1 to 12 follow the set-up number 1, whereas the rest of the participants follow the set-up number 2. Although, inside each different block of images and videos, the temporal order with which images and videos are displayed does not change

Figure 2.2: Block-based timeline of the CLAS experimental protocol for the two acquisition set-ups. Set-up 1 (participants 1–11) and Set-up 2 (participants 12–60) share the same tasks and stimulus blocks but differ in the ordering of the perceptive stimuli (image blocks vs video blocks). Each rectangle represents one protocol block and is labeled with its duration in seconds.

between the two set-ups. Before the beginning of the stimuli a 60s-baseline signal is recorded, during which the participant is asked to relax and an audio-video stimuli is displayed.

2.1.2 Sensors and signal acquisition

The physiological signals of the CLAS dataset have been recorded using wearable sensors, such as the Shimmer3 GSR+ Unit and the Shimmer3 ECG Unit. In particular the Shimmer3 GSR+ Unit measures the electrodermal resistance between two Ag/AgCl electrodes placed at the fingers of one hand. Additionally, this Unit also record the PPG signal from the ear lobe. Instead, the Shimmer3 ECG Unit is employed to record the ECG signal, using the Lead I configuration. For what concerns the three-axis accelerometer data, they were collected using the Shimer3 GSR Unit. All the physiological signals were acquired with a sampling rate of 256 Hz and a resolution of 16 bits per sample. Furthermore, the collected signals are synchronized, using a custom-build software.

2.1.3 Dataset Organization

In the CLAS dataset, which is publicly available, the records of 60 participants are available. The authors provided the signals in different formats, ranging from the uncut version to block-based subdivision and finally a subdivision by single tasks. For our purposes, the block-based subdivision of the recordings was selected and analyzed. In addition to physiological signals, the dataset provides a signal-quality annotation with three discrete levels (1–3). Specifically, level 1 denotes good-quality recordings, level 2 indicates noisy signals, and level 3 corresponds to poor-quality data. The most affected subjects are number 11, 13, 44, and 46. In detail:

- Subject 11 had the EDA signal labeled as level 3, except for the Math task (level 1) and the baseline recording (level 2).
- Subject 13 had the ECG signal labeled as level 3, except for the Math task and the baseline segment, which were labeled as level 2.
- Subject 44 had 24 task blocks labeled as level 3.
- Subject 46 had 20 task blocks labeled as level 3.

No PPG signal denotes a poor-quality data. Figure 2.1 reports the subjects for which at least one task has been annotated from the authors of the CLAS dataset as poor-quality. More specifically, for each signal, the table shows both the total number of task blocks labeled as quality level 3 (poor-quality level) and the corresponding number of blocks retained for this thesis. Despite this, it was

decided to keep these signals to avoid reducing too much the amount of available data and also to be able to train the network on signals of this type.

Participant	EDA poor-quality	EDA used	ECG poor-quality	ECG used
Subject 11	35	22	0	0
Subject 44	24	15	0	0
Subject 46	20	14	0	0
Subject 41	9	3	0	0
Subject 45	7	6	0	0
Subject 58	6	2	0	0
Subject 42	4	2	0	0
Subject 55	3	2	1	0
Subject 37	3	2	0	0
Subject 36	3	2	0	0
Subject 56	2	1	0	0
Subject 38	2	2	0	0
Subject 13	0	0	35	22
Subject 22	0	0	8	4

Table 2.1: Number of poor-quality task blocks (quality level 3) and corresponding blocks retained for this thesis after task block selection, separately for EDA and ECG signal.

For this thesis, we used the task blocks listed in Table 2.2. For each block, the corresponding duration is also reported.

The recordings corresponding to cognitive tasks, emotive tasks, and the initial baseline were extracted and annotated using a two-level labeling scheme. First, each segment was assigned a macro-label indicating the stressor category (0 for cognitive/interactive tasks, 1 for emotive/perceptive tasks, and 2 for baseline records). In addition, a task-specific label was stored (Math, Stroop, IQ/Logic, Video, or Photo) to enable task-wise analyzes and error breakdowns.

The classification problem addressed in this thesis is binary, distinguishing cognitive/interactive stressors from emotive/perceptive stressors. The baseline recordings are not used as an additional class in the classifier; instead, they are employed exclusively to perform a subject normalization, with the aim of reducing inter-subject variability.

For the image stimuli, the dataset provides four separate files per block (one per image). Since images within the same block are presented consecutively without interruptions, these four files were concatenated into a single continuous recording. As a result, four aggregated files were obtained, corresponding to the four image blocks in the protocol.

After organizing and labeling the recordings, the pipeline begins with the

preprocessing stage, where raw physiological signals are filtered and prepared for subsequent segmentation and embedding extraction.

Task category	Stimulus / task	Duration (s)
Baseline	Initial baseline recording	60
Interactive (cognitive)	Math test	168
	Stroop test	180
	IQ / Logic test	300
Perceptive (emotional)	Video blocks	240
	Image blocks	80

Table 2.2: Task blocks used in this thesis work and their durations expressed in seconds in the CLAS protocol (block-based segmentation).

2.2 MOMENT: a Foundation Model

MOMENT [21] is the first family of open-source large pre-trained time series models. This model is based on T5 encoder architecture. MOMENT has been trained on a large amount of time series data, the *Time-series Pile*, a massive and diverse collection of time-series data coming from different domains, ranging from healthcare to engineering to finance. This extensive quantity of datas allows the model to capture changes in intuitive time series characteristics such as trend, amplitude, frequencies, phases and auto-correlation information of time series. However, il cannot differentiate between vertically shifted time series as it normalizes each signal prior to modeling. It has been shown that MOMENT, even without dataset-specific fine-tuning, can learn distinct representations of the time series data in order to classify them into different classes. The core pre-training objective is a Masked Time-series Modeling (MTM) task, where a portion of the input patches is randomly masked, and the model is trained to reconstruct them. This process forces the network to learn robust structural and temporal dependencies within the data. To handle the high dimensionality and continuous nature of time-series data, the model employs a patching technique where the input series is divided into fixed-length, disjoint sub-sequences. These patches are projected into embeddings and processed by the Transformer layers.

2.2.1 MOMENT architecture

The MOMENT architecture is shown in Figure 2.3, adapted from the original MOMENT paper[21], to summarize the patching strategy and the masked modeling objective. MOMENT is pre-trained and released in three different sizes: Small, Base and Large. In our work the Base version of MOMENT is used, and below are the corresponding architecture of this determined model. The base version model uses a 12 layer Transform with hidden dimensions of size D = 768, 12 attention heads, and feed-forward networks of size 3072, resulting in approximately 125 million parameters. All weights are randomly initialized before pre-training. The model takes an input time series of length T = 512, breaking it into N = 64 disjoint patches of length P = 8. 30% of the patches are masked randomly during pre-training. The Adam optimizer with weight decay[26] with $\lambda = 0.05$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ is implemented. The gradient was clipped to 5.0, trained models using a batch size of 2048, and use cosine learning rate schedule with initial and final learning rates of $(1e^{-4})$ and $(1e^{-5})$, respectively. The authors used a gradient checkpointing [27] to improve training throughput and save memory, and trained the model in a mixed precision setting, using float-32 for numerically unstable operations, e.g. layer normalization, and bfloat-165, otherwise.

As mentioned above, MOMENT can be used in different modalities. For the aim of the work, MOMENT is employed as a Feature Extractor, with his weights frozen.

2.3 Proposed Methodology: Multimodal Foundation Model Framework

2.3.1 Overview of the framework

This study proposes a novel deep learning framework for physiological stress detection, with the aim of distinguishing stress stimuli, obtaining a supervised classification between the cognitive and the emotive tasks. The core of this work consists in the use of a Foundation Model, more specifically the MOMENT-1-base, to extract features from the signals and then using a Deep Multi-Layer Perceptron (MLP) network to train the model for classifying the tasks with a Late Fusion process of the three branches. The presented framework operates with the three physiological signals collected in the CLAS dataset individually and performs a Multimodal Late Fusion. An Early Fusion model is also tested for comparing it with the Late Fusion model. All the framework is developed on Colab, using Python. An overview of the different stages of the current framework is shown in the Figure 2.4 and the experimental setup combined with the training parameters is listed in Table 2.4.

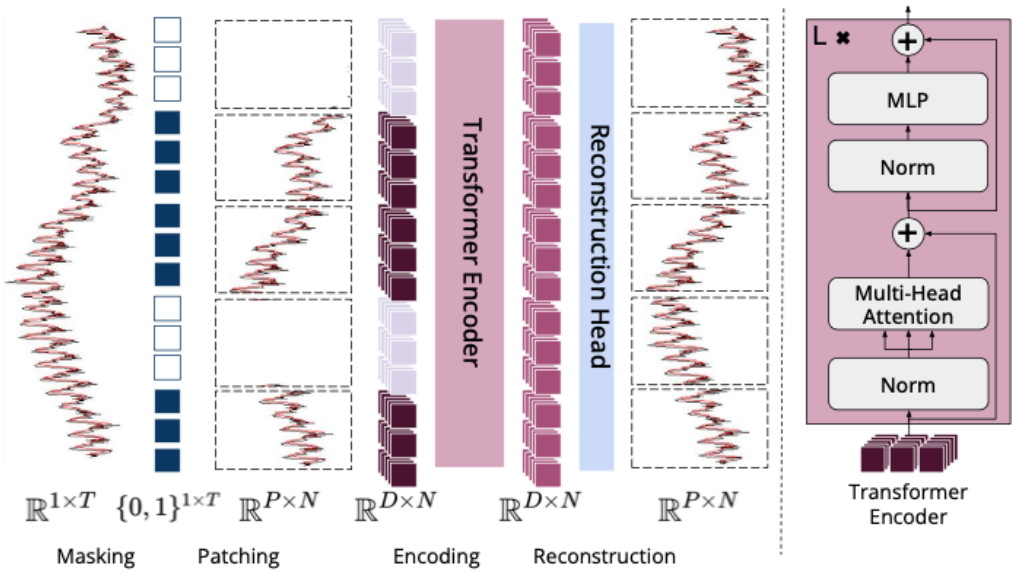


Figure 2.3: Overview of the MOMENT architecture. The input time series is divided into fixed-length patches, mapped to D-dimensional patch embeddings, and pretrained via masked time-series modeling by replacing a subset of patches with a special [MASK] embedding. Reproduced from Goswami et al.[21], Figure 3.

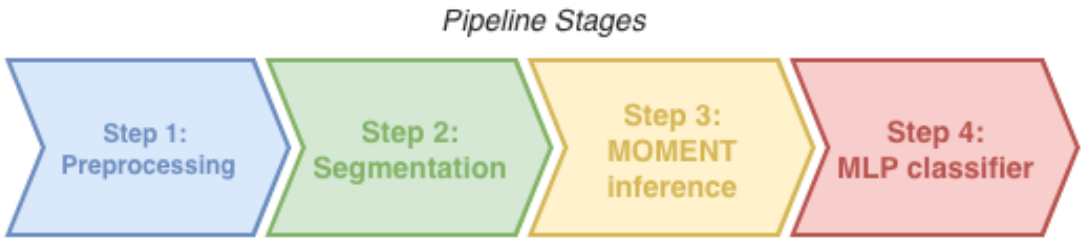


Figure 2.4: Overview of the main stages of the proposed framework. The framework consists in four main stages: (1) signal pre-processing, (2) segmentation into fixed length patches, (3) feature extraction with MOMENT inference, and (4) task classification using MLP head.

2.3.2 Pre-processing Stage

In the preprocessing stage, first the signals underwent a filtering process. The filters were designed with a double pass filtering (to avoid phase distortions on the signals), for removal of the drift, and noise. A Butterworth 2th order band-pass

filter is used through the *sosfiltfiltration* function, obtaining a 4th order filtering, for the EDA and ECG signal; the choice of the zero-phase filtering is motivated by the need of preserving the temporal morphology of the signals, especially for ECG signal where the morphology of the heartbeat is crucial for stress recognition. The EDA signal is filtered in a band-pass [0,05; 3] Hz: for this signal, it turns out, as described in many other studies, that the useful information component is typically under 3 Hz. For what conceirns ECG signal, the band-pass used is [0,5; 40] Hz, in accordance with best practice. Instead, the PPG signal is treated with the NeuroKit2 tools, a Python toolbox designed for neurophysiological signal processing, such as PPG signal.[28]. The 'elgendi' method was applied for signal cleaning[29], which consists of a band-pass 3th order Butterworth filter between 0,5 and 8 Hz. The Figure 2.5 shows the action of the filtering phase performed on the three signals for a random subject and a random task.

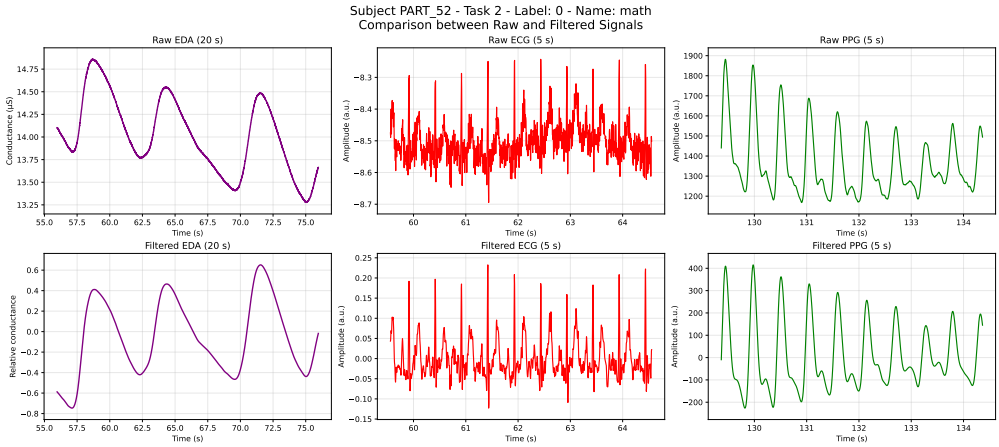


Figure 2.5: Comparison between raw and filtered physiological signals (EDA, ECG, PPG) for a representative subject (PART 52) during Math test (cognitive task). Each signal is shown over a fixed time window, with independent axis scaling to highlight morphological differences introduced by the preprocessing.

After the filtering phase, the signals underwent a downsampling phase. All signals are acquired at a sampling frequency of 256 Hz, but this phase is essential for the use of MOMENT for the aim of this work, because the input of MOMENT is a 512 point signal (independently of the sampling frequency of the input signal) and with downsampling a wider temporal window can enter in MOMENT. The original signals are sampled at a frequency of 256 Hz. The target frequencies for each signal were defined as follow:

- 8 Hz for EDA signal;
- 64 Hz for PPG signal;

- 128 Hz for ECG signal.

The reason for keeping the sampling frequency for the ECG signal higher is to preserve its morphology, particularly for the QRS complex. The difference of sampling frequencies used leads to a different number of points in a determined length of signal; the EDA signal is taken as the signal reference as it is the one with the lowest sampling frequency. Considering a segment of length 512 points, the corresponding duration (in seconds) for each signal is: 64 seconds for EDA segment, 8 seconds for PPG segment and 4 seconds for the ECG segment. The code automatically handles potential length mismatch in between signals before entering in the downsampling block to make sure the signals have the same length and are aligned.

The EDA signals are normalized just after the downsampling with a Z-score normalization: This strategy of normalizing the EDA signal inside the task turns out to be more efficient in terms of preserving the global trend of the EDA signal, which is a very low frequency signal and otherwise would be lost with a normalization inside the patches.

The signals are then segmented into patches with a fixed length of 512 points, required by the MOMENT foundation model. The data are organized in "macro-patches", with a fixed duration of 64 seconds. Due to the different sampling frequencies, EDA macro-patches are composed of 1 patch, while PPG macro-patches count 8 patches and ECG macro-patches count 16 patches. A detailed description of the segmentation of the three signals is shown in Figure 2.6.

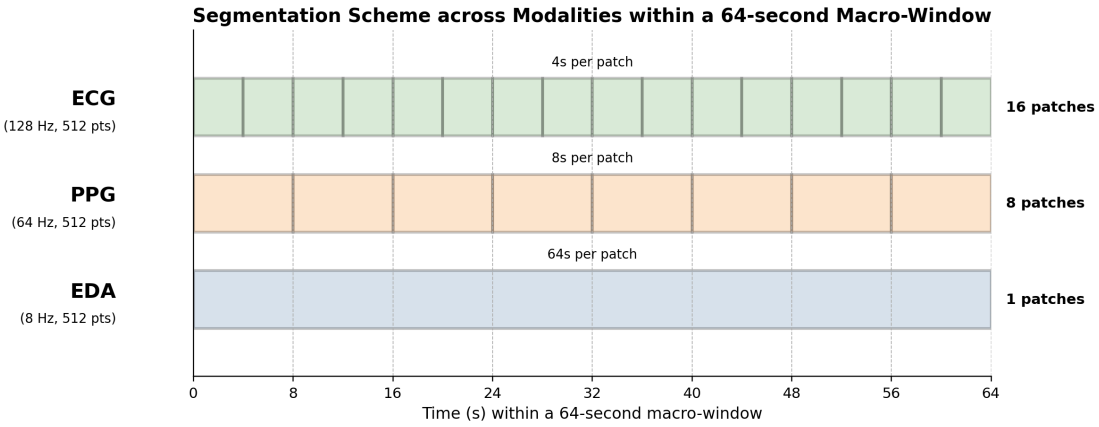


Figure 2.6: Scheme of the segmentation of a 64-second window length after downsampling for the three modalities. Each patch contains 512 samples. Due to the modality-specific sampling rates, a 64-second macro-window corresponds to 1 EDA patch (64 s), 8 PPG patches (8 s each), and 16 ECG patches (4 s each).

To mitigate the inherent data scarcity of the dataset and especially to robustly

train the deep learning neural network architecture, a data augmentation strategy based on a sliding window was implemented. This is a common and easy way to attempt data augmentation. Consequently, rather than partitioning the signals into non-overlapping segments, a stride is introduced, resulting in the creation of patches with an overlap with the previous patch. A different stride between interactive tasks and perceptive tasks is required in order to obtain more segments and keep the two classes balanced in terms of number of labels. In our framework, a 6 s stride and a 10 s stride were chosen for interactive and perceptive tasks, respectively. At the end of the segmentation process, a total of 6947 patches for the cognitive tasks and 6818 patches for the emotive tasks were created.

At this point of the pipeline, a local normalization of the ECG and PPG signals occurred. Again, a Z-score normalization was chosen, but this time it was applied inside each single patch; ECG and PPG are signals in which the key information lies in the morphology and the rhythm, less in the amplitude. It has been shown that parameters such as heart rate (HR), heart rate variability (HRV) or RR interval are highly discriminant for stress detection.

For what concerns the baseline signals, the duration of the baseline task is 60 seconds, which means that the EDA signal has less then 512 points (around 480 values). To ensure the good dimensionality for MOMENT application, a paddling strategy is applied, in modality 'reflect' that reflects the boundaries values of the signal, extending it at the correct dimension of 512. This lead to the creation of one single window for the EDA signal for each subject, for a total of 60 EDA baseline windows. Instead, the PPG and ECG baseline signal are processed the same way of the ones of the cognitive and perceptive tasks, with a stride of 256 points, causing an overlap of 50% and creating around 14 segments for PPG baseline signal and around 29 for ECG baseline signal. This leads to a total of 821 segments for the PPG signals and 1708 segments for the ECG signals across all subjects.

2.3.3 MOMENT inference

Afterwards, all the patches, from baseline, interactive and perceptive tasks enter in MOMENT. The foundation model is used in a frozen state, so its weights are not updated during back propagation. The model is set in the 'embedding' modality, allowing the model to learn representations from time-series input without any pre-training. This choice to use the foundation model in a freezed configuration is to significantly lower the computational cost of the framework. For each patch or signal segment of dimension 512, MOMENT gives back an embedding, which is a tensor of dimension 768 representing the features of the input time series. Embeddings representing interactive, perceptive, and baseline tasks are obtained. The baseline embeddings of ECG and PPG are processed and collapsed into one single embedding, by calculating the median value. The median is preferred over the

mean, as it is more robust to outliers. In this way, for each subject, three baseline embeddings are obtained, each corresponding to a biological signal baseline.

2.3.4 Attention Pooling and Baseline correction

For every 64 seconds of the PPG and ECG signal (corresponding to one macro-patch), 8 and 16 embeddings are obtained respectively. To resume this information into just one embedding, for each signal and each macro-patch, an Attention Pooling layer is employed. This layer assigns a different weight to different parts of the signal, depending on its relevance. Higher weights are assigned to more significant segments, while lower weights are assigned to noisy or less-informative segments.

The attention mechanism computes a scalar weight α_k for the k-th patch embedding $\mathbf{h}_k$ using a learnable compatibility function. The process is defined as follows:

First, the unnormalized attention score s_k is computed by projecting the embedding through a non-linear transformation:

$$s_k = \mathbf{w}^T \tanh(\mathbf{V}\mathbf{h}_k^T + \mathbf{b}) \tag{2.1}$$

where $\mathbf{V} \in \mathbb{R}^{L \times D}$ is a weight matrix that projects the input into a hidden attention space of size L, $\mathbf{b}$ is the bias vector, and $\mathbf{w} \in \mathbb{R}^L$ is a learnable context vector. The hyperbolic tangent (tanh) activation function is employed to handle non-linear relationships in the feature space.

Subsequently, the scores are normalized across the sequence using the Softmax function to obtain a probability distribution α_k, ensuring that the weights sum to unity:

$$\alpha_k = \frac{\exp(s_k)}{\sum_{j=1}^{N} \exp(s_j)} \tag{2.2}$$

Finally, the task-specific latent vector $\mathbf{z}$ is computed as the reliability-weighted sum of the input patches:

$$\mathbf{z} = \sum_{k=1}^{N} \alpha_k \mathbf{h}_k \tag{2.3}$$

To mitigate the impact of inter-subject variability, which represents a significant challenge for stress detection task, a normalization strategy is applied on the extracted embeddings. Since the signals, and so their embeddings are heavily biased by individual traits such as basal skin conductivity or resting heart rate variability, it is very important to center the data around the neutral state of the subject. For each subject, a Z-score normalization of the embeddings is performed,

using the statistical moments (mean $\boldsymbol{\mu}_B$ and standard deviation $\boldsymbol{\sigma}_B$) of his own baseline embedding, according to the following equation:

$$\text{Emb}^{(j)} = \frac{\text{Emb}^{(j)}_{task} - \mu^{(j)}_{base}}{\sigma^{(j)}_{base} + \epsilon}, \quad \forall j \in \{1, \ldots, D\} \tag{2.4}$$

The ϵ value prevents the division by zero to happen, in the event that $\boldsymbol{\sigma_{baseline}}$ equals zero. This standardization procedure mitigates the impact of inter-subject variability, rendering the feature space largely independent of individual biometric traits. Consequently, the classifier is conditioned to learn from the physiological deviation induced by the stressor (reactivity), rather than being biased by the subject's idiosyncratic resting baseline.

The full explained pipeline is resumed in Figure 2.7.

2.3.5 Task Classifier Architecture

The core of the proposed framework relies on a Deep MultiLayer Perceptron (MLP) architecture acting as the Task Classifier. Its primary objective is to learn a robust, non-linear decision boundary in the physiological latent space to differentiate between interactive (cognitive) and perceptive (emotional) stress.

The input to the Task Classifier is the rich latent vector of dimension 768 extracted by the frozen MOMENT Foundation Model. To handle potential class imbalances and force the network to learn subtle discriminative patterns, a Focal Loss function (with parameter $\gamma = 2$) was adopted for optimization. Unlike standard Cross-Entropy, Focal Loss dynamically scales based on prediction confidence, reducing the relative contribution of easily classifiable samples and forcing the model to focus on "hard" examples:

$$L_{task} = -(1 - p_t)^{\gamma} \log(p_t) \tag{2.5}$$

where p_t is the model's estimated probability for the true class. This classifier acts as the primary Baseline architecture (Source-Only) for evaluating the intrinsic generalization capabilities of the extracted embeddings.

2.3.6 Adversarial Evaluation Module (DANN)

To rigorously investigate whether the MOMENT Foundation Model inherently extracts subject-independent features, or if it requires explicit domain alignment, an experimental Adversarial Evaluation Module was introduced.

This module transforms the baseline architecture into a Domain Adversarial Neural Network (DANN) by adding a secondary branch connected via a Gradient Reversal Layer (GRL). The goal of this branch is to predict the subject's identity

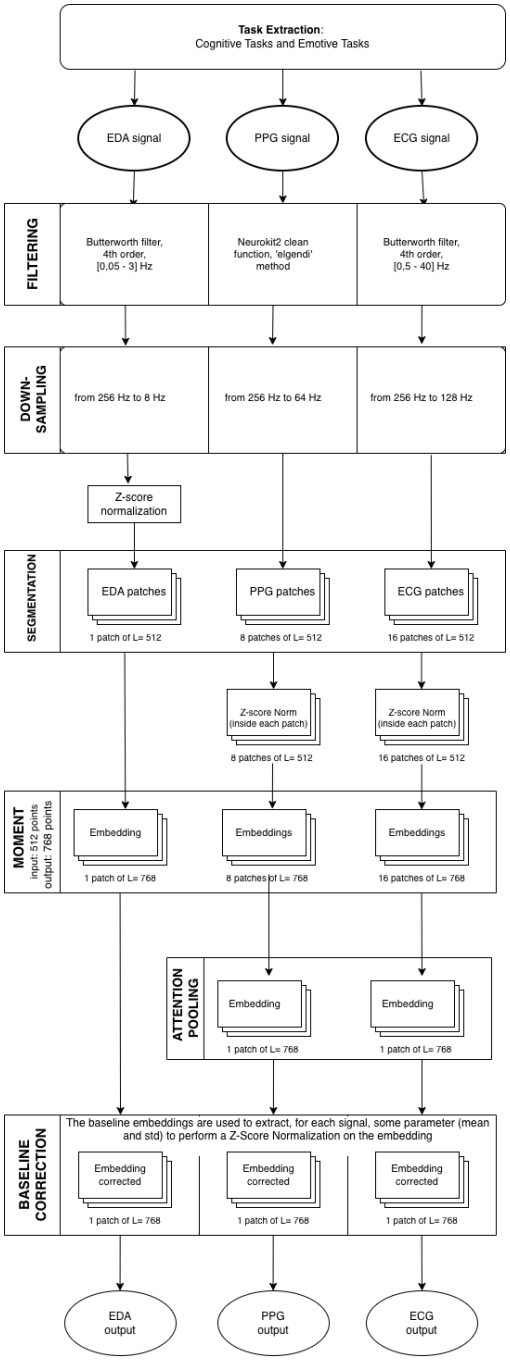


Figure 2.7: Pipeline of the presented framework.

(Domain Discriminator). During backpropagation, the GRL reverses the sign of the gradient flowing from the discriminator back to the feature extractor. This minimax game penalizes the feature extractor if the embeddings contain biometric information, forcing the latent space to become subject-agnostic.

Because different physiological signals possess varying degrees of biometric signatures, a **signal-adaptive strategy** was designed to modulate the discriminator's capacity:

- **Strong Biometric Signals (ECG and PPG):** These signals act as strong biometric fingerprints, making it extremely easy for the discriminator to recognize the subject. To prevent the discriminator from immediately overpowering the feature extractor, its capacity was strictly limited (only one hidden layer of 64 neurons) and the dropout was heavily maximized ($p \approx 0.65$).
- **Weak Biometric Signals (EDA):** Conversely, the EDA signal inherently lacks a strong biometric signature. To force the network to learn subtle subject-specific features and ensure a fair adversarial training, the discriminator was enhanced with a deeper architecture (two layers of 256 and 128 neurons) and the dropout was lowered ($p = 0.2$).

Comparing the performance of the Baseline framework against this DANN-enhanced version allows for a scientific assessment of the intrinsic robustness of Foundation Models against the Domain Shift problem.

2.3.7 Multimodal Late Fusion Strategy

To maximize the diagnostic power of the physiological signals, the framework employs a Multimodal Late Fusion strategy. Instead of concatenating the raw signals or early embeddings, the three modalities (EDA, PPG, ECG) are processed by independent classifier branches. This allows each sub-network to learn the optimal feature representations specific to its signal domain, without cross-modal interference.

Once the three independent models ($M_{eda}, M_{ppg}, M_{ecg}$) are trained, their outputs are aggregated during the inference phase. Each model m produces a probability distribution vector $P_m \in \mathbb{R}^2$ via the Softmax function:

$$P_m(x) = \text{Softmax}(z_m) = \frac{e^{z_m}}{\sum_j e^{z_j}} \tag{2.6}$$

where z_m are the logits produced by the classifier for modality m. The final fused probability vector P_{final} is computed as a linear combination of the individual probabilities:

$$P_{final} = w_{eda} \cdot P_{eda} + w_{ppg} \cdot P_{ppg} + w_{ecg} \cdot P_{ecg} \tag{2.7}$$

Component / Hyperparameter	EDA branch	PPG branch	ECG branch
Task classifier (MLP) settings			
Hidden layers (units)	$256 \to 128 \to 2$	$256 \to 128 \to 2$	$256 \to 128 \to 2$
Classifier dropout	$0.6 \to 0.5$	$0.6 \to 0.3$	$0.7 \to 0.5$
Focal loss γ	2	2	2
Learning rate (start)	1×10^{-4}	1×10^{-5}	1×10^{-5}
Weight decay	5×10^{-2}	5×10^{-2}	5×10^{-2}
Max epochs	70	70	70
Early stopping patience	12	15	15
Early stopping $\Delta_{\min}$	1×10^{-4}	1×10^{-4}	1×10^{-4}
Batch Size	64	64	128
Domain discriminator (DANN) settings			
GRL maximum α	0.3	0.8	1.0
Discriminator hidden units	$256 \to 128$	64	64
Discriminator dropout	$0.2 \to 0.2$	0.60	0.65

Table 2.3: DANN implementation details and hyperparameters for each modality-specific branch. For each signal, the table reports the MLP task-classifier configuration and the domain discriminator settings used in the adversarial training setup.

The scalar weights w_m (where $\sum w_m = 1$) determine the contribution of each modality to the final decision. To maximize robustness, these weights were not set arbitrarily but were derived empirically from the Validation Accuracy of each single-modality model. Let Acc_m be the accuracy of model m on the validation set. The weight for modality m is calculated as:

$$w_m = \frac{Acc_m}{\sum_{k \in \{eda, ppg, ecg\}} Acc_k} \tag{2.8}$$

This data-driven weighting scheme ensures that the fusion mechanism automatically trusts the most reliable signal for the given subject. For instance, if the ECG signal provides a clear separation of stress states (high validation accuracy) while the EDA signal is noisy or inconclusive, the fusion logic assigns a dominant weight to the ECG prediction, thereby filtering out the uncertainty introduced by the weaker modality. Finally, the predicted class $\hat{y}$ is obtained by selecting the index with the maximum probability in the fused vector:

$$\hat{y} = \underset{c \in \{0,1\}}{\operatorname{argmax}}(P_{final}[c]) \tag{2.9}$$

By aggregating modality-specific probability estimates with validation-driven

weights, the framework reduces single-sensor failure modes and yields a more robust stressor discrimination model than any individual modality.

2.3.8 Training Protocol and Performance Monitoring

To rigorously evaluate the generalizability of the framework across unseen subjects, a Subject-Independent Cross-Validation protocol was adopted. Unlike standard Leave-One-Subject-Out (LOSO) approaches, this study used a strict Train-Validation-Test split to prevent data leakage and ensure unbiased performance estimation. For each fold of the cross-validation (targeting a specific subject S_{test}):

- **Test Set:** The target subject S_{test} is completely isolated and accessed solely for the final inference, to ensure that the network never encounters their data during training.

- **Validation Set:** A subset of 3 distinct subjects (different from S_{test}) is left out of the training pool. This set is used exclusively to monitor convergence, tune hyperparameters, and trigger the *Early Stopping* mechanism.

- **Training Set:** The remaining $N - 4$ subjects constitute the training corpus used to update the model weights through backpropagation.

The cross-validation protocol, which is summarized in Figure 2.8, is repeated for every subject, resulting in a total of 60 folds per signal-branch.

During training, several metrics were tracked to monitor the gradient descent and network behavior. For the baseline model, only the L_{task} was minimized. For the adversarial evaluation setup, the losses were tracked separately:

- **Task Loss ($Loss_{task}$):** Measures the error made by the network during the stress classification task. A progressive decrease indicates that the network is learning class-separability.

- **Domain Loss ($Loss_{domain}$):** Measures the error in subject recognition. In the adversarial architecture, this parameter is expected to decrease initially. Subsequently, due to the GRL, the feature extractor actively tries to increase this loss (or stabilize it at high values), indicating that the features have become "confusing" regarding the subject's identity.

- **Total Loss ($Loss_{total}$):** The weighted sum of the two objectives, defined as $L_{total} = Loss_{task} + \lambda * Loss_{domain}$, where λ represents the adaptation factor.

To avoid overfitting, performance was evaluated at each epoch on the Validation Set. The *Validation Loss* was utilized to activate the Early Stopping function, which halts training if no improvements are registered for 15 consecutive epochs.

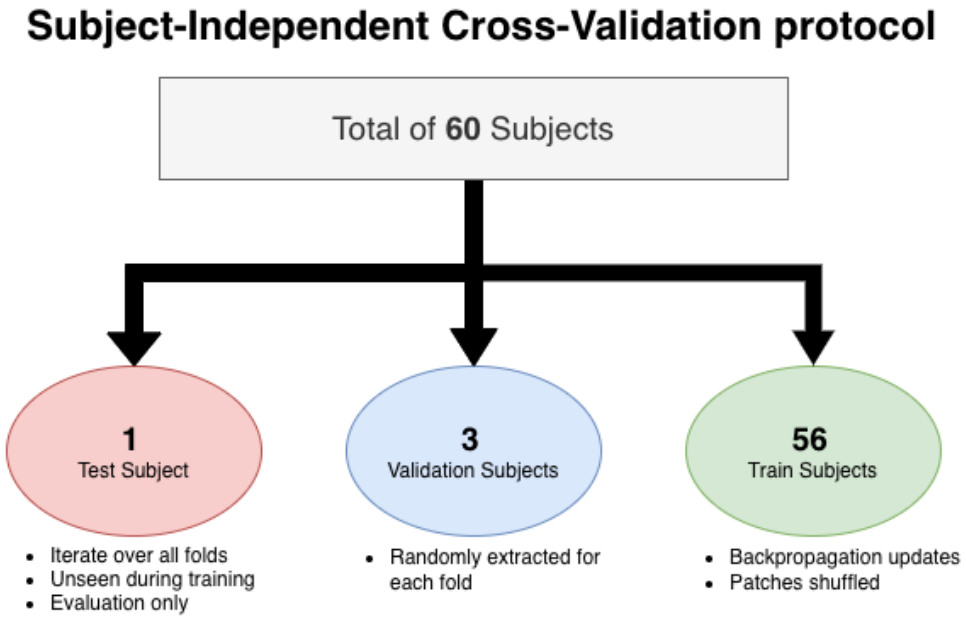


Figure 2.8: Train/validation/test partitioning scheme for the subject-independent cross-validation (60 folds): 56 training subjects, 3 validation subjects, and 1 held-out test subject per fold.

Furthermore, *Domain Accuracy* was monitored: a value close to the chance level (i.e., $1/N_{subjects}$) indicates successful adversarial alignment, meaning the discriminator is unable to distinguish between subjects.

2.3.9 Evaluation Metrics

To provide a comprehensive quantitative assessment of the proposed framework, classification performance was evaluated using standard statistical metrics:

Accuracy: The ratio of correct predictions to the total number of samples.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.10}$$

F1-Score: The harmonic mean of Precision and Recall. It provides a balanced view of the predictions, which is especially useful if class prediction are uneven.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2 * TP}{2 * TP + FP + FN} \tag{2.11}$$

The final results are presented as the mean across all test folds alongside the standard deviation, offering crucial insights into the stability and inter-subject robustness of the model. Finally, a Global Confusion Matrix was constructed by aggregating the test matrices of each fold. This metric reveals the cumulative distribution of True Positives, False Positives, and False Negatives, allowing for a highly detailed analysis of the framework's error patterns across the entire dataset.

Component	Setting
Evaluation protocol	Subject-independent LOSO cross-validation, 60 folds (one held-out subject per fold)
Per-fold split	Train: 56 subjects; Validation: 3 subjects; Test: 1 subject (fully unseen)
Input patch length (MOMENT)	512 samples per patch
Reference macro-window	64 s (EDA used as temporal reference)
Downsampling (from 256 Hz)	EDA: 8 Hz; PPG: 64 Hz; ECG: 128 Hz
Patch duration	EDA: 64 s; PPG: 8 s; ECG: 4 s
Patches per 64 s macro-window	EDA: 1; PPG: 8; ECG: 16
Segmentation (augmentation)	Sliding window with class-dependent stride
Stride (interactive tasks)	6 s
Stride (perceptive tasks)	10 s
Baseline handling	EDA baseline padded to 512 samples (reflect); PPG/ECG baseline segmented with 50% overlap
Baseline embedding aggregation	Median across baseline embeddings (PPG/ECG)
Embedding normalization	Subject-wise Z-score using baseline statistics (μ_{base}, σ_{base})
Representation model	MOMENT-1-base (frozen, embedding mode), output dimension $D = 768$
Classifier head	Multi-layer perceptron (MLP) for binary stressor discrimination
Loss function	Focal Loss ($\gamma = 2$)
Batch size	64
Max epochs	70
Early stopping	Patience = 20 epochs; minimum improvement $\Delta_{min} = 10^{-4}$ (validation loss)
Optimization	Adam (as implemented in PyTorch)

Table 2.4: Summary of the experimental setup and training hyperparameters used in this thesis for the Baseline model.

Chapter 3

Experimental Results

3.1 Evaluation Protocol and Experimental Setup

This chapter reports the experimental results obtained with a strict subject-independent evaluation protocol, projected to evaluate the generalization of the developed framework to unseen subjects. Specifically, a Leave-One-Subject-Out style cross-validation has been implemented, resulting in 60 folds for each branch. Unlike standard LOSO procedure, each fold was structured with an explicit Train-Validation-Test set separation, in order to prevent data leakage and enable unbiased model selection. For each folder, a single subject S_{test} was held out as the Test set and used esclusively for the final inference. From the remaining subjects, three subjects were randomly selected to form the Validation set and used only for monitoring the convergence of the learning of the model, hyperparameter selection, and early stopping. The random seed was set to the fold index to ensure reproducibility. The Training set consisted of the remaining $N - 4$ subjects, used to modify the classifier parameter through backpropagation and train the model. In Figure 2.8 is presented a visual representation of the test, validation, and train sets. The training lasted a maximum of 70 epochs and a batch size of 64 was employed. To avoid overfitting and reduce resource expenditure, an early stopping strategy was employed using the validation loss parameter as the stopping criterion. In detail, the training is interrupted if no improvements have been made by at least $\Delta_{\text{min}} = 10^{-4}$ for 20 consecutive epochs (patience = 20). This protocol is applied across all the different modalities explored during the experiments. This study used Focal Loss ($\gamma = 2$) rather than Cross Entropy. Focal Loss down-weights easy samples and emphasizes harder examples, which can be beneficial in physiological inference tasks where inter-subject variability and signal noise may create ambiguous segments. Performances were evaluated using Accuracy and F1-Score, computed on the Test subject for all folds. The results of the model are reported as the mean and

the standard deviation of the metrics across all folds, providing insight into both average performance and fold-to-fold variability. After the segmentation of both interactive and perceptive task signals, the number of segments associated with cognitive and emotive labels was, respectively, 6947 and 6818, creating a balanced dataset to train deep learning models.

3.2 Single-Modality Baselines Performances

In this section, the performance of the three different single-branch models using the Baseline model is presented. Each branch is evaluated individually, following the subject-independent protocol described in subsection 2.3.8. For each modality results are reported as mean ± standard deviation. The accuracy and the F1-score of the different modalities are resumed in Table 3.1:

Model branch	**Accuracy (mean ± std)**	**F1-score (mean ± std)**
EDA branch	0.6086 ± 0.1240	0.5660 ± 0.1537
PPG branch	0.6918 ± 0.1328	0.6405 ± 0.1808
ECG branch	0.7057 ± 0.1684	0.6506 ± 0.2229

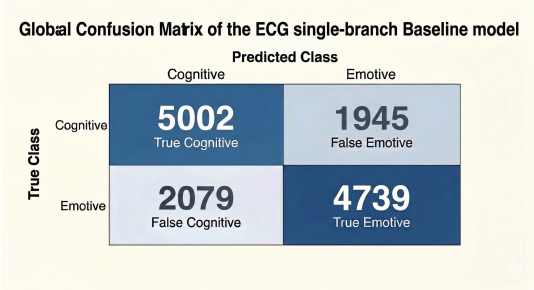
Table 3.1: Single-modality baselines performance across 60 subject-independent folds. Results are reported as mean ± standard deviation.

The ECG and PPG branches reported similar performances, with the PPG modality having slightly lower mean values in both metrics but more moderate standard deviations, compared to the ECG branch. The EDA signal turned out to be the weak signal in the stress discrimination classification task. In addition to the metrics, in figure 3.1 is reported the global Confusion Matrix (CM) of the three signal branches. The global CM is obtained by aggregating the confusion matrix across all 60 subject-independent cross-validation folds.

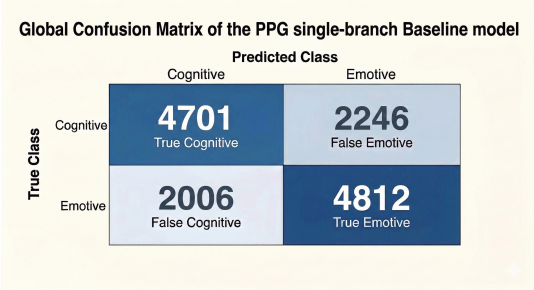
3.3 Multimodal Late Fusion strategy

The Multimodal Late Fusion strategy has turned out to be the one reaching the best performance, in terms of both accuracy and f1-score metrics. In detail, an accuracy of 0.7649 ± 0.1234 and a F1-score of 0.7639 ± 0.1244 were achieved. This performance is obtained with the following weights of the three single branches:

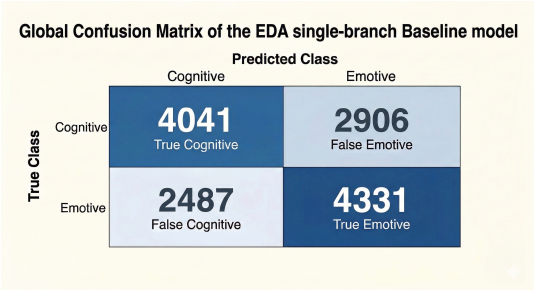
- $w_{\text{eda}} = 0.15$
- $w_{\text{ppg}} = 0.35$
- $w_{\text{ecg}} = 0.5$



(a) Global confusion matrix (ECG-only branch).



(b) Global confusion matrix (PPG-only branch).



(c) (Global confusion matrix (EDA-only branch).

Figure 3.1: Global confusion matrices for the single-branch baseline models (ECG, PPG, EDA), obtained by aggregating the per-fold confusion matrices across the 60 subject-independent test folds. Rows denote true labels and columns predicted labels for the two stressor classes (interactive/cognitive vs perceptive/emotional). This visualization highlights modality-specific error patterns and class confusions under strict cross-subject evaluation. Class 0 corresponds to interactive (cognitive-load) tasks and Class 1 to perceptive (emotional-arousal) stimuli.

The ECG branch is the channel with the highest weight, followed by the PPG signal and the EDA signal.

3.3.1 Late Fusion vs. Early Fusion

This section compares the two strategies of fusion explored: the early fusion and the late fusion. In the early fusion the embeddings of the three branches are concatenated before entering in the MLP classifier, while in the Late Fusion strategy three different modality-specific classifiers are trained independently and their output probabilities are merged at inference time. The final prediction is computed as a weighted sum of these these probabilities. The results obtained are summarized in Table 3.2.

In figure 3.2 are resumed the F1-score metrics across all modalities. The Multimodal Late Fusion improvements over the best single branch (that is the ECG-only branch) is statistically significant ($p< 0.05$). In addiction to average performance, the Multimodal Late Fusion has proven to affect stability by reducing the variability.

Fusion Strategy	**Accuracy**	**F1-Score**
Early Fusion	0.7266 ± 0.1548	0.6774 ± 0.2061
Late Fusion	0.7649 ± 0.1234	0.7639 ± 0.1244

Table 3.2: Comparison of Early Fusion and Weighted Late Fusion strategies in the Baseline architecture under subject-independent cross-validation. Results are reported as mean ± standard deviation of Accuracy and macro F1-score computed over the 60 test folds (one held-out subject per fold).

3.4 Domain Adversarial Training (DANN) as Robustness Probe

The introduction of Domain Adversarial Training (DANN) within the architecture was not conceived solely to maximize classification metrics, but to act as a robustness probe. Physiological signals, and in particular those of a cardiovascular nature such as ECG or PPG, carry with them a biometric "fingerprint". The goal of the DANN is to force the model to unlearn this subjective identity, focusing on the physiological variations caused by stress. The primary purpose of this implementation is to investigate whether the representations extracted by the frozen MOMENT foundation model are already sufficient robust to subject shift or whether an explicit domain-alignment method is required. This section analyzes the

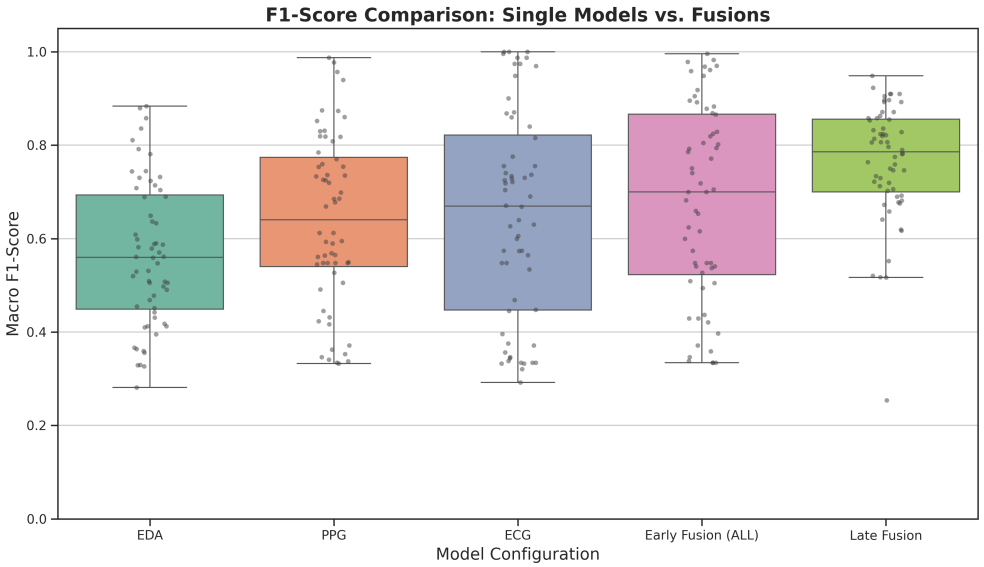


Figure 3.2: Boxplot distribution of F1-score metrics across all individual modalities (ECG, PPG, EDA), the Early Fusion Strategy, and the Multimodal Late Fusion strategy for the Baseline architecture. The overlapping points (stripplot) represent the performance obtained on individual subjects during the Subject-Independent (LOSO) Cross-Validation.

impact of such adversarial implementation on performance and learning dynamics and proposes a direct and in-depth comparison with the Baseline architecture.

In table 3.3 are shown the performance achieved by the DANN architecture across all different modalities. Once again, the best performance are obtained with the Late Fusion modality that outperformed all other modalities, reaching an accuracy of 0.7716 ± 0.1247 and a F1-score of 0.7708 ± 0.1254. Also this architecture confirmed the advantage of performing a Late Fusion instead of an Early Fusion strategy. Those metrics are achieved with the following weights of the three biological signals:

- $w_{\text{eda}} = 0.25$
- $w_{\text{ppg}} = 0.25$
- $w_{\text{ecg}} = 0.5$

Once again the ECG signal is the one with the highest weight. This time, EDA and PPG split the remaining half equally and each get the same weight.

Model branch	Accuracy (mean ± std)	F1-score (mean ± std)
EDA branch	0.6005 ± 0.1182	0.5549 ± 0.1484
PPG branch	0.6744 ± 0.1158	0.6430 ± 0.1494
ECG branch	0.6810 ± 0.1534	0.6404 ± 0.1935
Early fusion	0.7066 ± 0.1425	0.6600 ± 0.1902
Late fusion	0.7716 ± 0.1247	0.7708 ± 0.1254

Table 3.3: DANN architecture performance across different modalities. Results are reported as mean ± standard deviation.

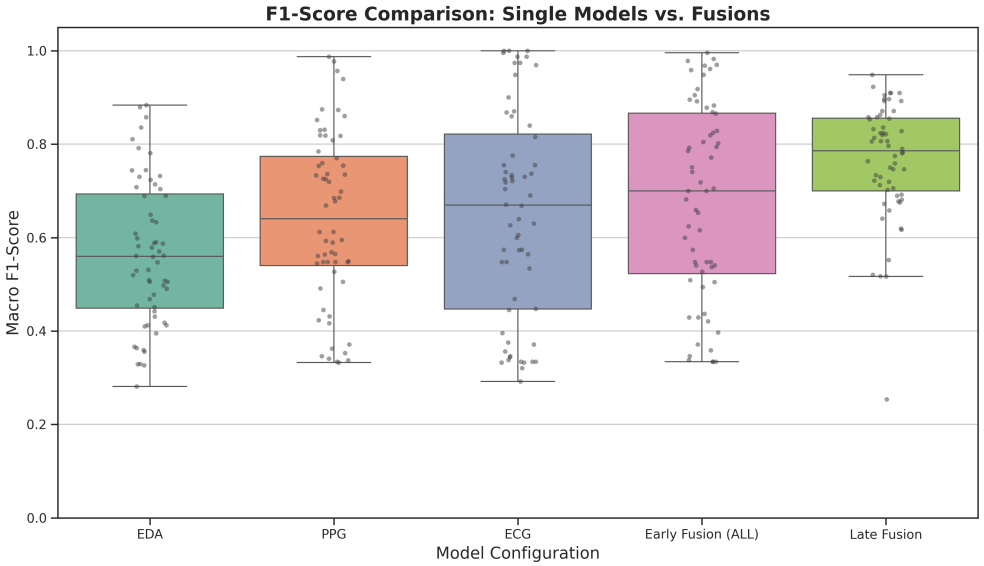


Figure 3.3: Boxplot distribution of F1-score metrics across all individual modalities (ECG, PPG, EDA), the Early Fusion Strategy, and the Multimodal Late Fusion strategy for the DANN architecture. The overlapping points (stripplot) represent the performance obtained on individual subjects during the Subject-Independent (LOSO) Cross-Validation.

3.4.1 Baseline vs DANN: Performance Comparison

This section shows a direct comparison between the Baseline architecture (non-adapted and trained uniquely by minimizing the Task Loss) against the DANN architecture. The purpose is to access whether explicit domain-adversarial implementation improves subjiect-independent stressor discrimination, compared to the baseline architecture.

As reported in Table 3.1 and Table 3.3, the DANN architecture yields a slight decrease of accuracy and f1-score in the single-modality performance. The only

Model branch	Δ Accuracy	Δ F1-score
EDA branch	+0.0081	+0.0111
PPG branch	+0.0174	−0.0025
ECG branch	+0.0247	+0.0102
Early Fusion	+0.0200	+0.0174
Late Fusion	−0.0067	−0.0069

Table 3.4: Performance differences between Baseline and DANN across branches. Δ is computed as (Baseline − DANN).

exception is for the PPG channel, where the DANN yields a f1-score of 0.6430 while the Baseline achieves 0.6405, indicating a small increase in performance. Overall, at the single-modality level, the implementation of the DANN does not provide consistent gains over the mean in accuracy and f1-score metrics. However, for what conceirns the standard deviation, a decrease is observed in all metrics for the three single-modality branches.

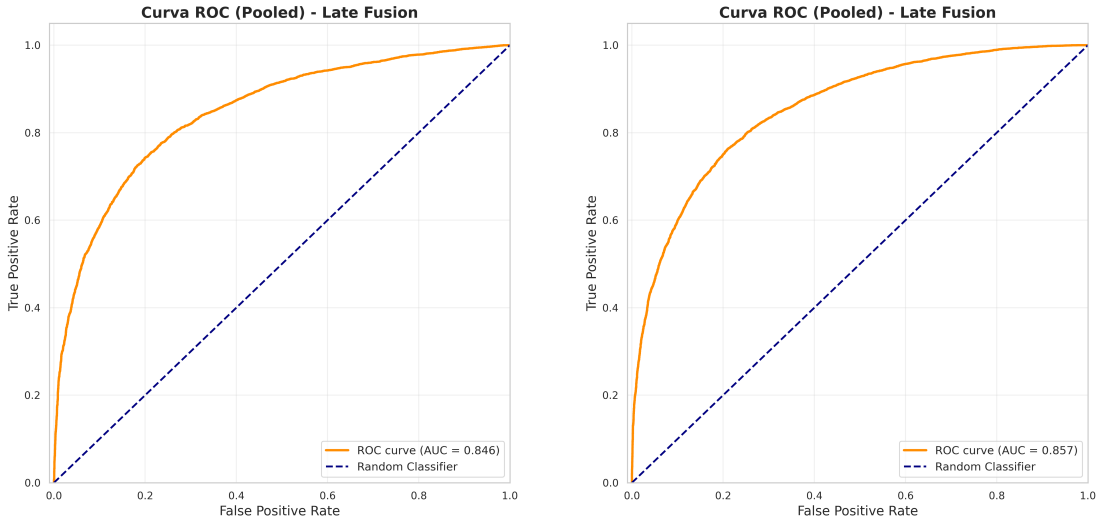
A direct comparison of the metrics between the two architectures is proposed in Table 3.4.

For the Late Fusion strategy, both architectures achieved similar performances in accuracy and f1-score metrics. The DANN yields a very slight increase in both metrics compared to the baseline ($\Delta Accuracy = +0.0067$, $\Delta F1-score = +0.0069$). Although, the non-parametric statistical test (Wilcoxon signed-rank test) did not show a significant difference in accuracy and F1-score metrics, with $p > 0.05$. The weight distribution for the Late Fusion showed, in both architectures, a dominant reliance on ECG signal with $w_{\text{ecg}} = 0.5$, followed by PPG signal and EDA signal (even if in the DANN architecture PPG and EDA have equal weight).

Although the increase in the overall average is marginal, the contribution of DANN is decisive in raising the lower limit of the predictive distribution, improving the worst case by more than 22 percentage points and lifting the entire lower tail ($5\hat{\text{th}}$ percentile), as shown in Table 3.5.

F1-Score	Baseline	DANN	Difference (Δ)
Mean (all folds)	0.7639	0.7708	+0.0069
Worst Fold	0.2540	0.4797	**+0.2257**
5° Percentile (Low Taila)	0.5206	0.5453	+0.0247
25° Percentile (1° Q.)	0.6998	0.6966	-0.0033
Critical Folds ($F1 < 0.50$)	1	1	-

Table 3.5: Robustness and stability Analysis evaluated on Late Fusion Strategy.

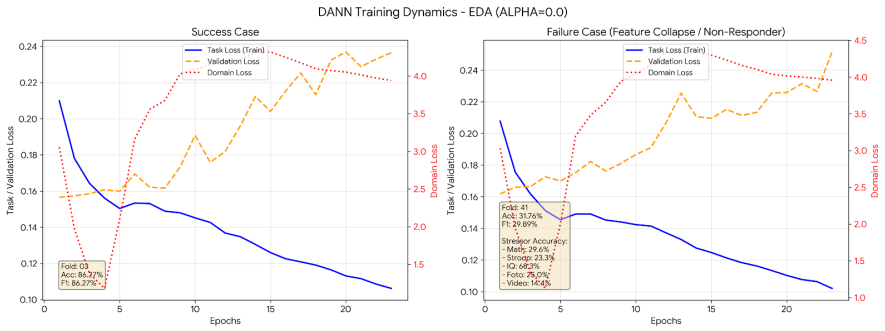


(a) Late Fusion: Baseline **(b)** Late Fusion: DANN

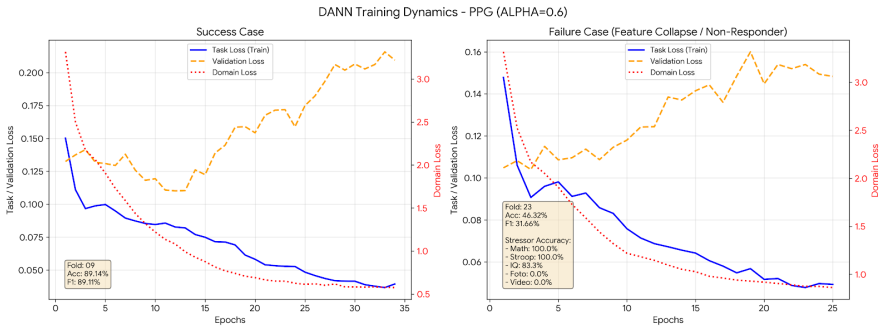
Figure 3.4: Comparison of ROC curves between the Baseline and the DANN Late Fusion frameworks.

3.4.2 Hyperparameter Sensitivity and Training Dynamics

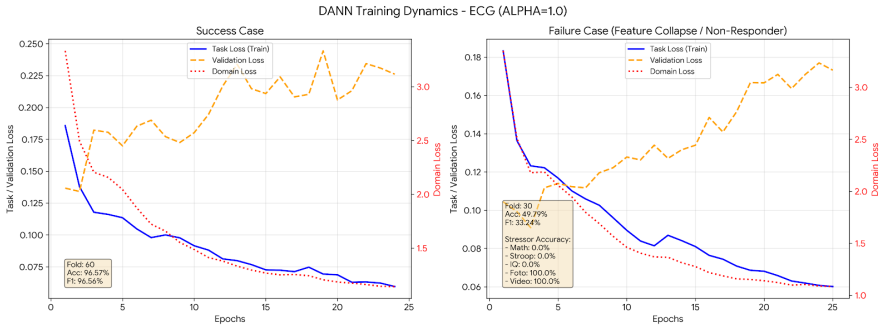
In this section, an analysis of the training logs has been performed to highlight how the three physiological signals react to adversarial training. To manage the different biological nature of these signals, the maximum limit of the GRL (parameter α) was dynamically scaled: $\alpha = 1.0$ for ECG (strongly biometric), $\alpha = 0.8$ for PPG (mixed), and $\alpha = 0.3$ for EDA (weak biometric). This parameter scales the strength of the GRL that forces the feature space to become subject-invariant while the domain discriminator aims to correctly predict subject identity.



(a) DANN training dynamics for the EDA branch (Task Loss, Domain Loss, and Validation Loss over epochs)



(b) (DANN training dynamics for the PPG branch (Task Loss, Domain Loss, and Validation Loss over epochs)



(c) DANN training dynamics for the ECG branch (Task Loss, Domain Loss, and Validation Loss over epochs)

Figure 3.5: Training dynamics of the Domain-Adversarial Neural Network (DANN) across physiological modalities. Each subplot reports Task Loss, Domain Loss, and Validation Loss over epochs. A decreasing task loss combined with domain accuracy approaching chance level indicates successful adversarial alignment and reduced subject identification in the latent representation.

In figure 3.5 are reported the trend of the parameters related to the single-modality for the three signals (EDA, PPG and ECG), illustrating both a successful fold and a failure/non-responsive fold for each branch. The success refers to a fold with accuracy higher than 80%, while in the failure case accuracy metric could not reach 40%. Each subplot shows the evolution of training task loss, validation loss, and domain loss over epochs. For ECG ($\alpha = 1.0$), the success case exhibits a monotonic decrease of task loss, while the domain loss decreases rapidly, indicating that the domain discriminator learns effectively under the strong biometric signature of ECG, and stabilizing after a few epochs. In the corresponding failure case, the task loss still decreases in the training set, but the validation loss increases early and continues to rise, resulting in poor fold-level test performance. This divergence between training and validation behavior suggests that, for certain held-out subjects, adversarial training can converge to representations that are not beneficial for task generalization. For PPG ($\alpha = 0.8$), the success case shows a stable convergence of the task objective: the training task loss decreases smoothly and the validation loss remains relatively contained in the first epochs, while the domain loss decreases steadily, suggesting that the discriminator is able to capture subject-related cues from the PPG embeddings. Compared to ECG, the decrease in domain loss is slightly slower and the curves tend to be more oscillatory, which is consistent with the mixed nature of PPG: it contains cardiovascular dynamics informative for stressor discrimination, but is also more sensitive to acquisition variability and residual noise. In the corresponding failure case, the task loss still decreases, but the validation loss exhibits a clear upward trend after only a few epochs, anticipating a low-performing fold. Notably, in this regime the domain loss still decreases, indicating that the discriminator continues to learn subject identity even when the task generalization collapses.

For EDA ($\alpha = 0.3$), the overall dynamics differ from the cardiovascular modalities. In the success case, the task loss gradually decreases, whereas the domain loss remains comparatively higher and less stable across epochs, suggesting a reduced ability of the discriminator to identify subjects from EDA-derived representations. This behavior is coherent with the weaker biometric signature of electrodermal activity: while EDA is highly sensitive to sympathetic arousal, it typically provides less subject-specific "fingerprint" information than ECG or PPG. In the failure case, a similar pattern is observed which the training objective improves but the validation loss increases, leading to poor test performance. Importantly, even with a relatively low adversarial strength, some folds still fall into a non-responsive regime, highlighting that fold-level generalization may be affected by subject-specific reactivity patterns and signal quality.

Across modalities, these logs highlight a common trend: the validation loss often reaches its minimum within the first epochs, followed by a progressive increase. This behavior indicates rapid convergence and a tendency to overfit

under strict cross-subject validation, especially given the limited validation set size (three subjects per fold) and the high expressiveness of the classifier operating on informative foundation embeddings. Therefore, early stopping plays a crucial role in selecting the best epoch for each fold, and the qualitative inspection of training dynamics provides useful evidence for understanding why some folds achieve high generalization while others fail.

3.5 Error Analysis and Model Behavior

To fully understand the potential and limitations of the proposed framework, it is not enough to evaluate aggregate metrics. This section breaks down the error of the model that reached better performance (the Multimodal Late Fusion) by analyzing the performance both at the level of global classification (cognitive and emotive classification) and at the finor level of the single task category. In this work we extracted signals captured during 5 different types of stressors:

- Math test, IQ test, and Stroop test are designed to elicit cognitive stress;

- Video and images stimuli, instead, aimed to induce a passive emotional stress.

In figure 3.6 is represented the global Confusion Matrix of the best model obtained, that of the Multimodal Late Fusion with no adversarial training. Across all 60 participants, 5068 patches out of 6947 has been correctly classified as 'cognitive' by the model, representing 72,9%, while around 17,1% have been wrongly classified as 'emotive'. For what conceirns the emotive tasks, the correct patches classified reached 5491, representing 80% of cases.

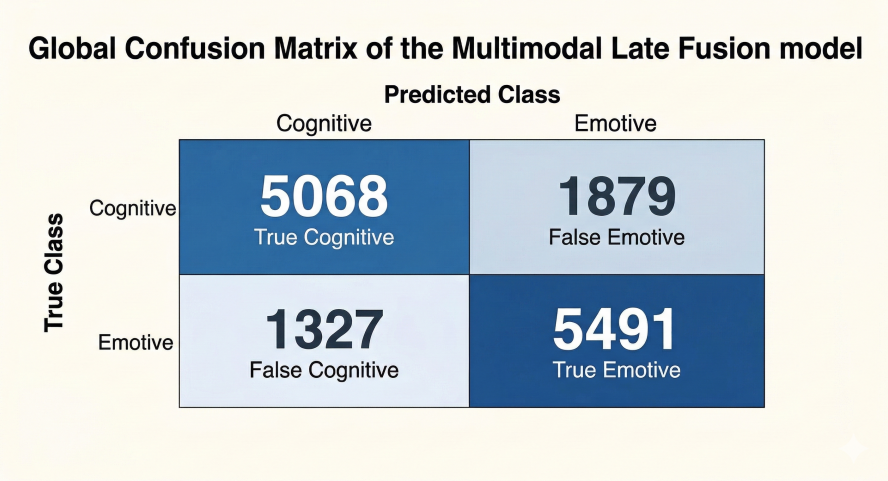


Figure 3.6: Global Confusion Matrix of the Multimodal Late Fusion model. The Confusion Matrix is obtained by summing the Confusion Matrix of all 60 folds.

3.5.1 Performance by Stressor Subtype

To assess whether all stressors are equally distinguishable, an evaluation of the model performance across the stressors stimuli has been performed. This breakdown provides a more nuanced view of the model behavior. The classification accuracy of the Late Fusion model across the five task categories of the CLAS protocol is reported in Figure 3.7.

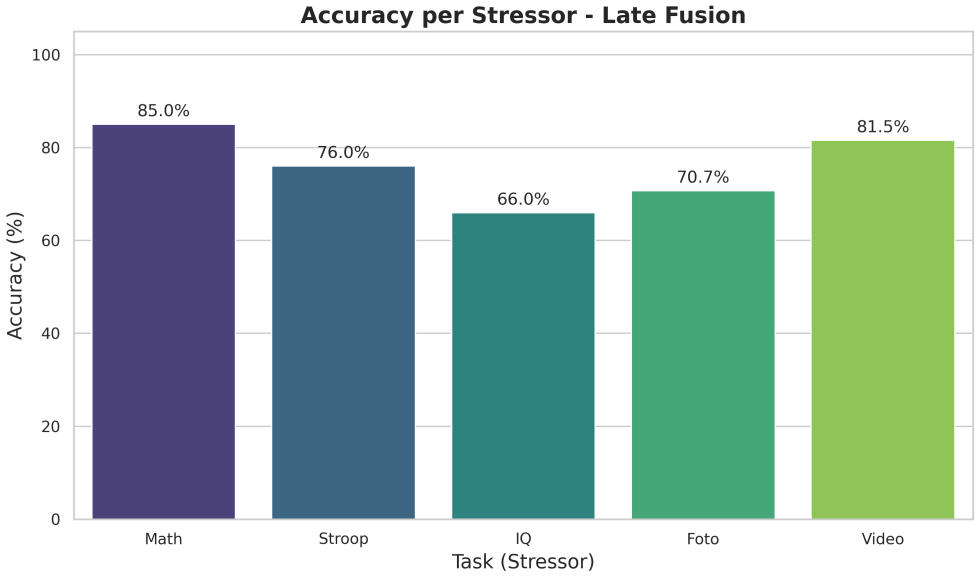


Figure 3.7: Task-wise classification accuracy of the Late Fusion baseline model across the five CLAS stressor categories.

Among interactive tasks, the model achieves the best accuracy in the Math test (85.0%), followed by the Stroop test (76.0%), while the IQ test shows the lowest performance (66.0%). For the emotive task, accuracy is 70.7% for images stimuli and 81.5% for videos stimuli. Overall, these results indicate that the task-level performance is heterogeneous across stressor subtypes, with math test and video being the most reliable classified categories.

3.6 Latent Space Visualization (Qualitative Analysis)

Finally, to complete the analysis of the relative work, a qualitative visualization of the learned space is provided, using the t-SNE method for two representative subjects: a best-case (Figure 3.8) and a worst-case (Figure 3.9). It is important

to note that t-SNE is not a proof of separability and can be sensitive to hyper-parameters such as perplexity. In this work, we set this parameter to a value of $perplexity = 30$. In those figures the ECG embeddings are visualized. Each figure reports three conditions: (i) no preprocessing (raw signals), (ii) preprocessing without baseline correction, and (iii) preprocessing with baseline correction. In the best case (Subject 22), that achieves an accuracy of 100.00%, the two classes (interactive and perceptive) form well-separated clusters. Instead, in the worst-case t-SNE visualization (Subject 27, $Accuracy = 39.06\%$) shown in Figure 3.9, it exhibits substantial overlap between classes across all conditions, indicating a limited class separability in the latent space.

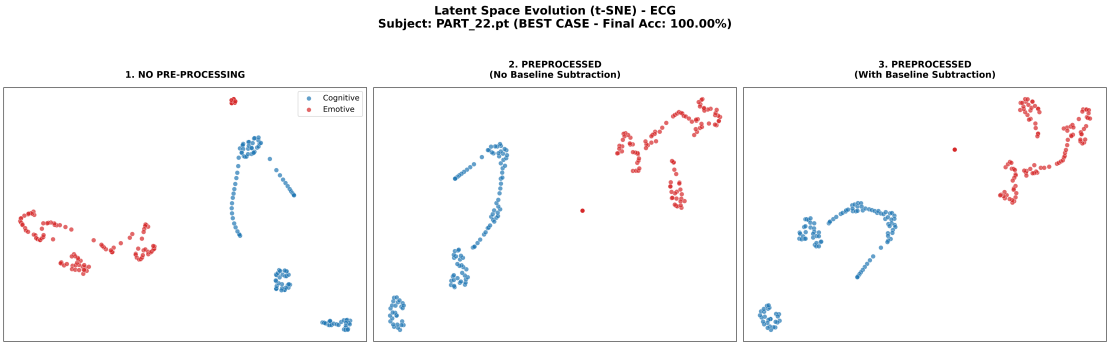


Figure 3.8: Latent space evolution (t-SNE) for the ECG branch – best-case subject. The figure shows a 2D t-SNE projection of ECG embeddings for the subject achieving the highest fold-level performance. Points are colored by class (interactive/cognitive vs perceptive/emotional). From left to right: (1) no preprocessing, (2) preprocessing without baseline correction, and (3) preprocessing with baseline correction.

Two more latent space visualization were performed, to better analyze and interpret the results obtained. Figure 3.11 shows the embeddings of the ECG signal in a latent space projected via t-SNE and colored by the ground-truth labels (cognitive tasks in blue, emotional tasks in red). Rather than forming two monolithic macro-clusters, the t-SNE visualization reveals the presence of pretty well-defined clusters in the embedding space. These clusters appear as separated "islands", each of which is associated to a different subject. This thesis is confirmed by the Figure 3.10, where the ECG embeddings of just 4 randomly chosen subjects were plotted. In this manner, it has been suggested and confirmed that the different "islands" seen in the previous plot belong to different subjects.

Critical insight emerges when inspecting the error distribution in Figure 3.12, where misclassified samples are highlighted. Errors are not uniformly scattered; instead, they concentrate into distinct regions, with entire micro-clusters being consistently misclassified. This clustering indicates that mispredictions are not

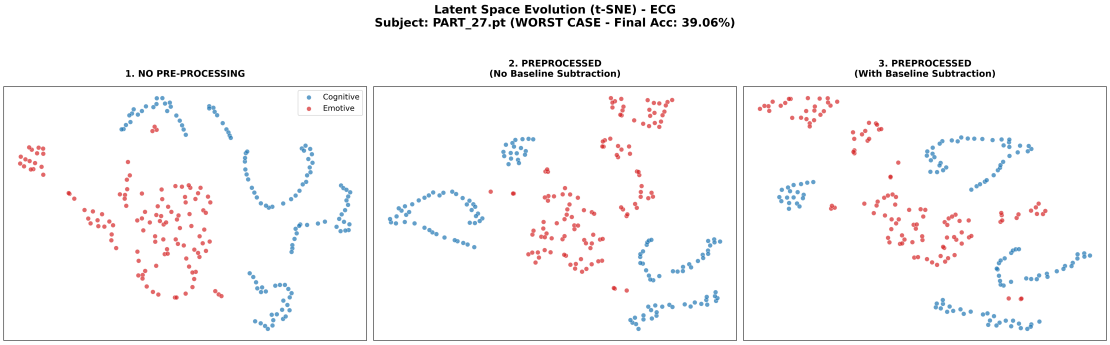


Figure 3.9: Latent space evolution (t-SNE) for the ECG branch – worst-case subject. The figure shows a 2D t-SNE projection of ECG embeddings for the subject achieving the lowest fold-level performance. Points are colored by class (interactive/cognitive vs perceptive/emotional). From left to right: (1) no preprocessing, (2) preprocessing without baseline correction, and (3) preprocessing with baseline correction.

primarily driven by isolated noisy patches, but rather by a systematic mismatch between the model's decision boundary and the physiological response patterns of specific individuals. In other words, for certain outlier subjects the learned representation leads to a coherent but incorrect mapping of the session into the wrong class, which aligns with the presence of the worst-case folds discussed in the previous section. Since t-SNE is a qualitative visualization, these observations are used here to support the fold-level performance analysis rather than as a stand-alone quantitative proof.

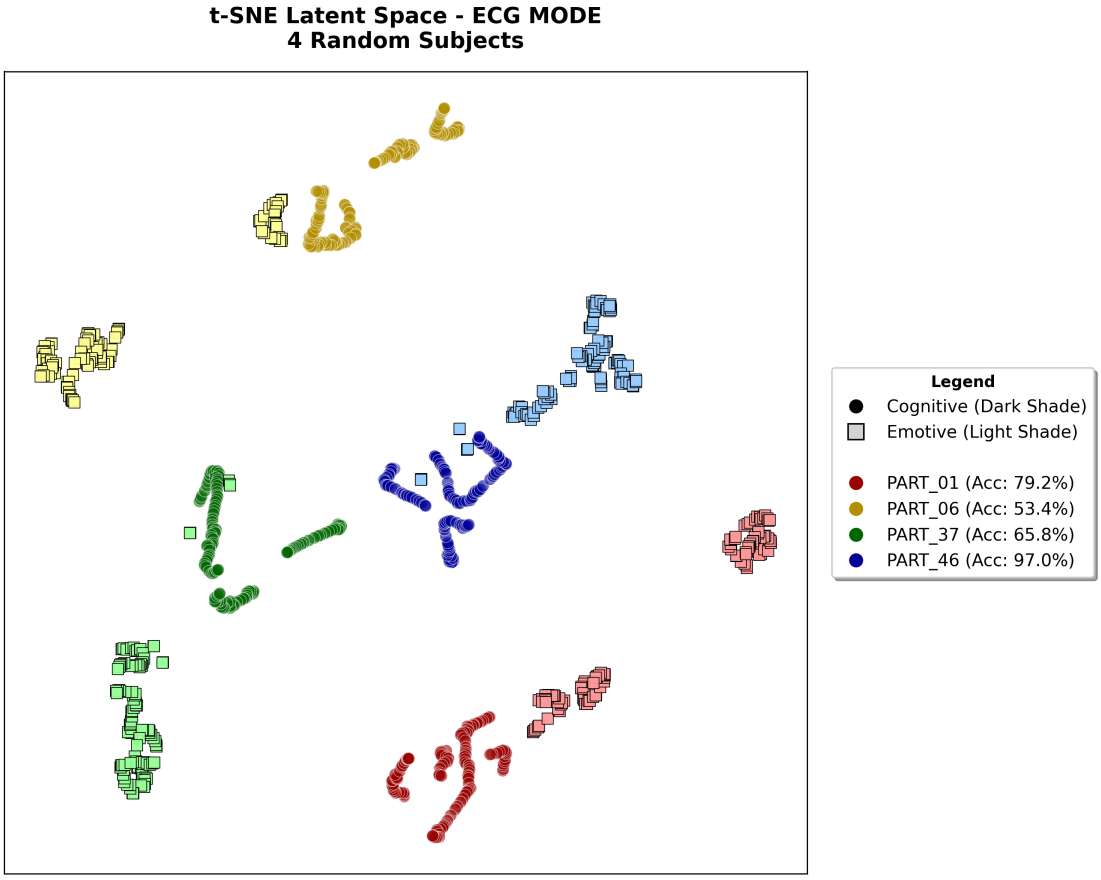


Figure 3.10: t-SNE representation of the ECG embeddings of 4 random subjects.

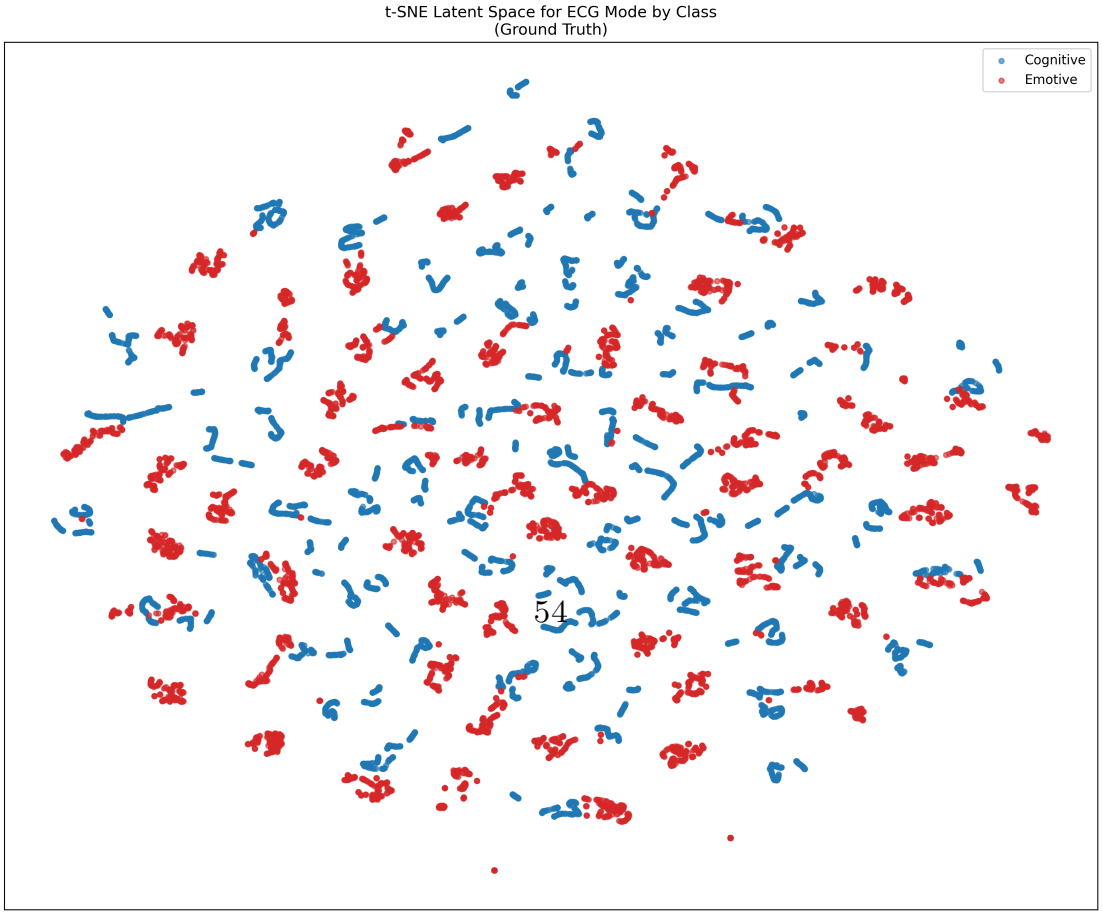


Figure 3.11: t-SNE projection of ECG embeddings for all samples, colored by

Figure 3.12: t-SNE projection of the same ECG embeddings, where misclassified samples are highlighted.

Chapter 4

Discussion

4.1 Summary of Main Findings

This thesis investigated subject-independent discrimination of stressor types (interactive/cognitive vs perceptive/emotional) from wearable physiological signals using a MOMENT-based embedding framework. In particular, the Foundation Model employed in this framework is used in a fully frozen setting, without any task-specific fine-tuning. This choice represents an unexplored approach in physiological signal analysis, where it is not guaranteed that a general-purpose time-series foundation model successfully extracts learning representations for a highly specific task such as stressor discrimination classification. Across all experiments, the evaluation was conducted under a strict subject-independent LOSO cross validation protocol with 60 folds, ensuring that each test subject was fully unseen during training and validation.

At the single-modality level, cardiovascular signals provided the strongest baselines. ECG achieved the highest mean performance among individual branches (Accuracy 0.7057 ± 0.1684, F1-score 0.6506 ± 0.2229), with PPG yielding comparable results (Accuracy 0.6918 ± 0.1328, F1-score 0.6405 ± 0.1808). In contrast, EDA showed clearly lower discriminative power for stressor-type classification (Accuracy 0.6086 ± 0.1240, F1-score 0.5660 ± 0.1537), suggesting that electrodermal dynamics alone are insufficient to robustly separate cognitive and affective stressors under cross-subject testing (Table 3.1, Figure 3.1).

Multimodal integration substantially improved performance. The proposed weighted Late Fusion strategy outperformed Early Fusion and all single-modality branches, achieving Accuracy 0.7649 ± 0.1234, F1-score 0.7639 ± 0.1244 (Table 3.2, Figure 3.2). The learned fusion weights emphasized ECG as the most reliable modality ($w_{\text{ecg}} = 0.50$), while still using complementary information from PPG and EDA ($w_{\text{ppg}} = 0.35$ and $w_{\text{eda}} = 0.15$).

Domain-adversarial training (DANN) was introduced primarily as a robustness probe to test whether explicit domain alignment improves subject-independent generalization beyond the baseline architecture. At the aggregate level, DANN did not produce statistically significant gains in mean performance (Wilcoxon signed-rank test, $p>0.05$), and single-modality branches generally showed comparable or slightly reduced averages under adversarial training (Tables 3.1, 3.3, 3.4). However, robustness analysis revealed that adversarial training improved the lower tail of the performance distribution for Late Fusion, substantially increasing the worst-case fold outcome and raising low-percentile performance (Table 3.5). Training logs further showed modality-dependent dynamics and occasional non-responder folds, emphasizing the sensitivity of adversarial optimization in this setting (Figure 3.5).

Finally, error analysis confirmed heterogeneous behavior across task categories. Late Fusion achieved higher accuracy on Math and Video stimuli (85.0% and 81.5%, respectively), while IQ/Logic and Photo stimuli remained more challenging (66.0% and 70.7%) (Figure 3.7). Qualitative t-SNE visualizations supported the presence of subject-dependent structure and clustered error regions, aligning with the observed variability between best-case and worst-case subjects under subject-independent evaluation (Figures 3.8 and 3.12).

4.2 Interpretation of Single-Modality Results

Single-modality baselines provide an important lens for understanding which physiological channel has been better interpreted from MOMENT in this specific task classification. As discussed in the previous chapters, there are a lot of particular issues when working on physiological biosignals compared to general time-series data used to train the MOMENT foundation model. In this regard, the cardiovascular branches (ECG and PPG signal) consistently outperformed EDA in both mean performance and practical reliability (Table 3.1), suggesting that the separation between interactive (cognitive-load) and perceptive (emotional-arousal) stressors is better decoded from cardiac dynamics signal rather than from electrodermal activity alone.

The ECG branch achieved the highest average accuracy among single modalities and, importantly, produced discriminative performance despite the strong domain shift inherent to subject-independent testing. A plausible explanation is that interactive stressors (time-pressured, goal-directed tasks) elicit relatively consistent cardiovascular responses, such as changes in heart rate dynamics and autonomic balance, that are sufficiently structured to be captured even after embedding extraction. ECG also provides richer and more stable cardiac information than PPG, as it is less affected by artifacts such as movement artifacts and offers a signal morphology that is highly informative for rhythm-related variations. At

the same time, the ECG results exhibit a relatively high standard deviation across folds (Table 3.1), indicating that generalization at the subject-level remains challenging. This variability is consistent with the idea that cardiac reactivity to stressors is not uniform: some subjects may show strong physiological differentiation between cognitive and affective stimuli, while others exhibit muted or atypical responses. In other words the ECG signal his highly informative when stressor-dependent reactivity is present, but it can still suffer from "non-responder" cases under subject-independent evaluation. A notable observation is that the ECG branch achieves the strongest performance despite operating on relatively short temporal segments of 4 seconds. This indicates that a substantial portion of the discriminative signal between interactive and perceptive stressors is encoded in short-term cardiovascular dynamics, such as rapid changes in heart rate and beat-to-beat variability, rather than requiring long recordings. From an application perspective, this finding supports the potential use of ultra-short HRV metrics (and other short-window cardiac markers) for near-real-time monitoring, where low latency is crucial. Nevertheless, HRV computed on ultra-short segments is inherently less stable than HRV derived from standard time windows, and should be interpreted as a proxy of short-term cardiac dynamics rather than a replacement for conventional HRV assessment.

PPG achieved performance close to ECG in mean terms, with slightly lower averages and a more moderate variance. This behavior is coherent with the nature of PPG: it captures cardiovascular dynamics through peripheral blood volume changes and is therefore sensitive to factors such as sensor contact, motion artifacts, and vasoconstriction. These confounders can reduce the fidelity of stress-related pulse dynamics for certain subjects or sessions, limiting peak performance. Nonetheless, the fact that PPG remains competitive suggests that a large portion of the discriminative information needed for stressor discrimination is present in cardiovascular dynamics that are observable even through wearable-friendly sensing. From an application standpoint, this result is meaningful: although ECG remains the strongest single modality, PPG provides a valuable alternative in wearable-first scenarios where ECG acquisition may be less practical, and it can contribute complementary information when combined with ECG and EDA.

EDA is a well-established marker of sympathetic arousal and is often highly informative for detecting whether arousal increases relative to baseline. However, the present task is not binary arousal detection but discrimination between two stressor categories. Both interactive and perceptive stressors can increase sympathetic activation, and therefore EDA may show partially overlapping patterns between classes. In addition, electrodermal signals evolve slowly and are heavily influenced by subject-specific baseline skin conductance, electrode contact, and individual sweating dynamics. Moreover, this signal, compared to the other two signals, is the one that is more affected by noise artifacts. In fact, during the CLAS protocol,

the EDA signal is captured from a device worn on the fingers, which is more likely to be affected by movement artifacts. These factors reduce the separability of stressor types under subject-independent evaluation, even after preprocessing and embedding extraction. The global confusion matrices (Figure 3.1) support this interpretation by showing modality-dependent error patterns: compared to ECG and PPG, EDA exhibits a higher confusion rate between the two classes, indicating weaker class-specific structure in the electrodermal representation. Importantly, this does not imply that EDA is useless; rather, it suggests that EDA is better suited as a complementary modality in a multimodal framework, where it can add arousal-related cues that may be informative for specific subjects or stimulus types.

4.2.1 Implications for Multimodal Design

Overall, the single-modality results motivate two design choices that are central to this thesis. First, the ECG embeddings extracted from the frozen time-series foundation model MOMENT turned out to be the anchor modality for stressor discrimination, as they provide the strongest single-channel evidence. Second, they support the rationale for multimodality: because the embeddings of each channel capture a different aspect of autonomic activation and exhibit distinct failure modes, combining modalities offers a principled strategy to mitigate subject-level variability and sensor-specific noise. This conclusion directly motivates the fusion analysis presented in the Results chapter 3.3 and provides the basis for discussing why Late Fusion becomes the most effective integration strategy.

4.3 Late Fusion vs. Early Fusion

The results in Table 3.2 and Figure 3.2 show a clear advantage of the Late Fusion strategy over Early Fusion under strict subject-independent evaluation. While Early Fusion yields a mean F1-score of 0.6774 ± 0.2061, the Late Fusion framework reaches 0.7639 ± 0.1244, with a visibly higher median and a substantially tighter distribution in the boxplot. Those differences indicate that Late Fusion strategy improves both average discrimination capability and cross-subject reliability, supporting this preference modality for real-world deployment applications.

A key reason for the observed gap lies in how the two fusion strategies handle modality-specific noise and variability. In Early Fusion, embeddings from EDA, PPG, and ECG are concatenated into a single vector before classification. This design allows the classifier to model cross-modal interactions, but also makes the decision boundary sensitive to the weakest modality: if one signal is noisy, poorly informative for a specific subject or affected by artifacts, its features are injected into the shared space and may degrade the prediction of the classifier. This issue becomes even more enhanced in subject-independent settings, where each test

subject can present a different "dominant" modality or different noise patterns. Late Fusion strategy, instead, mitigates this issue by decoupling representation learning across modalities. Each branch is trained independently in order to produce class probabilities, and fusion is performed only at decision level. As a consequence, failure remains more localized: a noisy modality can yield uncertain predictions without contaminating the representations learned and the overall output of the model. The empirical results in Figure 3.2 are consistent with this interpretation: Late Fusion not only achieves better median F1-score, but also reduces the standard deviation, indicating improved robustness to cross-subject heterogeneity.

In the Multimodal Late Fusion framework the late fusion is implemented as a weighted soft-voting mechanism, combinng the probabilities of the three modality-specific classifiers. The fusion weights were not selected as averages across all folds; instead, they were obtained through a grid search that optimized the F1-score metric. The resulting optimal configuration assign a dominant weight to ECG ($w_{\text{ECG}} = 0.5$), a substantial contribution to PPG ($w_{\text{PPG}} = 0.35$), and a smaller but non null contribution to EDA ($w_{\text{EDA}} = 0.15$). This distribution is coherent with the single-modality results: ECG turned out to be the strongest individual branch, PPG provides complementary cardiovascular information, and EDA contributes secondary arousal-valence related cues that can still be beneficial in some cases. The fact that the optimal weight for EDA is not zero suggests that electrodermal signal information, even if it is weaker on average in the single-modality, helps improve performance in the final fused decision when combined with cardiovascular modalities. Existing literature suggests that EDA is generally a less informative signal than cardiovascular signals for stress-related classification tasks. The authors Radhika and Murthy Oruganti[30] demonstrated that models based on ECG-derived features consistently outperformed those relying only on EDA features on both, the ASCERTAIN and the CLAS datasets. Likewise, Ninh et al.[31], in their studies on the WESAD dataset, reported superior performance for models based on BVP compared to EDA-only approaches. This pattern is found in several classifiers tested, such as Random Forest, Support Vector Machine, or Neural Network architecture. These results indicate that cardiovascular signals provide more discriminative information for stress recognition, whereas EDA may lack sufficient specificity when used alone.

4.4 Latent Space Structure and Subject Bias

To complement the quantitative results, we analyze the structure of the learned representation space through t-SNE visualizations. While t-SNE does not provide a formal measure of separability and is sensitive to hyperparameters, it is useful for developing qualitative intuition about (i) how class information is organized

in the embedding space and (ii) whether errors arise as isolated misclassifications or as subject-level failure modes. In this work, t-SNE was computed with a fixed perplexity (= 30) to ensure consistency across plots.

The best-case and worst-case subject analyses (Figures 3.8 and 3.9) illustrate how subject variability can dominate model behavior even within the same modality (ECG). In the best-case subject (Accuracy = 100%), interactive and perceptive samples form well-separated clusters, indicating that stressor-specific structure is strongly expressed in the representation space. Across the three preprocessing conditions of (i) no preprocessing, (ii) preprocessing without baseline correction, and of (iii) preprocessing with baseline correction, the two classes remain largely separable, suggesting that for this subject the physiological response patterns are consistently discriminative and robust to changes in preprocessing.

In contrast, the worst-case subject (accuracy = 39.06%) exhibits substantial overlap between cognitive and emotive points across all three conditions. This persistent overlap implies that, for certain individuals, the physiological responses elicited by interactive and perceptive stressors are not easily separable in the learned embedding space. Notably, baseline correction may re-center the embedding distribution and reduce some subject-dependent offsets, but it cannot create separability when the underlying physiological response is intrinsically ambiguous or weakly expressed. These observations align with the fold-level performance variability and help contextualize why subject-independent evaluation yields "non-responder" folds even when the overall mean performance is competitive.

Beyond best/worst cases, Figures 3.11 and 3.12 provide a global view of the ECG latent space across subjects. When colored by class (Figure 3.11), the embedding space does not form two monolithic macro-clusters; instead, it shows a fragmented topology with multiple micro-clusters and filament-like structures. This pattern suggests that class-related information is present but intertwined with other variability sources (most plausibly subject-specific physiological traits, session-dependent factors, and signal quality differences) creating locally coherent "islands" in the representation space.

More importantly, visual inspection suggests that many of these micro-clusters correspond to individual subjects. This outcome is highlighted in Figure 3.10 where the embeddings of just 4 subjects are plotted with different colors, in order to immediately visualize if the micro-clusters seen in the entire latent space derived from different subjects. This Figure suggests to confirm this sentence, since it appears that the embeddings belonging to the same subject tend to form compact and isolated regions in the latent space, indicating that subject identity is a dominant organizing factor in the learned representation.

Critical insight emerges when errors are highlighted (Figure 3.12). Misclassifications are not uniformly scattered as stochastic noise; rather, they concentrate in distinct regions, with entire micro-clusters being consistently misclassified. This

indicates that failures often occur at the level of coherent segments of the latent space, consistent with a subject-level or session-level mismatch rather than occasional patch-level artifacts. In practical terms, when the model's learned decision boundary is misaligned for an outlier physiological profile, a large fraction of that subject's samples may be mapped into the wrong class, producing the observed worst-case folds and lower-tail behavior reported in the Results chapter.

These findings highlight both the potential and the limitations of MOMENT foundation model when employed in its frozen settings to extract embeddings from physiological time-series data for a specific task. While MOMENT seems to be able to extract meaningful representations without any task-specific fine-tuning, the resulting embeddings remain strongly influenced by subject-dependent characteristics. This suggest that, despite its general-purpose nature, MOMENT do not inherently provide subject-invariant representations when applied to physiological time-series data. Based on this observation, DANN was introduced to explicitly mitigate subject-dependent representations and promote the learning of subject-invariant features from the extracted embeddings.

4.5 DANN as a Robustness Probe

The domain-adversarial component was introduced in this thesis work primarily as a robustness probe rather than as a performance booster. Physiological time series encode subject-specific characteristics that can dominate the learned representation and degrade cross-subject generalization. By adding a domain discriminator with a Gradient Reversal Layer (GRL), the DANN objective explicitly penalizes subject-identifiable features and encourages the classifier to rely on stressor-related physiological deviations that are more transferable across individuals.

At the aggregate level, the comparison between the baseline Late Fusion and the DANN Late Fusion models shows very similar mean performance. The baseline Late Fusion architecture achieves an accuracy of 0.7649 ± 0.1234 and a F1-score of 0.7639 ± 0.1244, while the DANN Late Fusion reaches 0.7716 ± 0.1247 accuracy and 0.7708 ± 0.1254 F1-score. Although the DANN mean values are slightly higher ($\Delta F1 = +0.0069$, $\Delta Accuracy = +0.0067$), paired Wilcoxon signed-rank tests confirm that these differences are not statistically significant ($p > 0.05$). Therefore, under strict subject-independent evaluation, the two approaches should be regarded as statistically comparable in terms of average discrimination performance.

This outcome aligns with the hypothesis that the embeddings extracted from MOMENT already provide a relatively structured representation space. If the baseline representation were strongly dominated by subject-dependent features, one would expect the adversarial strategy to increase the mean performance. This further supports the idea that the main limitaiton of the Baseline model is not the

absence of discriminative information, but rather its entanglement with subject-specific variability.

While mean performance remains comparable, DANN exhibits a clear advantage when performance is examined from a robustness perspective, i.e., focusing on the lower tail of the fold distribution. As reported in Table 3.5, the worst-case fold improves from $F1 = 0.2540$ (in the baseline Late Fusion model) to 0.4797 (in the Late Fusion DANN architecture), corresponding to a gain of +0.2257. Similarly, the 5th percentile increases from 0.5206 to 0.5453, indicating that DANN raises not only the single worst case, but also the typical behavior of the lowest-performing folds.

This pattern is consistent with the visual impression provided by the two boxplots (Figures 3.2 and 3.3): the overall distribution remains similar, but DANN tends to mitigate extreme collapses for a subset of subjects. In practical terms, this matters because real-world stress-aware systems must handle "difficult" users reliably; improving the worst-case behavior can be more valuable than marginally increasing the mean in controlled settings.

Adversarial training often affects not only accuracy at a fixed decision rule (argmax), but also the quality of the confidence scores. In this regard, pooled ROC curves provide a threshold-independent view of discrimination. The Late Fusion ROC AUC increases from 0.846 (baseline) to 0.857 (DANN) (Figures 3.4), suggesting a modest improvement in ranking/calibration of predicted probabilities even when mean Accuracy/F1 remain statistically unchanged. This supports the interpretation that DANN can refine the decision geometry of the classifier without reliably shifting the average performance at the default threshold.

Finally, the training-log analysis presented in Section 3.4.2 highlights that DANN introduces fold-dependent dynamics, with both successful convergence and failure/non-responder regimes. This observation provides a mechanistic explanation for why robustness improvements may concentrate in the lower tail: adversarial alignment can prevent severe subject-specific failures in some folds, while leaving average performance largely unchanged when the baseline embeddings are already robust.

4.6 Task-Wise Behavior and Error Patterns

An in depth-analysis of the classification is actuated to evaluate the effect of the task type classification performance. In the CLAS dataset, interactive and perceptive tasks include different type of stressors that are not homogeneous. Interactive tasks aggregate math test, stroop test, and IQ test, while perceptive aggregate images and video stimuli. These subtypes of stimuli differ in cognitive demands, affective content, and duration, which can lead to partially distinct physiological responses

and consequently to different degrees of separability in the learned representation.

The Late Fusion baseline model exhibits its highest accuracy on Math test (85.0%) and Video (81.5%), with Stroop test reaching 76.0% (Figure 3.7). A plausible interpretation is that math under strict time constraints induces a relatively consistent and sustained cognitive load, often accompanied by clear cardiovascular reactivity patterns. Similarly, video stimuli provide temporally continuous affective content, which may elicit a more stable arousal trajectory compared to shorter or more heterogeneous perceptive stimuli. In other words, both Math and Video tasks can produce "coherent" physiological signatures over time, which improves separability in the embedding space.

These observations also align with the global confusion matrix of the best-performing model (Figure 3.6), where the model achieves a higher true positive rate for emotive samples (80.0% correctly classified as emotive) than for cognitive samples (72.9% correctly classified as cognitive). This asymmetry suggests that perceptive stimuli (especially videos) often generate distinctive patterns that the model can reliably capture across subjects.

The lowest performance is observed for the IQ task (66.0%) and photo stimuli are also comparatively harder (70.7%) than videos (Figure 3.7). A key reason is likely heterogeneity in both stimulus processing and subjective evaluation. IQ tasks may vary in perceived difficulty among participants: for some individuals, the task may not induce substantial cognitive stress, while for others it can trigger frustration or disengagement. This produces variability in physiological reactivity that is not strictly tied to the "interactive" label, increasing overlap with the perceptive class.

A similar argument applies to photo stimuli. Compared to videos, affective images are typically shorter and can elicit a more heterogeneous response depending on personal sensitivity, habituation, and semantic interpretation. Consequently, the physiological trajectory may be less sustained and the resulting windows can be more ambiguous. This is consistent with the overall findings of stressor discrimination: some stressors sit near the boundary between cognitive load and affective arousal, and their physiological signatures can partially overlap.

4.7 Limitations

Despite the promising results obtained in this thesis work, several limitations should be considered when interpreting the findings and assessing their potential for real-world applications.

First, all experiments were conducted on the CLAS dataset, which was collected under a controlled experimental protocol with predefined tasks and stimuli. Although the protocol includes both interactive and perceptive stressors, it remains confined to laboratory conditions. Consequently, the reported results may not

directly generalize to real-world scenarios, where contextual variability is higher and physiological signals are often noisier and less structured.

Second, this work focuses on discriminating between stressor categories (interactive/cognitive vs perceptive/emotional), rather than distinguishing between stress and non-stress conditions. This classification depends on the assumptions underlying the experimental protocol, namely that the selected tasks reliably elicit the intended cognitive or affective states. However, the same stimulus may induce different responses across subjects, leading to variability that is not fully captured by the labeling scheme.

Another limitation concerns the partial use of the available data f the CLAS dataset. In addition to ECG, PPG, and EDA, the dataset includes three-axis accelerometer signals, which were not integrated into the proposed framework. Motion information could have been exploited to identify segments affected by movement artifacts, particularly for PPG and EDA signals, which are more sensitive to acquisition noise. Incorporating such information could improve signal quality assessment and potentially enhance model robustness.

Furthermore, the dataset does not include self-report measures of perceived stress or affective state. The availability of subjective ratings after each task would have provided valuable complementary information, allowing verification of the effectiveness of the stimuli at the individual level. In stress and affective computing research, self-reports (e.g., perceived stress, arousal, or valence ratings) are often considered a reference standard for interpreting physiological responses and validating experimental conditions.

Finally, the MOMENT foundation model was employed in a fully frozen setting, without any task-specific fine-tuning. While this choice allows for evaluating the transferability of general-purpose representations, it may limit the model's ability to capture task-specific patterns relevant for stressor discrimination. Future work could explore fine-tuning strategies to adapt the pretrained model to the specific characteristics of physiological signals and the target classification task, potentially leading to improved performance and better subject-invariant representations.

Chapter 5

Conclusion

This thesis work addressed the problem of subject-independent stressor discrimination from wearable physiological signals, shifting the focus from the common binary *stress vs. non-stress* paradigm toward a more concrete goal: distinguishing interactive stressors from emotional ones. To this end, a multimodal framework based on the MOMENT foundation model is developed. MOMENT is used in a frozen setting without any fine-tuning step. This represents a novelty in literature, as this type of application of a foundation model has never been tested on physiological signals. The role of MOMENT is to act as a feature extractor of time-series segments taken from the three biological signals (EDA, PPG, and ECG). The extracted embeddings enter then a lightweight classifiers (one for each channel) based on MLP and a final Mulitmodal Late Fusion. All experiments were evaluated under a strict Leave-One-Subject-Out cross-validation protocol with 60 folds, ensuring that each test subject remained unseen during training and validation phases.

Overall, the results show that cardiovascular dynamics are the strongest single-modality for this stress discrimination task, with ECG and PPG outperforming EDA. The ECG branch emerged as the strongest individual modality, achieving the highest mean accuracy and F1-score. This indicates that the distinction between cognitive load and passive emotional arousal is robustly encoded in short-term cardiac dynamics. The PPG signal proved to be a highly reliable alternative, confirming its value for wearable applications where ECG acquisition might be obtrusive. Conversely, the EDA branch yielded the lowest discriminative performance, suggesting that while EDA is a known marker of general sympathetic arousal, it is insufficient for separating specific stressor categories on its own.

More importantly, this work demonstrates that multimodal Late Fusion yields the most reliable performance, improving both average and variability metrics, compared to Early Fusion and single-modalities models. This result supports the conclusion that stressor discrimination benefits from combining heterogeneous

physiological signals, while keeping mode-specific decision paths separate until the inference stage. The optimal weights assigned a dominant role to the ECG signal ($w_{\text{ECG}} = 0.50$), followed by the PPG ($w_{\text{PPG}} = 0.35$) and EDA signals ($w_{\text{EDA}} = 0.15$), confirming the need for a multimodal approach for stressor discrimination tasks.

A secondary objective of this thesis is to evaluate the impact of Domain Adversarial Training as an experimental probe to test whether explicit domain adaptation improves performance beyond MOMENT embeddings. While improvements in mean metrics were not statistically significant, DANN provided measurable gains in worst-case and low-percentile behavior, highlighting its potential role as a robustness mechanism for difficult subjects.

Further analysis revealed that the model's performance is heterogeneous across different types of stressors. Tasks that induce sustained and consistent physiological responses, such as the Math test and Video stimuli, were classified with higher accuracy. In contrast, stressor stimuli such as IQ test and image stimuli proved to be more challenging in stress discrimination task. This task-wise variability was visually supported by qualitative analysis of the latent space, which confirmed that misclassifications often occur as clusters rather than isolated patches, indicating a systemic subject-level misunderstanding of the physiological response.

The present study opens several avenues for future work:

- Ecological validation: test the framework on in-the-wild datasets or collect naturalistic data to evaluate robustness to motion, context changes, and label noise.

- Richer supervision: integrate self-report measures (arousal/valence or perceived stress intensity) to validate stimulus effectiveness per subject and reduce label ambiguity.

- Context and artifact handling: incorporate accelerometer signals to detect motion artifacts and condition the classifier on activity context, especially for PPG and EDA.

- Uncertainty and reliability-aware fusion: replace fixed weights with dynamic reliability estimation (e.g., confidence calibration, modality dropout, missing-modality handling).

- Personalization without calibration burden: explore lightweight adaptation strategies (e.g., few-shot calibration, test-time adaptation) to address non-responder subjects while preserving scalability.

- Foundation model adaptation: evaluate when partial fine-tuning or parameter-efficient adaptation of MOMENT improves performance, and how this interacts with domain shift.

In conclusion, this thesis provides evidence that time-series foundation models can serve as strong subject-independent representation engines for physiological stressor discrimination, reducing the reliance on complex handcrafted pipelines and heavy domain adaptation. While inter-subject variability remains a fundamental challenge, the combination of pretrained embeddings in a zero-shot setting and multimodal Late Fusion offers a robust baseline for future physiological HCI systems aimed at stress-aware, adaptive, and user-centered interaction.

Bibliography

[1] Hans Selye. *The Stress of Life*. New York: McGraw-Hill, 1956.

[2] Giorgos Giannakakis et al. «Review on Psychological Stress Detection Using Biosignals». In: *IEEE Transactions on Affective Computing* 13.1 (Jan. 2022), pp. 440–460. ISSN: 1949-3045, 2371-9850. DOI: 10.1109/TAFFC.2019.2927337. URL: https://ieeexplore.ieee.org/document/8758154/ (visited on 03/01/2026).

[3] Lili Zhu et al. «Stress Detection Through Wrist-Based Electrodermal Activity Monitoring and Machine Learning». In: *IEEE Journal of Biomedical and Health Informatics* 27.5 (May 2023), pp. 2155–2165. ISSN: 2168-2194, 2168-2208. DOI: 10.1109/JBHI.2023.3239305. URL: https://ieeexplore.ieee.org/document/10024755/ (visited on 03/02/2026).

[4] Ritu Tanwar, Ghanapriya Singh, and Pankaj Kumar Pal. «A hybrid transposed attention based deep learning model for wearable and explainable stress recognition». en. In: *Computers and Electrical Engineering* 119 (Nov. 2024), p. 109551. ISSN: 00457906. DOI: 10.1016/j.compeleceng.2024.109551. URL: https://linkinghub.elsevier.com/retrieve/pii/S0045790624004786 (visited on 03/01/2026).

[5] Seyedmajid Hosseini et al. «A multimodal sensor dataset for continuous stress detection of nurses in a hospital». en. In: *Scientific Data* 9.1 (June 2022), p. 255. ISSN: 2052-4463. DOI: 10.1038/s41597-022-01361-y. URL: https://www.nature.com/articles/s41597-022-01361-y (visited on 03/01/2026).

[6] Rafael A Calvo and Sidney D'Mello. «Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications». In: *IEEE Transactions on Affective Computing* 1.1 (Jan. 2010), pp. 18–37. ISSN: 1949-3045. DOI: 10.1109/T-AFFC.2010.1. URL: http://ieeexplore.ieee.org/document/5520655/ (visited on 03/02/2026).

[7] Stephen H. Fairclough. «Fundamentals of physiological computing». en. In: *Interacting with Computers* 21.1-2 (Jan. 2009), pp. 133–145. ISSN: 09535438. DOI: 10.1016/j.intcom.2008.10.011. URL: https://academic.oup.com/iwc/article-lookup/doi/10.1016/j.intcom.2008.10.011 (visited on 03/02/2026).

[8] Philip Schmidt et al. «Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection». en. In: *Proceedings of the 20th ACM International Conference on Multimodal Interaction*. Boulder CO USA: ACM, Oct. 2018, pp. 400–408. ISBN: 9781450356923. DOI: 10.1145/3242969.3242985. URL: https://dl.acm.org/doi/10.1145/3242969.3242985 (visited on 03/02/2026).

[9] Alexios-Fotios A. Mentis, Donghoon Lee, and Panos Roussos. «Applications of artificial intelligence-machine learning for detection of stress: a critical overview». en. In: *Molecular Psychiatry* 29.6 (June 2024), pp. 1882–1894. ISSN: 1359-4184, 1476-5578. DOI: 10.1038/s41380-023-02047-6. URL: https://www.nature.com/articles/s41380-023-02047-6 (visited on 03/02/2026).

[10] James A. Russell. «A circumplex model of affect.» en. In: *Journal of Personality and Social Psychology* 39.6 (Dec. 1980), pp. 1161–1178. ISSN: 1939-1315, 0022-3514. DOI: 10.1037/h0077714. URL: https://doi.apa.org/doi/10.1037/h0077714 (visited on 03/02/2026).

[11] Ydwine Jieldouw Zanstra and Derek William Johnston. «Cardiovascular reactivity in real life settings: Measurement, mechanisms and meaning». en. In: *Biological Psychology* 86.2 (Feb. 2011), pp. 98–105. ISSN: 03010511. DOI: 10.1016/j.biopsycho.2010.05.002. URL: https://linkinghub.elsevier.com/retrieve/pii/S0301051110001353 (visited on 03/02/2026).

[12] S. Koelstra et al. «DEAP: A Database for Emotion Analysis ;Using Physiological Signals». In: *IEEE Transactions on Affective Computing* 3.1 (Jan. 2012), pp. 18–31. ISSN: 1949-3045. DOI: 10.1109/T-AFFC.2011.15. URL: http://ieeexplore.ieee.org/document/5871728/ (visited on 03/02/2026).

[13] Aleksandr Ometov et al. «Stress and Emotion Open Access Data: A Review on Datasets, Modalities, Methods, Challenges, and Future Research Perspectives». en. In: *Journal of Healthcare Informatics Research* 9.3 (Sept. 2025), pp. 247–279. ISSN: 2509-4971, 2509-498X. DOI: 10.1007/s41666-025-00200-0. URL: https://link.springer.com/10.1007/s41666-025-00200-0 (visited on 03/01/2026).

[14] Zeeshan Ahmad and Naimul Khan. «A Survey on Physiological Signal-Based Emotion Recognition». en. In: *Bioengineering* 9.11 (Nov. 2022), p. 688. ISSN: 2306-5354. DOI: 10.3390/bioengineering9110688. URL: https://www.mdpi.com/2306-5354/9/11/688 (visited on 03/02/2026).

[15] Exarchos TP Lazarou E. «Predicting stress levels using physiological data: Real-time stress prediction models utilizing wearable devices.» In: *AIMS Neurosci* (2024).

[16] Lan-lan Chen, Ao Zhang, and Xiao-guang Lou. «Cross-subject driver status detection from physiological signals based on hybrid feature selection and transfer learning». en. In: *Expert Systems with Applications* 137 (Dec. 2019), pp. 266–280. ISSN: 09574174. DOI: 10.1016/j.eswa.2019.02.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S0957417419301046 (visited on 03/02/2026).

[17] Maciej Dzieżyc et al. «Can We Ditch Feature Engineering? End-to-End Deep Learning for Affect Recognition from Physiological Sensor Data». en. In: *Sensors* 20.22 (Nov. 2020), p. 6535. ISSN: 1424-8220. DOI: 10.3390/s20226535. URL: https://www.mdpi.com/1424-8220/20/22/6535 (visited on 03/02/2026).

[18] Yaroslav Ganin et al. «Domain-Adversarial Training of Neural Networks». In: *Domain Adaptation in Computer Vision Applications*. Ed. by Gabriela Csurka. Cham: Springer International Publishing, 2017, pp. 189–209. ISBN: 9783319583464. DOI: 10.1007/978-3-319-58347-1_10. URL: http://link.springer.com/10.1007/978-3-319-58347-1_10 (visited on 03/02/2026).

[19] Yuxuan Wang et al. *Deep Time Series Models: A Comprehensive Survey and Benchmark*. 2024. DOI: 10.48550/ARXIV.2407.13278. URL: https://arxiv.org/abs/2407.13278 (visited on 03/02/2026).

[20] Yuxuan Liang et al. «Foundation Models for Time Series Analysis: A Tutorial and Survey». en. In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. Barcelona Spain: ACM, Aug. 2024, pp. 6555–6565. ISBN: 9798400704901. DOI: 10.1145/3637528.3671451. URL: https://dl.acm.org/doi/10.1145/3637528.3671451 (visited on 03/02/2026).

[21] Mononito Goswami et al. *MOMENT: A Family of Open Time-series Foundation Models*. 2024. DOI: 10.48550/ARXIV.2402.03885. URL: https://arxiv.org/abs/2402.03885 (visited on 03/02/2026).

[22] Pradeep K. Atrey et al. «Multimodal fusion for multimedia analysis: a survey». en. In: *Multimedia Systems* 16.6 (Nov. 2010), pp. 345–379. ISSN: 0942-4962, 1432-1882. DOI: 10.1007/s00530-010-0182-0. URL: http://link.springer.com/10.1007/s00530-010-0182-0 (visited on 03/02/2026).

[23] Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. «Multimodal deep learning for biomedical data fusion: a review». en. In: *Briefings in Bioinformatics* 23.2 (Mar. 2022), bbab569. ISSN: 1467-5463, 1477-4054. DOI: 10.1093/bib/bbab569. URL: https://academic.oup.com/bib/article/doi/10.1093/bib/bbab569/6516346 (visited on 03/02/2026).

[24] Vasileios-Rafail Xefteris et al. «A Multimodal Late Fusion Framework for Physiological Sensor and Audio-Signal-Based Stress Detection: An Experimental Study and Public Dataset». en. In: *Electronics* 12.23 (Dec. 2023), p. 4871. ISSN: 2079-9292. DOI: 10.3390/electronics12234871. URL: https://www.mdpi.com/2079-9292/12/23/4871 (visited on 03/02/2026).

[25] Valentina Markova, Todor Ganchev, and Kalin Kalinkov. «CLAS: A Database for Cognitive Load, Affect and Stress Recognition». In: *2019 International Conference on Biomedical Innovations and Applications (BIA)*. Varna, Bulgaria: IEEE, Nov. 2019, pp. 1–4. ISBN: 9781728147543. DOI: 10.1109/BIA48344.2019.8967457. URL: https://ieeexplore.ieee.org/document/8967457/ (visited on 02/16/2026).

[26] Ilya Loshchilov and Frank Hutter. *Decoupled Weight Decay Regularization*. 2017. DOI: 10.48550/ARXIV.1711.05101. URL: https://arxiv.org/abs/1711.05101 (visited on 03/03/2026).

[27] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. DOI: 10.48550/ARXIV.2103.00020. URL: https://arxiv.org/abs/2103.00020 (visited on 03/03/2026).

[28] Dominique Makowski et al. «NeuroKit2: A Python toolbox for neurophysiological signal processing». en. In: *Behavior Research Methods* 53.4 (Aug. 2021), pp. 1689–1696. ISSN: 1554-3528. DOI: 10.3758/s13428-020-01516-y. URL: https://link.springer.com/10.3758/s13428-020-01516-y (visited on 03/03/2026).

[29] Mohamed Elgendi et al. «Systolic Peak Detection in Acceleration Photoplethysmograms Measured from Emergency Responders in Tropical Conditions». en. In: *PLoS ONE* 8.10 (Oct. 2013). Ed. by Vladimir E. Bondarenko, e76585. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0076585. URL: https://dx.plos.org/10.1371/journal.pone.0076585 (visited on 03/03/2026).

[30] Radhika K and V Ramana Murthy Oruganti. «Transfer Learning for Subject-Independent Stress Detection using Physiological Signals». In: *2020 IEEE 17th India Council International Conference (INDICON)*. 2020 IEEE 17th India Council International Conference (INDICON). New Delhi, India: IEEE, Dec. 10, 2020, pp. 1–6. ISBN: 9781728169163. DOI: 10.1109/INDICON49873.2020.9342505. URL: https://ieeexplore.ieee.org/document/9342505/ (visited on 03/19/2026).

[31] Van-Tu Ninh et al. «An Improved Subject-Independent Stress Detection Model Applied to Consumer-grade Wearable Devices». en. In: *Advances and Trends in Artificial Intelligence. Theory and Practices in Artificial Intelligence*. Ed. by Hamido Fujita et al. Vol. 13343. Cham: Springer International Publishing, 2022, pp. 907–919. ISBN: 9783031085291. DOI: 10.1007/978-3-031-08530-7_77. URL: https://link.springer.com/10.1007/978-3-031-08530-7_77 (visited on 02/16/2026).