IMPERIAL

Rethinking Stress Monitoring: Convenient Modular Early-Onset Multimodal Stress Detection with Attention Score Caching

Author

W. F. Powell

CID: 01738637

Supervised by

Prof. D. Farina

Dr. A. Spiers

A Thesis submitted in fulfilment of requirements for the degree of
Master of Science in Applied Machine Learning

Department of Electrical and Electronic Engineering
Imperial College London

2024

Abstract

This paper proposes a robust, modular multimodal method for the early-onset detection of stress using several innovative, consumer friendly devices, which have not been previously evaluated in the literature, namely mastoid based sEMG and EEG and single-channel consumer-grade, pre-frontal cortex fNIRS. The mastoid provides a discrete, convenient monitoring site through headphones, while consumer fNIRS offers a portable, underexplored stress monitoring modality. We introduce the MUSED dataset, utilizing the novel devices, in addition to popular modalities for stress monitoring. A virtual-reality based trier social stress test is used to induce stress. We collected data from 18 participants (10m, 8f), aged 26.33 $\pm$ 4.77. While key sEMG and EEG biomarkers related to stress were identified at the mastoid, their statistical significance was limited due to noise and signal robustness challenges. Features derived from fNIRS showed strong predictive capabilities but suffered from high inter-participant variability.

We designed a truly modular architecture to adapt to varying combinations of available devices, utilizing an ensemble of bidirectional cross-attention and self-attention blocks for temporal alignment of different modalities. Extensive testing on two well-known datasets, WESAD and UBFC-Phys, demonstrates that our cross-attention network effectively learns and maintains an intermediate representation that generalize well across various modality combinations. This modular approach resulted in only a 2% reduction in performance compared to a fixed model trained exclusively on the same modalities and combinations used during testing. We also employ personalization techniques to address inter-participant variability issues and reduce computational overhead. For early-onset detection, we employed a sequence-to-sequence attention model and attention score caching mechanism, reducing the required window size by a factor of five compared to existing methods, without compromising performance, due to a significant reduction in computational complexity of our mechanism. We outperform several state-of-the-art methods while utilizing a fraction of the window size and enabling true modularity between modalities.

Declaration of Originality

I hereby declare that the work presented in this thesis is my own unless otherwise stated. To the best of my knowledge the work is original and ideas developed in collaboration with others have been appropriately referenced.

Copyright Declaration

Acknowledgments

I would like to thank Professor Dario Farina, who has been continually patient and helpful through planning, submitting ethics, conducting and interpreting the results of the study, as well as the Farina Lab who have been very welcoming. I would also like to thank Dr Ad Spiers who kindly accepted the role of supervisor for my self-proposed thesis despite his numerous other students, offering me clear and helpful guidance on the evaluation and presentation of signal processing and machine learning approaches. I am grateful to Dr. Aaron Zhao for his time and advice, determining the plausibility of the attention and caching mechanisms proposed in this thesis. Additionally, to Han Yu for his support in expanding on his implementation and architecture of the Modality Fusion Network. I commend Saskia Steel and Professor Gregor Domes for developing the Open TSST-VR protocol and helping with its adaption into English. On a personal note, I'd like to thank Daria for her support this year and my parents for their support and emphasis on a good education as I conclude my academics. Finally, I would like to thank both BrainPatch and SilverLine Research for granting me the opportunity to investigate this topic on their behalf.

Contents

List of Acronyms

ECG Electrocardiogram

EDR ECG Derived Respiration

EEG Electroencephalogram

sEMG Surface Electromyography

PPG Photoplethysmography

BVP Blood Volume Pulse

HRV Heart Rate Variability

SNS Sympathetic Nervous System

PNS Parasympathetic Nervous System

EDA Electrodermal Activity

GVS Galvanic Vestibular Stimulation

HPA Hypothalamic-Pituitary-Adrenal

TSST Trier Social Stress Test

VR Virtual Reality

STAI State-Trait Anxiety Inventory

PANAS Positive and Negative Affect Schedule

SSSQ Short Stress State Questionnaire

AS Attention Score

fNIRS Functional Near-Infrared Spectroscopy

RESP Respiration

ACC Acceleration

KV Key-Value

LOSO-CV Leave-One-Subject-Out Cross-Validation

fNIRS Functional Near-Infrared Spectroscopy

BCSA Bi-directional Cross and Self Attention

IBI Interbeat Intervals

List of Figures

1 Introduction

Contents

1.1 Overview

Continuous exposure to psychological stress can adversely affect an individual's mental and physical health; however, automated stress detection has shown to alleviate these symptoms [1]. Smart wearables that can achieve accurate measurements of physiological data are becoming increasingly available and recent advancements in wearable sensors, signal processing and machine learning have enabled a surge in consumer metrics for tracking health metrics. However, the biosensors utilised in these devices are prone to noise and movement artefacts, often serving as a random number generator when only one biosignal is used exclusively. Reliably and consistently measuring metrics such as psychological stress have proven challenging across mainstream wearables and to combat this, multi-modal monitoring systems should be employed to obtain more reliable measurements. However, this arrangement is not commonly used by consumers due to the inconvenience of wearing two devices.

1

1.2 Motivation and Rationale

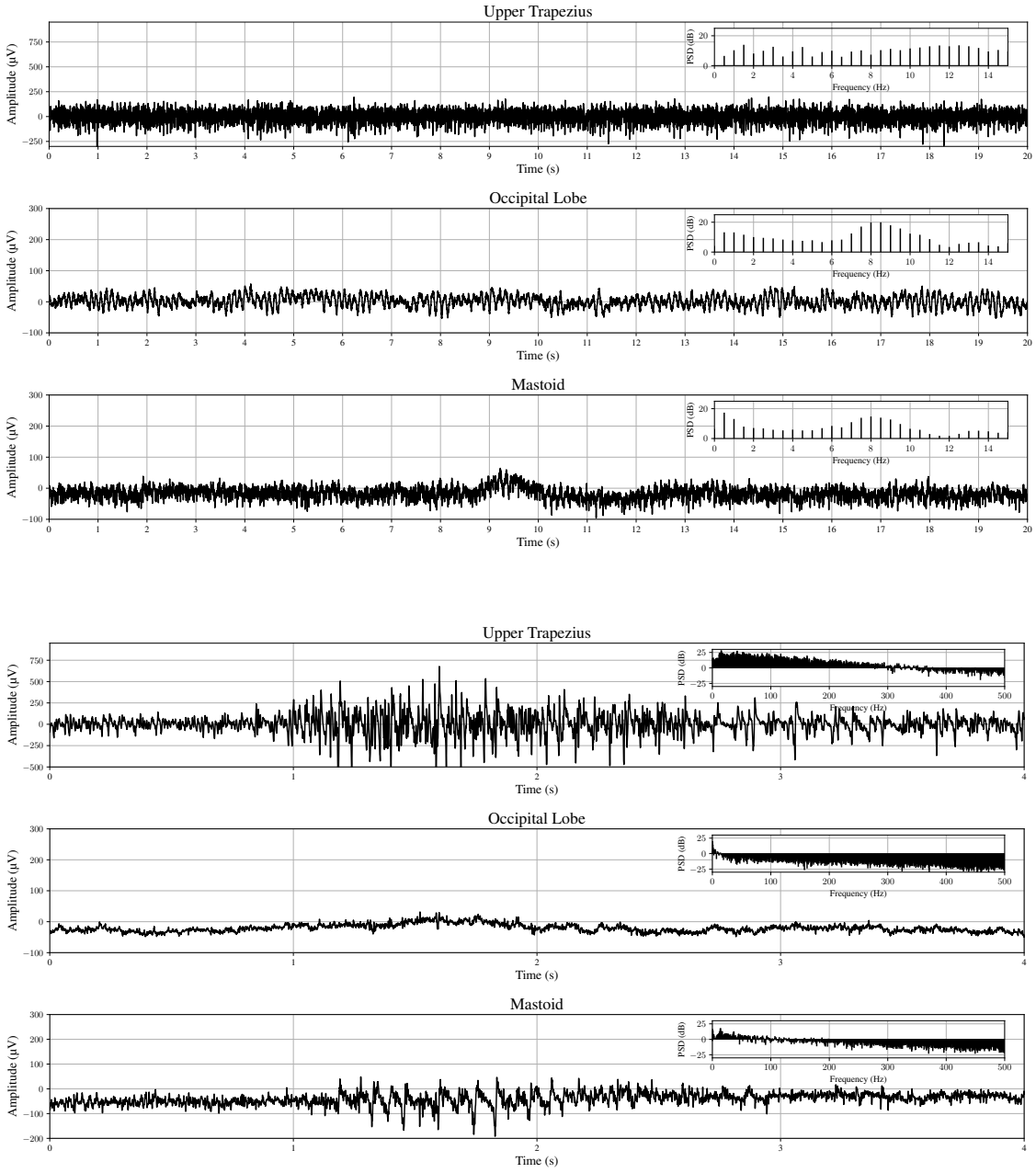
Excessive stress accumulation can cause burn-out [2], [3] that currently affects one in five of the UK workforce [4]. Long-term effects include suppression of the immune system, and increase the likelihood of chronic diseases such as high blood pressure, heart disease, and cancer [5]. According to the World Health Organization, there has been a 25% surge in anxiety and depression cases globally due to the COVID-19 outbreak [6], exposing the current lack of tools to diagnose and treat mental health issues. Automated stress detection synthesised with stress mitigation techniques such as GVS, a type of non-invasive brain stimulation, shows great potential for diminishing these effects [7]. Additionally, through early-onset stress detection, the user will be able to move out of the stressor where possible, which would be particularly beneficial for those who are sensitive to sensory overload, such as those with autism.

This project has been motivated by BrainPatch, an Imperial spinout and now neurotechnology company, who aim to combat stress through galvanic vestibular stimulation. Its effectiveness has been exemplified by a number of studies, particularly Pasquier et al. 2019[7], who demonstrate that it is able to modulate anxiety and corroborates the involvement of the vestibular system in the emotional process. The electrodes are located in headset form, see Figure 1.1.

Figure 1.1: The BrainPatch GVS headset featuring wet electrodes (in white) designed for placement on the mastoids [8].

Previous studies have shown that the ear offers a good location for monitoring EEG [9] and ECG [10], however the devices come with multiple components that need to be correctly positioned, i.e., in-ear EEG electrodes or ECG electrodes on the earlobe. These devices are more complex and require greater effort to wear properly compared to user-friendly wearables like the BrainPatch

headset, which are simple to put on and already familiar to most users. Neurable's Alcaide et al. study demonstrated that headphones have the capability of measuring EEG, however this was



(b)

Figure 1.2: a). Pilot tests showing that alpha waves can be viewed during a resting, eyes closed condition not only in the occipital lobe, but also the mastoid process, using bipolar silver-silver chloride electrodes. b) Muscular artefacts from sEMG are visible in both the upper-trapezius region and the mastoid process using the same electrodes during the tensing of the shoulder.

Initial pilot results, shown in Figure 1.2, indicate that both EEG and sEMG signals can be successfully recorded from electrodes positioned at the mastoids. In both the occipital lobe and mastoid regions, pronounced power spectral density (PSD) of alpha waves are observed, suggesting the feasibility of EEG monitoring from these sites. While the sEMG signal from the mastoids is less pronounced compared to that from the upper trapezius muscle, exhibiting approximately one-quarter of the amplitude, the muscle contractions are still detectable. Notably, such contractions were not observed in the occipital lobe region. Additional evidence from supplementary testing is provided in Figures A.1 (a) and (b). If further research demonstrates the electrodes' robustness, this would enable the use of the same device for both stimulation and stress monitoring without additional hardware. Headphones serve as a familiar and unobtrusive medium for users to wear, and this project seeks to test the feasibility of this idea by exploring the potential of a novel multi-modal system that would eventually use the mastoids as a strategic site for monitoring the sympathetic nervous system. This is achieved through the detection of sEMG signals from the upper trapezius muscle and the integration of EEG measurements located at the vestibular nerve, thus potentially offering a more user-friendly and efficient means for continuous stress and physiological monitoring.

BrainPatch have also partnered with SilverLine Research to utilize their single channel, consumer-ready fNIRS device. This device monitors the hemodynamics of the prefrontal cortex, and its capabilities have shown to be an under-investigated area of research. This device could then be utilised as part of the BrainPatch ecosystem as part of the multi-modal system.

This project will thus serve as a feasibility study for the use of a headphone multi-modal system that record concurrently sEMG and EEG at the mastoid, and fNIRS for the integration of stress measurement into the BrainPatch ecosystem. The end goal of integrating stress monitoring into a closed-loop brain stimulation device, to detect stress onset and evaluate the effectiveness of the neurostimulation being administered.

1.3 Intellectual Contribution

Consumer Mastoid sEMG and fNIRS Based Stress Detection Feasibility: A comprehensive review of the relevant literature reveals that no prior studies have demonstrated the feasibility of stress detection via the mastoid region. Consequently, the application of sEMG measurements at the mastoids to detect stress biomarkers associated with the upper trapezius muscle [12], [13], along-

side EEG recordings to capture stress-related brain activity, remains unexplored. Moreover, while stress monitoring through commercial devices such as PPG ECG and EDA are well-established, the potential of a commercial single-channel fNIRS device for stress detection has not yet been evaluated. Previous research has predominantly employed multi-channel laboratory equipment for this purpose [14], [15].

New Dataset for Multimodal Stress Detection

In this paper, we introduce the MUSED dataset, which includes simultaneous monitoring of mastoid sEMG and EEG activity, along with single-channel fNIRS measurements using a commercially available device. This dataset is designed to complement existing datasets, such as WESAD [16] and UBFC-Phys [17], by employing a well-documented protocol for stress induction and capturing a broad range of biosignals. The study protocol has received a favourable opinion from Imperial's Science, Engineering and Technology Research Ethics Committee (SETREC). Upon publication, the MUSED dataset, along with its accompanying documentation, will be made publicly available, and we encourage the research community to utilize it for algorithm development, benchmarking, and advancing knowledge in the field of stress detection.

Novel Modular-Based Biosignal Architecture Multimodal approaches have proven essential for robust stress monitoring [18]. While various models have been explored for imputing or recovering missing data [19]–[21], mitigating noise in biosignals [22], and addressing binary modularity [23], to the best of our knowledge, no existing literature has evaluated a modular architecture specifically for stress monitoring. Furthermore, there is no research investigating whether cross-attention mechanisms applied to multimodal biosignals could serve as a modular alternative to traditional data imputation or decision-fusion techniques.

This study advances the field of multimodal biosignal classification by implementing a truly modular architecture. The proposed system leverages a pre-trained, generalized model that learns from individual sensors and is subsequently fine-tuned for modularity. This reduces the dependency on specific sensory inputs within a multimodal framework, allowing the model to adapt seamlessly between scenarios where only a single modality is available (e.g. PPG wrist-wearable) and those involving multiple modalities (e.g., sEMG headphones and PPG wrist-wearable).

Novel Attention Score Caching Mechanism for Early-Onset Stress Detection To maximize the model's ability to contextualize historical context of the biosignals, whilst ensuring the latency of inference is minimized, a novel sliding attention score caching mechanism with bidirectional encoder-decoder attention is proposed. The attention and caching mechanism is

inspired by the Transformer architecture, whereby its decoder employs key-value (KV) caching for its autoregressive text generation. Unlike traditional text generation, where outputs must maintain cohesion, our approach applies KV caching within the model's encoder, allowing for dynamic re-evaluation of past signals in light of new inputs. Both previous key-value projections and attention scores are stored, typically discarded in standard Transformer models. This method enables the model to maintain a richer contextual memory without requiring a larger window size, thereby reducing computational complexity and improving inference speed.

1.4 Report Structure

The structure of this paper is as follows. We begin by summarising the background of stress detection research in Chapter 1, both through first principle analysis of biomarkers, and the standpoint of experimental designs for stress detection, devices used and existing datasets. To address the issues that arise with detecting stress, existing state-of-the-art machine learning architectures focusing on modularity, data fusion, and personalization techniques will also be discussed. Following from this, in Chapter 3, we detail the experimental design which outlines the protocol and data collection techniques employed to create the MUSED dataset.

Chapter 4, shows the implementation of our novel modular multimodal architecture proposed in this paper, alongside the overview of the machine learning pipeline required for this. To ensure the most clear comparison, we extensively test the proposed model through the widely researched WESAD and UBFC-Phys datasets [16], [17] in addition to the MUSED dataset curated. Extensive experiments into the new modalities and the new architecture are presented and analysed. Chapter 5 consolidates these results and discusses the wider implications, and proposes limitations and future work. Lastly, Chapter 6 concludes the paper, summarizing our approach and findings and linking it back to the aims of our research.

1.5 Published Material

The MUSED dataset, designed to be similarly structured to the WESAD dataset for easy comparison, will soon be publicly available after its accompanying paper is published.

The complete codebase for the analysis of MUSED aims to be as clear and easy to use as possible. It is open-source to combat the machine learning research that claims to have achieved a

certain accuracy without transparent details of the exact features, hyperparameters or architectures used. This can be accessed at: https://github.com/WillPowellUk/Modular-Multimodal-Stress-Detector and will be referenced throughout Chapter 4.

1.6 GPU Energy Consumption

To encourage researchers to declare how much compute was used during their research projects and raise awareness of the impacts machine learning research has on the climate, we have disclosed my GPU energy usage. For the extensive training and testing of this project, it was approximately 120 kWh. If a user was to perform LOSO-CV inference once using the proposed machine learning pipeline in this paper, using the MUSED dataset, the GPU energy usage would be approximately 0.25kWh.

1

2

Background & Related Work

Contents

Currently, there is no automatic, continuous, and unobtrusive method for early stress detection. Given the multimodal nature of stress, any effective approach will need to integrate multiple types of measurements [24]. The nuances of the stress response are highlighted below, along with the commonly used sensors for its detection and the protocols used for inducing it. Then, machine learning approaches for early detection of stress onset are discussed.

2.1 Biomarkers and Biosignals of Stress

As described in Rochette et al. [25], stress was first defined by Hans Selye as "the nonspecific response of the body to any demand made upon it" [26] and has since had varying definitions [27]. Selye's clear emphasis on the biological aspect, neglects the psychological side involved. To ensure a concrete definition for the remainder of this paper, a definition associated with a well-researched psychological stress protocol will be used, see Section 2.2.

2

The biological response to stress is controlled by the hypothalamic-pituitary-adrenal, Hypothalamic-Pituitary-Adrenal (HPA), axis and sympathetic–adrenal–medullary axis of the Sympathetic Nervous System (SNS), and are responsible for regulating the release of hormones such as cortisol, adrenaline, norepinephrine, and alpha-amylase [28]. A stressor activates the brain, triggering the hypothalamus to simultaneously initiate the rapid SNS pathway, releasing catecholamines (adrenaline and noradrenaline) for immediate physiological changes, and the slower HPA axis releases cortisol to sustain prolonged alertness and energy [29], [30]. These hormones increase heart and breathing rates, suppress the immune system, and prioritize energy towards vital physiological functions, preparing the body to survive a perceived threat [31], [32]. Conversely, the activation of the parasympathetic nervous system pathway through stimulation of the vagus nerve indicates a mental relaxed state [33]. It is evident that cortisol would be the gold standard biomarker for stress measurement, but despite recent advancements in its sensing [34], it is not feasible for mass commercial adoption due to the requirement of disposable patches or saliva sampling.

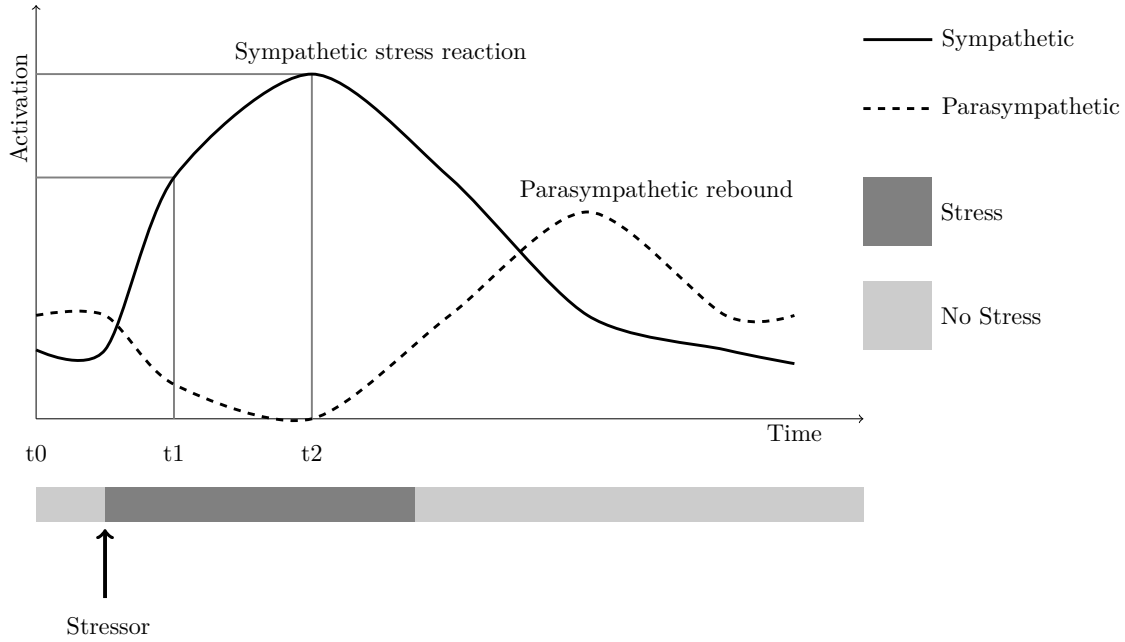


Figure 2.1: The non-instantaneous sympathetic reaction to a mild, acute stressor demonstrates the requirement for large windowed samples to be taken to ensure the correct prediction is made. The parasympathetic rebound, a phenomenon caused by a period of suppression in the PNS which elicits an exaggerated response [35], also showcases the need for a contextualized understanding of the signal to make an informed prediction, again requiring a large window size.

The majority of studies examining stress induce strong, acute mental stress which cause a large sympathetic activation and parasympathetic withdrawal and is thus easily detectable if a large window-size is used, i.e. from t0 to t2 in Figure 2.1. However, for early-onset detection, i.e. detecting at “t1”, the onset of the sympathetic stress reaction must be identified after a much

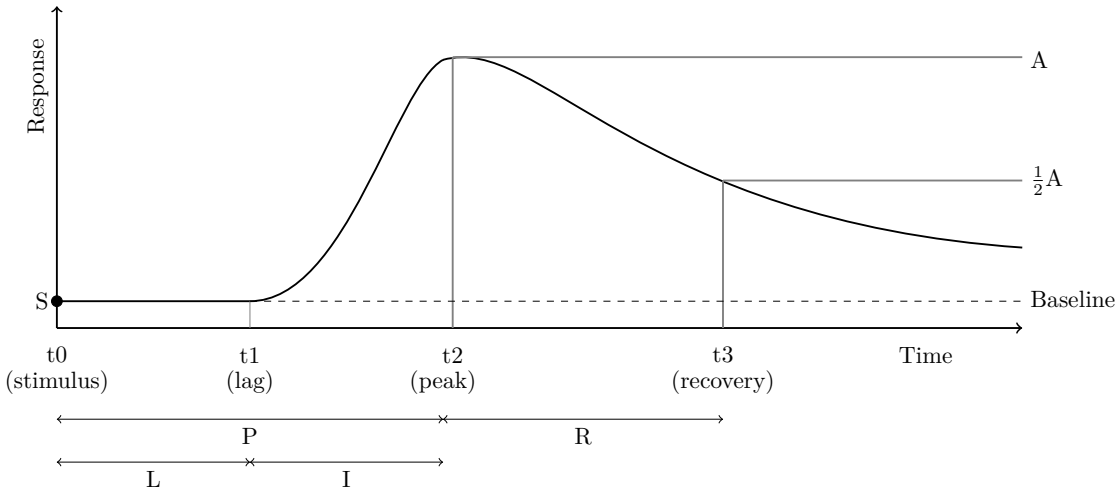
smaller threshold. This therefore yields two difficult problems: firstly, a high temporal resolution of the signals and its features are required to predict this initial response, and secondly, more subtle changes within and between features must be identified. Thus, utilising several biosignals through widely adopted wearables will enable a more reliable measurement for early-onset stress detection.

Wrist-based wearables are the most commonly adopted means of stress detection, typically utilising PPG, otherwise known as BVP, to detect through optics the volumetric changes in blood which fluctuate in correspondence to the heart rate. PPG, whilst enabling the detection of metrics such as Heart Rate Variability (HRV) that correlate with sympathetic nervous system responses shown in the meta-analysis by Castaldo et al. [36], it cannot be solely relied on. Thus, other methods are gaining popularity, such as ECG, a signal generated by the electrical activity of the heart, and less commonly EDA, a sensor that measures the electrical conductance of the skin, which varies with its moisture level. Blood pressure measures the systolic pressure (pressure during the heart's contraction) and diastolic pressure (pressure during the heart's relaxation). ECG, whilst offering a richer, more precise feature set including EDR, it requires two electrodes to be placed either side of the heart, typically using either hands, like the use in an Apple Watch [37], or via chest strap ECG such as Polar H10 [38].

In addition to the non-invasive sensors mentioned, EEG, which measures the electrical activity of the brain has been shown to detect stress. EEG Alpha waves, of frequency 8–12.99 Hz [39], are associated with a relaxed state and thus conversely, a decrease in alpha wave activity has shown to be a strong biomarker of stress [40]. To detect alpha-waves, an EEG cap is typically placed on the prefrontal cortex, a modality that is not suitable for daily use [41]. Since the mastoid offers an excellent site for the measurement of alpha waves, placed in headphone form, this device could contribute well to a multimodal stress detection system.

fNIRS is a non-invasive neuroimaging technique that uses near-infrared light to measure brain activity by monitoring changes in blood oxygenation and flow [42]. It is favoured over other modalities for its high spatial resolution, portability, and superior tolerance to movement artefacts compared to EEG, which is often compromised by movement in everyday settings [42], [43]. However, fNIRS has low temporal resolution, typically between seconds and minutes, which limits its ability to track rapid neural dynamics [44]. By emitting near-infrared light absorbed by hemoglobin and myoglobin, fNIRS measures capillary and intracellular oxygenation levels [45], governed by the modified Beer-Lambert Law [46]. As noted by Cui et al. [45], the delayed hemodynamic response to neural activation poses challenges for early-onset detection using machine learning, as models must account for this latency to accurately predict neural activity.

To bring the most certainty to automated stress detection, several multimodal biosensors mentioned above are typically used in experimental environments, with a clear summary of recent papers summarised by Naegelin et al. [18]. However, the biosignals mentioned, though activated by the SNS and HPA pathways, incur different temporal delays during their activation, saturation and recovery to an acute stressor. For instance, a stressor activating the sweat glands does not align well with the cardiovascular response, as the activation of the sweat glands via the release and binding of the neurotransmitter acetylcholine, following SNS activation, is inherently slower than the rapid electrical and muscular response of the heart [47].



S Stimulus	**Cortisol** (Salivary Conc.)	**EDA** (Skin Cond.)	**FNIRS** (Oxy/De-oxy Conc.)
L Time lag	L: 15-30 sec [48]	L: 1-3 sec [50]	L: 2-6 sec [45]
I Increase	I: 15-45 sec [49]	I: 0.5-3.5 sec [50]	I: 2-5 sec [51]
P Peak	P: 30-45 sec [49]	P: 1.5-6.5 sec [50]	P: 5-8 sec [51]
R Recovery	R: 1+ hour(s) [49]	R: 1-10 sec [50]	R: 10-15 sec [51]
A Amplitude	A: 40-100 nmol/L [49]	A: 0.01-5µS [50]	A: 0.1-0.4 oxy-Hb (%) [45]

Figure 2.2: Typical response curve to an acute stressor of cortisol, EDA and fNIRS demonstrates the large temporal misalignment which must be accounted for during data fusion.

As shown in Figure 2.2, whilst the activation, saturation, and recovery of cortisol, EDA and fNIRS do not align well amongst each other, this is further exaggerated by more immediate electrical biosignal responses, namely ECG PPG, sEMG and EEG. Despite this, these biosignals are frequently used together, and their temporal alignment is often overlooked and instead rely on large window sizes. Ensuring proper temporal alignment across modalities is essential to facilitate synergy between the biosignals and prevent conflicting data, as further explored in Section 2.3.

To add further complexity, inter-participant differences in stress response are extremely high and unpredictable. A few anecdotal examples from various studies are depicted in Table 2.1 to demonstrate this. For example, during laboratory conditions, 10% of subjects did not show any signs of sweating due to an external stimulus, termed “non-responders” [52]. However, this has

had some deliberation with another study which did not identify this but rather hypo-responsive periods during the stressor [53]. Similarly, the detection of tensing in the neck and shoulders through sEMG monitoring of the upper-trapezius would yield a strong biomarker in some users and a weak biomarker in others. This variability was demonstrated by a large interquartile range and significant differences in the mean and median frequencies of short periods of muscular relaxation during the stress response [54].

Sensor	Biomarker Of Stress	Inter-individual Examples
Salivary / Blood / Urine Sampling	Cortisol	52% of Participants Showed Habituation, 16% Showed Response Sensitization [55]
EDA	Electrodermal Activity	10% non-responders [52]
ECG/PPG/fNIRS	Heart Rate	Heart-rate recovery indicates that individuals who ruminate have slower heart-rate recovery after stress [56]
	Heart Rate Variability	Females have a decreased mean RR interval and SDNN compared to males [57]
Respiration (RESP) / ECG/ fNIRS	Respiration Rate	Standard deviation of 2.35 breaths per minute during rest, whilst stressor elicits a mean increase of 0.20 during a stressor [58].
fNIRS / EEG	Prefrontal Activity	Cognitive heavy tasks increases individual differences in activity [59]

Table 2.1: Examples of inter-individual differences in the biosignals involved in the detection of stress demonstrate the reductive qualities of a generalized model and the importance of considering individual biomarkers through personalized approaches.

Examples of inter-individual differences in stress response were highlighted in a review article on psychoneuroendocrinology by Kudielka et al. [60], which detailed a wide range of intra- and inter-individual variability in salivary cortisol response samples. The inter-individual factors included age, sex, mental disorders, genetic factors, early life experiences (pre- and postnatal stress), social hierarchy. Even cortisol, which is considered closest to the ground truth due to its direct connection to the HPA axis, showed significant variance both at different times of measurement within the same participant and between participants. Whilst intra-individual differences cannot be adequately addressed using standardized stress protocol studies, this further emphasizes the need for personalized detection mechanisms to ensure inevitable individual differences are addressed.

2.2 Experimental Designs to Induce Stress

Defining, measuring and inducing stress are very heated areas of research with many conflicting arguments, however standardized tests allow for consistent definitions and protocols to induce and measure stress. The cold water pressor test is a classical method of inducing stress, due to its ability to reliably trigger a series of physiological responses involving sympathetic activation through HPA axis activation [61] including a reliable adrenaline response. However, to combat the poor realism of this protocol, the TSST has been favoured in recent times, consisting of a preparation period, a free

speech, and mental arithmetic task in front of an audience [62]. A meta-analysis of laboratory-based acute psychological stress tests, found that tasks combining social-evaluative threat with uncontrollability elicit the strongest HPA axis response. Due to its combination of the social-evaluative public speaking task and a mental arithmetic task inducing high mental load, the TSST shows to elicit stress reliably [63] and has resulted in numerous datasets utilising this protocol, most notably WESAD [16], Ulm-TSST [64] and UBFC-Phys.

The VR equivalent of the TSST protocol, VR-TSST, initially introduced by Kelly et al. [65], was proposed to combat the inevitable variance induced by individual behaviour, which poses challenges for replication and standardization. Additionally, utilising a VR environment removes the requirement of substantial resources, like the two separate rooms, various props and the personnel to act as judges.

Kelly et al. [65] noted an elevation in salivary cortisol following stress induction via VR, though these stress responses were less intense compared to those elicited by the in-vivo TSST. However, since then, Zimmer et al. conducted a study comparing the real-world TSST with a VR simulation that closely mimicked the in-vivo environment [66]. This research found that physiological stress markers were nearly identical across both the real and VR versions of the TSST. Moreover, it highlighted that reactions to stress in the VR-TSST were more significant and pronounced than those observed in the original study. Several other studies have further shown that VR-TSST reliably induces acute stress [67]–[70], with two publicly available environments available [71], [72]. For this project, the VR-TSST environment is well suited due to the limited availability of personnel to act as the judges, whilst still ensuring scientific vigour and comparison with other studies and databases.

To ascertain the ground truth of the subject during the TSST, subjective tests are used to evaluate the participants' mental state. Most commonly, the State-Trait Anxiety Inventory (STAI) [73], which is a self-report scale that assesses separate dimensions of "state" and "trait" anxiety [74] and Short Stress State Questionnaire (SSSQ) which identifies the type of stress (worry, engagement or distress) [75]. Positive and Negative Affect Schedule (PANAS) [76] and Self-Assessment Manikin (SAM) [77] can also be used to evaluate affection based on the likert scale. Of course, these subjective questionnaires, though burdensome for everyday monitoring, are typically considered the ground truth for stress labelling, and can be further used to validate the effectiveness of a protocol on eliciting stress in addition to adding self-reported labelled samples for active learning via Ecological Momentary Assessment (EMA), like in Tazarv et al. [78].

Whilst some studies utilise perceived measurement of stress, enabling the consideration of subjective and personalised stress scores, this leads to a loose definition and thus offers poor reproducibility and comparisons. Thus perceived stress will not be considered as the stress label in this paper, however self-evaluations are provided in the dataset to follow, see Chapter 3.

2.3 Machine Learning Methods

The initial research for stress detection using biosignals dates to the 1990s in the form of speech classification and hidden Markov models [79]. Through increased availability and affordability of biosensing equipment, in the early-2000s, EDA and ECG were then commonly utilised with feature extraction algorithms and either traditional machine learning methods such as SVM, LDA or simple neural networks such as MLPs, GRNN and ANFIS [80]–[83]. These studies showcased varying levels of performance and were highly dependent on the nuances of the dataset. With the popularity of wrist wearables emerging, the 2010s focused considerable research into heart rate metrics, particularly HRV, with an increase in popularity for random forest as an effective multimodal decision model [84], [85]. Most recently, deep learning based methods have thrived not only as classification methods but also in the automatic feature extraction process and have significantly improved over conventional models on popular datasets [86], [87]. Automatic feature extraction has been particularly useful for biosignals that have a larger information context. Although their complexity prevents these models from being commonly used in commercial devices, simpler approaches typically rely on manual feature extraction of key biomarkers and simplistic, interpretable machine learning models, though the specific methods are rarely disclosed.

Since this research is focused on a pragmatic way of detecting the early onset of stress, the architecture used to detect stress must be lightweight, allowing compute to either occur on edge devices or on the cloud with minimal latencies and computational costs, whilst ensuring enough contextual information is extracted from the signals for accurate classification. Firstly, the current state-of-the-art traditional and lightweight deep learning methods will be discussed, followed by a detailed literature review on architectures for data fusion and personalization. These aspects are critical to not undermine the per-subject variance in the response to stress via their biomarkers and subsequently biosignals, as explained in Section 2.1.

2.3.1 Contextual Learning

We believe that the accuracies showcased by stress datasets, such as WESAD, demonstrate that even with a relatively small selection of manually extracted features from pre-processed biosignals, the reliability of stress is sufficient without computationally-expensive automated feature extraction. However, we also recognise that a large window size is usually required to contextualize the biosignals captured, thus reducing the model's ability to detect a stress response immediately and reliably. Thus, we propose that sequence-to-sequence models, which interpret the historical context of the signal, are more desirable for this task as they are able to capture greater subtleties in the temporal aspects of the stress response, increasing the robustness of early onset stress detection.

Recalling the slow, yet dynamic nature of a sympathetic stress reaction and the parasympathetic rebound illustrated in Figure 2.2, obtaining a broader context of the biosignals with a high temporal resolution would significantly improve the reliability of detection. Throughout our literature review, we identify that a large window size is typically used (see Tables 5.5 and 5.6) during feature extraction to serve as a low frequency buffer to reduce noise, capturing more reliable information and aligning predictive features more effectively. However, this comes with three key disadvantages. Firstly, the latency increases proportionally to the window size, since the pipeline must capture large samples before inferring, and thus early-onset detection is not possible. Secondly, it results in a higher computational cost to extract features, and thus a larger throughput of the system from these longer samples. Both these aspects, are exemplified by the study by Jaén-Vargas et al. [88] who examined the trade-offs of window size for real-time human activity recognition for deep learning models. Lastly, large window sizes function as a low-pass filter when manual feature extraction is employed, removing temporal nuances from the biosignal which are often crucial for determining the current state of the autonomic nervous system.

Sequence-to-sequence Learning

To prevent the neglection of temporal features, which are key to gathering context of the sympathetic nervous system, sequence-to-sequence (Seq2Seq) models are a valid approach used widely in biosignal classification [89]–[91]. In Seq2Seq models, commonly implemented using recurrent neural networks (RNNs) or variants such as Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), the architecture consists of two main components: an encoder and a decoder. The encoder processes the input sequence and transforms it into a fixed-length

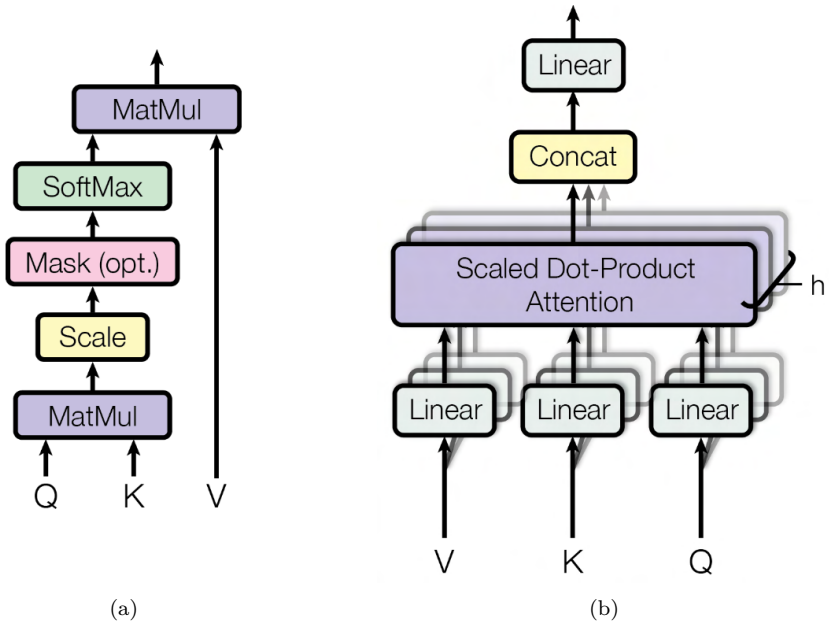
representation known as the context vector or hidden state. This context vector captures the essential features of the input sequence, including temporal dependencies. Subsequently, the decoder takes this context vector and generates an output sequence, leveraging the encoded information to produce temporally structured predictions. LSTMs and their bidirectional variant, BiLSTMs, which allow for awareness of both past and future temporal data, have demonstrated the ability to capture rich temporal features in the domain of stress detection, as shown by recent studies whose results surpass traditional machine learning methods [89], [90]. However, the methods discussed do not fully address the issue of information loss over long sequences, nor do they fully leverage parallel computing. The LSTMs hidden state is dependent on sequential processing, preventing the ability to process timestamps in parallel.

In a recent study, Xia et al. [91] introduced hierarchical temporal attention to the LSTM model, aimed at capturing early onset stress through EEG signals. They devised a two branch system, whereby one branch focuses on capturing intraslice temporal dependencies and one for interslice dependencies. Applied to both branches, the hierarchical temporal attention allows the model to prioritize the most critical time periods or data points using attention weights, aggregating discriminate temporal features in the sequence into an attention score. Its ablation study demonstrated that employing both intra-slice inter-slice slice branches with this attention were most effective, reiterating the need for maximising the contextualization of a signal. Whilst the study achieved significant cost savings by electrode configuration and the parallelization of the two branches, the BiLSTM is still bound to sequential processing on a per-branch basis.

Attention-Based Learning

Transformers have gained popularity in recent time for effectively contextualizing information for natural language processing tasks, particularly for text generation, which was exemplified in the pivotal paper, “Attention is all you Need” [92]. The critical finding of this paper is that the exclusive use of self-attention mechanisms was effective at capturing long-term dependencies without the reliance on recurrent or convolutional layers. The implications of this was that the computational cost drastically reduced, due to its parallelizable nature, allowing for more efficient training on larger datasets and inference. It also means that the architecture enabled processing of significantly longer sequences at a given computational latency, effectively increasing the model’s context window. After the release of this paper in 2017, the exclusive use of the attention mechanism has propagated to other applications, including biosignals.

Self-attention blocks have shown their performance in sequence-to-sequence tasks in the biosignal domain [23], [93]. The attention mechanism is similar to how the human brain's neurological system highlights important sounds while filtering out background noise. In deep learning, this mechanism enables neural networks to assign different levels of importance to various parts of the input sequence, greatly enhancing their ability to capture key information. The Transformer model, as proposed by Vaswani et al., comprises three types of attention: encoder attention, encoder-decoder attention, and masked decoder attention [92]. These components are crucial in sequence-to-sequence tasks like text generation. Encoder attention occurs within the encoder, where all tokens in the input sequence attend to each other through self-attention, capturing dependencies and building comprehensive token representations. Encoder-decoder attention, or cross-attention, occurs between the encoder's output and the decoder's input sequence, allowing the decoder to focus on relevant parts of the encoded input. Here, keys and values are derived from the encoder's output, while queries come from the decoder's current state. Masked decoder attention is employed in the decoder during generation, ensuring that each token is generated based on previously generated tokens. In addition to the attention mechanisms discussed, another notable approach in the literature is co-attention, which jointly learns attention weights for two input sequences, often originating from different modalities or sources. Unlike cross-attention, which asymmetrically combines two sequences by treating one as the query and the other as the key/value, co-attention focuses on the symmetric interaction and alignment of features across modalities.



(a) (b)

Figure 2.3: a) Scaled Dot-Product Attention. b) Multi-Head Attention consists of several attention layers running in parallel [92].

Self-attention blocks undergo scaled dot-product attention, shown in Figure (a). Given an input sequence of length n with each element represented by a d-dimensional embedding, let X be the matrix of these embeddings, where $X \in R^{n \times d}$. The input matrix X is linearly transformed into three different matrices: Queries Q, Keys K, and Values V which are obtained by the linear projection of learned weight matrices W_Q, W_K, and W_V:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V \tag{2.1}$$

Here, W_Q, W_K, and W_V are weight matrices of dimensions $d \times d_k$, $d \times d_k$, and $d \times d_v$, respectively. The attention scores are computed using the dot product of the query and key matrices.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.2}$$

Where QK^T computes the dot products between all queries and keys, resulting in a matrix of attention scores. These scores are scaled by $\sqrt{d_k}$ to prevent large values that could lead to vanishing gradients. The softmax function normalizes these scores to probabilities. The resulting matrix after applying the attention scores to the values V gives the final output of the self-attention mechanism.

To enhance the model's ability to focus on different parts of the input, Transformers employ multi-head attention, which is a concatenation of multiple self-attention layers. This involves splitting the input into multiple sets, and performing attention for each set independently with different learned linear projections, as shown in Figure (b). The outputs of these multiple attention heads are concatenated and linearly transformed to produce the final output. For h attention heads this can be written as:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_h)W_O \tag{2.3}$$

where each attention head, h, is computed as $\text{head}_h = \text{Attention}(XW_{Q_h}, XW_{K_h}, XW_{V_h})$.

The complexity, for which governs the efficiency of a multi-head attention, is $O(h \cdot T^2 \cdot d)$, where h is the number of attention heads, T is the sequence length, and d is the dimensionality of the embeddings. However, it is worth noting that various optimization techniques have been developed to address the quadratic complexity of self-attention, such as sparse attention mechanisms, linear attention variants or sliding window attention [94]–[96].

Applying attention to non-generation tasks such as biosignal classification, which do not require autoregressive properties, the encoder self-attention has been utilised in recent literature [23],

[93], [97]. This is accomplished first by either passing pre-processed raw data into the model [91], manually extracted features [23] or automatically extracted features through deep learning approaches [93]. Firstly, the signal is segmented into temporal slices and subsequently temporal embeddings, then all slices in a batch then attend to one another in a bidirectional manner. The attention mechanism highlights the most salient features by putting more emphasis, by the use of attention weights, on certain features in a feature set. However, as demonstrated in Chapter 4, this approach is inefficient for real-time classification tasks due to the high computational load required to attend to each token in a batch. Therefore, we propose a more efficient method to achieve this, which is detailed in Section 4.4.2. Attention-based methods are highly effective not only for aligning and processing biosignals efficiently but also in multimodal settings, and we continue our discussion on this in Sections 2.3.2 and 2.3.3.

2.3.2 Multimodal Ensemble Learning and Data Fusion Techniques

As discussed in Section 2.1, monitoring multiple biosignals as part of a multimodal system shows to be promising, since it brings more certainty to the current state of the sympathetic nervous state. However, as depicted in Figure 2.2, combining biosignals from various sources introduces the challenge of temporal misalignment between the biomarkers. Addressing this issue, along with minimizing noise interference, is critical to enabling sensors to effectively integrate data from different sources and predict symbiotically more accurate predictions than those derived from individual signals.

To achieve higher accuracy in detecting stress levels, the fusion of multimodal physiological signals is a promising approach. Current methods can be broadly categorized into two types: (i) early fusion, which allows for joint representation learning but often struggles when sensor data is noisy, and (ii) late fusion, which can improve estimates over individual branch classifiers [16], [98], but do not leverage features in the joint feature space of the different modalities and thus have limited temporal alignment. In late fusion, alignment is typically achieved through a large window size in the aim to synchronize the physiological signals recorded from different devices, see Tables 5.5 and 5.6. However, other techniques that have been introduced to combat this rudimentary approach.

One effective method for ensuring temporal alignment in late fusion models is the SELF-CARE fusion approach, which enhances traditional late fusion techniques by employing a context-aware, selective fusion strategy [22]. This strategy dynamically adapts the sensor fusion process in real

time, based on data from wrist-worn sensors. While traditional late fusion methods often rely on a simple voting mechanism across multiple classifiers to determine the final output, where each classifier is specialised for its particular set of input data [99], SELF-CARE advances beyond these methods by incorporating a Kalman filter-based approach. This approach integrates temporal dynamics to effectively combine the outputs from various classifiers according to the current motion context, as detected by accelerometer data. Unlike other learned late fusion methods, such as those discussed in [100], the Kalman filter mechanism in SELF-CARE allows the system to selectively fuse sensor data that is most relevant to the current context. This reduces the impact of noise and improves classification accuracy, leading to state-of-the-art performance on the WESAD dataset for wrist-based modalities. However, it should be noted that this high performance might partly be attributed to the method's sensitivity to motion artefacts, which could be seen as a form of data-snooping in the context of interview-based protocols. Furthermore, the modularity of the SELF-CARE architecture has not yet been thoroughly explored.

Another method explored as a fusion technique are cross-attention mechanisms. For multimodal attention based models, which are built up the encoder-decoder attention mechanism, the literature is mainly focused on combining computer vision and NLP tasks such as semantic segmentation [101], visual-question answering [102] and image captioning [103]. Nonetheless, these methods have filtered down to other domains, including emotion recognition and stress detection. For example, Bhatti et al. [97] proposed a cross-modal attention mechanism to enhance multimodal learning for stress detection using EDA and ECG data. They argued that cross-attention could be applied to share intermediate feature representations between convolutional neural network layers; this module would serve to align and fuse signals together in order to obtain joint representations which would be interpreted and condensed by subsequent layers. Bhatti et al. demonstrated that adding their proposed mechanisms in multiple different locations within the multimodal pipeline was more advantageous than fusing features at a single point. Additionally, bidirectional cross-attention, resulted in a significant improvement in accuracy to when either individual modalities were used or single-directional attention on the WESAD dataset.

Progressing from this, Zhang et al. [93], proposed the Bi-directional Cross and Self Attention (BCSA) module which incorporates self-modal attention after each cross-modal attention, or cross-attention block. This self-modal attention is argued to be essential for minimizing noise and redundancy introduced by multimodal data and the cross-modal attention process. Each BCSA module contains one cross-modal attention and one self-modal attention mechanism, working together to refine and correlate information across different modalities. In the cross-modal attention

mechanism, the primary goal is to align and correlate information between different modalities. This process involves bidirectional information flow between the modalities, ensuring that the features from one modality inform and are informed by the features of the other modality, irrespective of their positioning in the sequence. Specifically, one modality's features, such as BVP, are used as query terms, while the other modality's features, such as EDA, serve as key and value terms to calculate the cross-modal attention. For the self-modal attention, the aim is to highlight the most important features and reducing the impact of noise both in the features of the signal and the noise introduced during the cross-attention process, i.e. through contradicting features. Multiple pairs of these self- and cross-modal attention blocks were proposed, and it was demonstrated that these improved performance to a certain extent, helping the model to lean different levels of abstract information, though increasing the latency of the system.

2.3.3 Modularity

Whilst widespread adoption of smart wearable devices may be limited by current signal processing techniques [36], [104], computational costs [22] and individual differences [60], a challenge that has not had sufficient consideration in the literature is modularity; the discrepancy between the sensor modalities or devices used by consumers after deployment and those used during model training is a clear but neglected issue.

Although Seq-to-Seq methods have demonstrated effectiveness in stress classification, they typically require consistent sensor modalities during both the training and inference phases. Alternatively, when one or more sensors are absent, architectures typically rely on either imputing missing data [20] or apply decision-level fusion on the active branches to make an aggregated prediction [100]. More sophisticated approaches to achieving true modularity include techniques such as knowledge distillation. For instance, the "More2Less" framework, proposed by Yang et al. [105], employs a cooperative learning strategy to enhance modularity by selectively transferring knowledge from stronger to weaker sensor modalities using an adaptive regularizer to ensure positive transfer; each sensor modality is assigned a dedicated neural network trained not only on its own data but also through knowledge distillation, where weaker networks learn from stronger ones by minimizing representation differences in the embedding space. Compared to standard LSTM models, which vary in performance based on the sensor used, the More2Less framework leverages multimodal data during training and maintains strong performance even with fewer sensors during testing.

As discussed in Section 2.3.1, cross-attention enables the exploitation of complementary information between modalities and avoids redundant, misleading modality features by enabling information transfer during the hidden layers. Cross-attention and a similar technique, co-attention, has been applied in the verbal and non-verbal semantics domain for multimodal modularity. For example, Shi et al. [106] proposes a co-attention network which places more emphasis on the role of text as the central modality and introduces a specialized gating mechanism to handle modality gaps in video and audio. Each modality (text, video, and audio) is initially processed separately to extract relevant features. Then, the Cross-modal Modulation Module based on Co-attention (CMMC) inside this network, enables these modalities to interact and influence each other's representations. The Cascading Modular Multimodal Cross-attention Network (CMMCN) proposed by Han et al. [107] introduces a novel approach for rumour detection in the multimodal domain by leveraging cross-attention mechanisms in a modular fashion. Similar to Shi et al.'s model, CMMCN processes text and image modalities separately, extracting features using pre-trained models (BERT for text, ResNet50 for images) before employing a cross-attention module to allow these modalities to interact. Crucially, the cascading modular structure of the network, where cross-attention units are stacked to progressively refine multimodal interactions at different layers, ensures that deeper, more informative joint representations of textual and visual data are formed. Modularity is enhanced by the independent feature extractors and attention mechanisms, which facilitate precise multimodal fusion, comparable to the gating mechanism used by Shi et al. to address modality gaps. Both papers are comparable to how cross-attention was employed in the papers by Zhang et al. [93] and Bhatti et al. [97] where intermediate joint representations were formed, but with the addition of modularity achieved through a gating mechanism. This begs the question: can cross-attention of multimodal biosignals be utilized in a modular fashion as a replacement to data imputation or decision-fusion techniques? To the best of our knowledge, this question remains unanswered and has thus motivated our research.

2.3.4 Personalization

As discussed in depth in Section 2.1, to address the widely documented issue of inter-participant variability in stress monitoring, showcased in Table 2.1, personalization techniques must be employed to obtain inter-individual nuances that generalized models cannot capture.

Personalized classification systems have shown to be particularly effective in self-attention based models. Liu et al. developed a model to enhance recommender systems by generating personalized

representations of users and items, termed "personalized attention" [108]. The paper found that by using query vectors on an individual or item level was beneficial for learning a precise representation of users and items that could not be captured by a generalized model. In the field of stress monitoring, Yu et al. developed a Modality Fusion Network that employed both a generalized attention model and a personalized attention model to detect stress [23]. The study collected data from 41 participants over a span of 8 days using self-reported stress labelling. It was found that the model trained on personalized data only resulted in a slight performance increase in comparison to the generalized model, and only required one self-attention network - a 70% reduction from the four self-attention networks required to train the generalized model. This study clearly showcases the redundancy of parameters in the generalized model when it used on an individual; the extra parameters are required to learn nuanced features specific to the participants it was trained on. This study also strongly dismisses the need for using a combination of transfer learning and fine-tuning to adapt a one-size-fits-all model to a personalized model trained on each subject's data, an idea proposed in a previous study [109].

Personalization in model design is essential to mitigate the large parameter requirements associated with training generalized models. By employing personalized attention and fusion mechanisms, the model no longer needs to account for inter-participant variability in stress responses. Instead, it can focus on individualized biomarker patterns, both at the single-biomarker and multi-biomarker levels. This approach builds on the work of Yu et al., who applied personalization at the sensor level to achieve a 70% reduction in the model size [23]. However, they did not address the inter-participant variability between modalities. In contrast, our model leverages personalized cross-attention to align multiple modalities more precisely to the unique characteristics of each user. Personalized self-attention further emphasizes the intra-modality features that are most relevant to the individual. Finally, by incorporating personalization into the late fusion stage, the model ensures that predictions are weighted toward the modality providing the most reliable indicators of stress for each user.

3

Experimental Design

Contents

3.0.1 Criterion and Bias Statement

The study protocol has received a favourable opinion from Imperial's Science, Engineering and Technology Research Ethics Committee (SETREC). The inclusion and exclusion criterion were made to firstly ensure the safety of the participants (i.e. those with mental disorders or prone to anxiety) and secondly, to prevent participants from showing stress responses that were abnormal (i.e. through pregnancy, smoking, drug use, intensive exercise prior to the study – all of which have been shown to drastically alter a normal stress response). The aim of this study is to invoke a typical stress response for each participant so that it may act as a benchmark in testing the efficacy of new biosensing modalities on stress detection. After these findings have been made, a further study will be conducted, aiming at collecting data that is more generalizable to real-world scenarios and thus aid in a stress detection system that can tolerate stress responses made by participants that suffer from anxiety, have recently exercised, or are pregnant, for example. This will then serve to construct as a consumer ready device. However, for this study, we hope this justifies the participant recruitment made. The consent form and accompanying information sheet given to participants are available and can be given upon request.

Figure 3.1: The study protocol where the red boxes refer to filling in self-report questionnaires, blue represents baseline (no-stress) conditions and orange represents the stress conditions. The stress conditions intend to induce a different type of stress: anticipatory stress, social stress, mental stress.

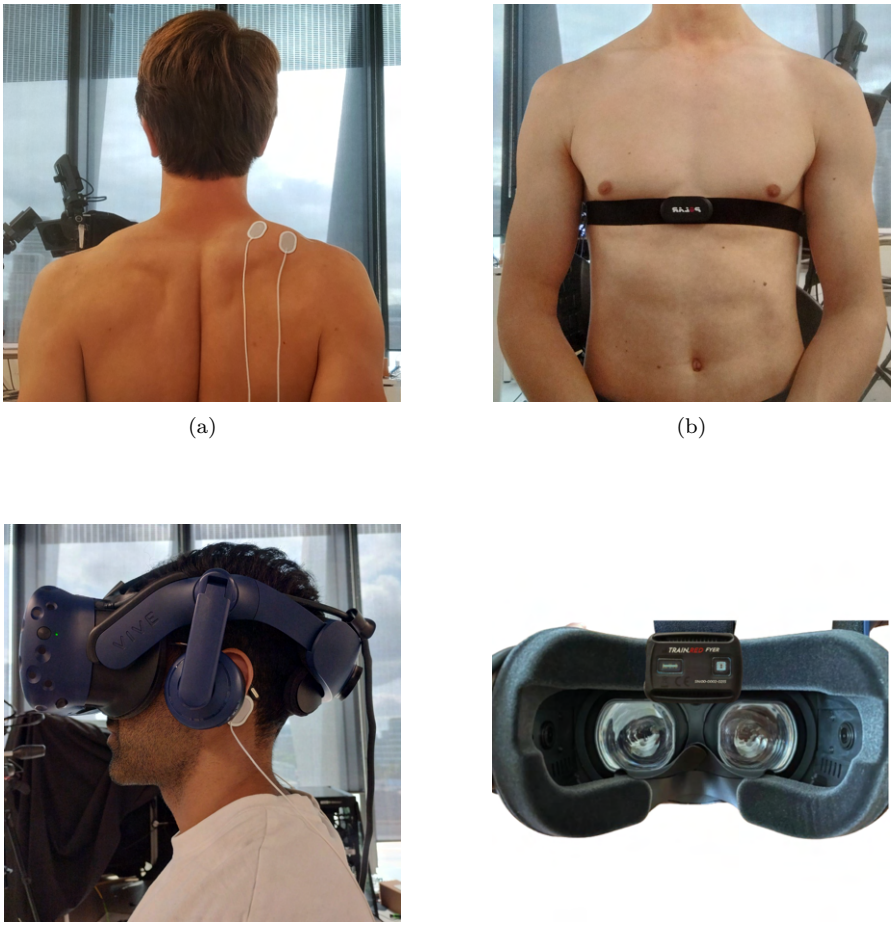
3.0.2 Protocol

The protocol is taken from the paper by Linning et al. describing the creation and evaluation of the Open TSST-VR discussed in Section 2.2 [72]. In addition to the original protocol, participants were provided with the following supplemental instructions before the experiment began:

- Participants were informed to complete their baseline tasks (sitting and standing) facing the radiator; some participants during piloting thought that the previous setup facing the desk and microphone was slightly stressful. During anticipation they were instructed to turn to face the desk.
- Participants were informed of the overall structure of the protocol and were told that the interview would be interactive, i.e to expect to talk and respond according to the interview questions.
- The remainder of the instructions were presented to them in audio and subtitle form between tasks. Since the initial study was in German, we devised a translation into English, and utilized a convincing AI voice-over - a more relaxed voice for the instructions, and a more domineering voice for the interview. The audio were then edited with Audacity, matching the previous acoustics such as, reverb, room-size configuration, reverb etc. The English audio files will soon be contributed to the supplementary materials of the Open TSST-VR paper [72].

Here's the rewritten version without any formatting:

Similar to the WESAD protocol, the short-form of PANAS was used, along with the STAI immediately before and after the protocol, to assess the effectiveness of the stressor. The SSSQ was also conducted after the protocol, to distinguish the type of stress induced. The ground truth, cortisol, was not measured, however the effectiveness of the protocol was validated in the original study introducing the Open TSST-VR, as discussed in Section 2.2.



(a) (b)

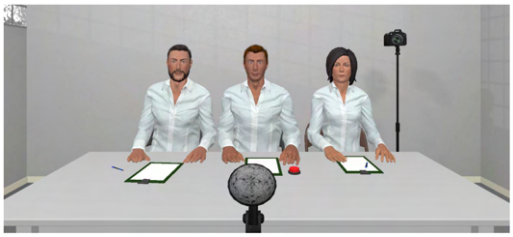
(c) (d)

Figure 3.2: a) Bipolar electrode for sEMG placed on right upper trapezius of subject. b) Polar H10 ECG chest strap positioned at the level of the sternum. c) Bipolar electrodes for sEMG placed on left mastoid of subject - other electrode is placed on right mastoid). d) fNIRS device is located at the centre of the pre-frontal cortex and is held in place by the VR headset.

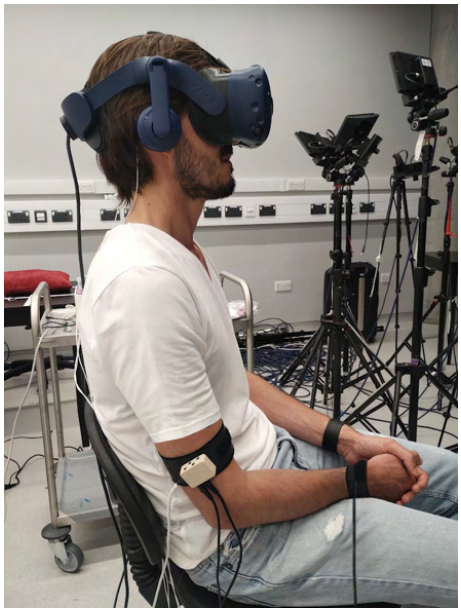
The Empatica E4, seen on the left wrist of the participant in Figure 3.3 (c), is a medical-grade, wrist-based wearable and its robustness has been exemplified in thousands of studies. It was fitted onto the non-dominant hand of the participant in the hope of reducing the number of motion artefacts. It gathers data at varying frequencies, including 32 Hz for 3-axis accelerometer readings, 64 Hz for photoplethysmography, and 4 Hz for both electrodermal activity and temperature.

Each subject was seated, and after wiping the area with an alcohol wipe, a pair of pre-gelled Ag/AgCl bipolar electrodes were placed on the right shoulder, midway between the acromion and vertebra C7, following the guidelines of SENIAM. The electrodes were spaced 40 mm apart, with a wet reference strap placed on the dominant wrist. Another pair of electrodes were placed on the mastoid, as shown in Figure apparatus-setup mastoid. The two channels were connected to OT Bioelettronica's Quattrocento EMG recording device to capture sEMG at 2048 Hz with 16-bit

(a)



(b)



(c)

(d)

Figure 3.3: a) experimenter's control panel for the TSST1 (job interview) to control the actions of the judges using Open TSST-VR [72]. b) the corresponding environment with virtual interviewers that can be triggered remotely by the experimenter. c) subject during baseline sitting with wet reference strap on right wrist, Empatica E4 on left wrist. d) subject during anticipation task, where the experimenter screen can be seen, mirroring the subject's view.

resolution. Shown in Figure 1.2, to ensure that any potential EEG activity at the mastoid, in addition to sEMG signals, is captured, a high-pass filter of 0.7 Hz and a low-pass filter of 500 Hz was applied.

The electrodes of the Polar H10 chest wearable were slightly wetted and fitted to the participant, see Figure 3.2 (b), streaming ECG data at 130 Hz with 14-bit precision via Bluetooth. Whilst this device operates at a lower sampling frequency to the one employed in the WESAD study, inter-beat intervals are sampled at a much larger 1 kHz range. This allows for precision measurement of the R-waves of the signal and can be deemed as the ground truth in relation to the other cardiovascular and hemodynamic measurements signals that derive heart-rate, namely BVP, ECG

and fNIRS. The chest device also records accelerometer data at 25 Hz with 16-bit precision, to monitor the body accelerometer data, which may help distinguish between body movements and wrist movements captured by the Empatica E4.

The Train.Red FYER Muscle Oxygen device, which utilizes the MyndSens streaming platform from SilverLine Research, was employed to record single channel fNIRS on the prefrontal cortex. The device was placed on the forehead, fixed by the padding of the HTC Vive Pro VR headset which was tightened to a comfortable but secure amount, see Figure vr-myndsens. The data was streamed over Bluetooth with LabStreamingLayer (LSL) at 10Hz. The device provides three signals: O2Hb (oxygenated hemoglobin), HHb (deoxygenated hemoglobin), and brain oxygenation level.

Synchronization across these devices was achieved using a double-tap signal pattern, and the synchronized data is provided in pre-processed files, ensuring coherence across different sensor modalities for stress detection analysis. Nonetheless, it was identified that in several subjects' data, the modalities were either unable to be synchronized or became unsynchronized due to imprecision in the stated sampling rate and timestamps. Whilst synchronization analysis will thus be difficult for some recordings, the signals from the modalities still render useful.

3.0.3 Data Collection and Protocol Validation

Phase	Type	Duration	Total Duration
Baseline Sit	Non-Stress	1hr 31 mins (19.56%)	3hrs 05 mins (39.72%)
Baseline Stand	Non-Stress	1hr 34 mins (20.15%)	
Anticipation	Stress	1hr 38 mins (21.03%)	
Interview	Stress	1hr 33 mins (19.99%)	4hrs 41 mins (60.28%)
Arithmetic	Stress	1hr 30 mins (19.26%)	

Table 3.1: Overview of data collected during the study, including the duration of each phase and corresponding class imbalance.

A total of 18 subjects (10 males, 8 females) ranging from age 22 to 41 with a mean age of 26.33 and standard deviation of 4.77 took part in the study. The summary of the data collection is presented in Table 3.1 and is compared against the well-established WESAD and UBFC-Phys datasets in Table 3.2.

Subject details and qualitative notes were taken from each participant, and are available in the subject's data folder readme file, see Section A.3 for details. Technical issues which affected data reliability, such as poor signal quality in PPG, sEMG, and fNIRS measurements and device battery failures, were also noted.

	WESAD			UBFC-Phys			MUSED		
Number of subjects	15 (12 males, 3 females)			56* (46 females, 10 males)			18 (10 males, 8 females)		
Age of subjects	27.5 ± 2.4			21.8 ± 3.11			26.33 ± 4.77		
Modalities + Sampling Rate	ACC (700 Hz), ECG (700 Hz) EMG (700 Hz), TEMP (700 Hz), RESP (700 Hz),W_ACC (32 Hz), BVP (64 Hz), W_EDA (4 Hz) EDA (700 Hz), W_TEMP (4 Hz)			BVP (64 Hz), W_EDA (4 Hz)			ACC(20 Hz), ECG (130 Hz), Mastoid EMG (2048 Hz), Upper Trap EMG (2048 Hz), W_ACC (32 Hz), BVP (64 Hz), W_TEMP (4 Hz), fNIRS (10Hz) W_EDA (4 Hz)		
Protocol	**Task**	**Duration (mins)**	**Class Labels (%)**	**Task**	**Duration (mins)**	**Class Labels (%)**	**Task**	**Duration (mins)**	**Class Labels (%)**
	Sitting/standing (baseline)	20	54.79	Sitting (baseline)	3	33.33	Sitting/Standing (baseline)	10	39.71
	Speech and arithmetic (stress)	10	27.40	Speech (social stress)	3	33.33	Preparation (anticipatory stress)	5	21.03
	Funny video task (amusement)	6.5	17.80	Arithmetic (cognitive stress)	3	33.33	Speech (social stress)	5	19.99
							Arithmetic (cognitive stress)	5	19.26

Table 3.2: Comparison of the three stress datasets used in this study. * The UBFC-Phys dataset comprises data from 56 subjects; however, this study utilizes data from only 15 subjects. The remaining subjects were excluded based on the criteria outlined by the dataset authors in their publication.

The qualitative findings reveal a range of emotional and physiological responses during interviews and arithmetic tasks. Several participants reported stress or nervousness, especially when the tasks were unfamiliar or challenging, such as struggling with arithmetic or language barriers. Others appeared mildly stressed or unfazed, indicating variability in anxiety levels (especially among those who rated low on being "scared of interviews". Behavioural observations included frequent hand gestures, possibly contributing to motion artefacts in signals, and signs of flustered or embarrassed behaviour during arithmetic tasks. These findings underscore the role of individual differences and task familiarity in both stress responses and signal quality.

	PANAS		STAI
	Positive	Negative	
Pre-protocol (baseline)	31.4 ± 6.09	12.9 ± 2.47	12.9 ± 3.74
Post-protocol (stress)	28.6 ± 10.1	19.1 ± 7.24	14.2 ± 2.02

Table 3.3: Evaluation of the questionnaires completed pre- and post-protocol indicate that the subjects' state was influenced by the stress condition.

Engagement	Distress	Worry
4.28 ± 0.787	3.61 ± 0.850	3.56 ± 1.10

Table 3.4: Evaluation of the post protocol SSSQ indicates that subjects were strongly engaged during the interview and arithmetic tasks and had a comparable amount of moderate worry and distress.

The complete T-test statistical results for the STAI and PANAS questionnaires can be viewed in Tables B.1 and B.2 respectively. The statistical analysis reveals that the stress protocol significantly influenced participants' emotional and anxiety states. Pre- and post-protocol comparisons on the PANAS and STAI scales, shown in Table 3.3 showed a decrease in positive affect (31.4 ± 6.09 to 28.6 ± 10.1) and a substantial increase in negative affect (12.9 ± 2.47 to 19.1 ± 7.24). The STAI results indicated heightened anxiety, with significant increases in feelings of jitteriness (T = -4.68, P = 0.000), worry (T = -2.96, P = 0.009), and nervousness (T = -3.00, P = 0.008), alongside a significant decrease in relaxation (T = 2.53, P = 0.022). From the SSSQ, we observe a strong post-protocol engagement (4.28 ± 0.787), although participants experienced moderate levels of distress (3.61 ± 0.850) and worry (3.56 ± 1.10). These findings demonstrate that the stress protocol effectively induced emotional discomfort and heightened anxiety, while maintaining cognitive engagement. However, these findings contradict the high inter-participant variability found in the qualitative findings of the readme's, where some subjects reported little stress, particularly during the anticipation task but also during the interview and arithmetic tasks.

3

4 Methodology

Contents

4.1 Proposed Model Overview

The model architecture, depicted in Figure 4.1, showcases a modular multimodal model that leverages three strengths utilised in the decoder of the Transformer for early-onset stress detection. These are the early fusion and multimodal temporal alignment capabilities of cross-attention blocks, noise attenuation and feature importance selection through self-attention blocks, and Key-Value (KV) and attention-score caching of previous temporal slices to act as a memory state for sequence-to-sequence classification.

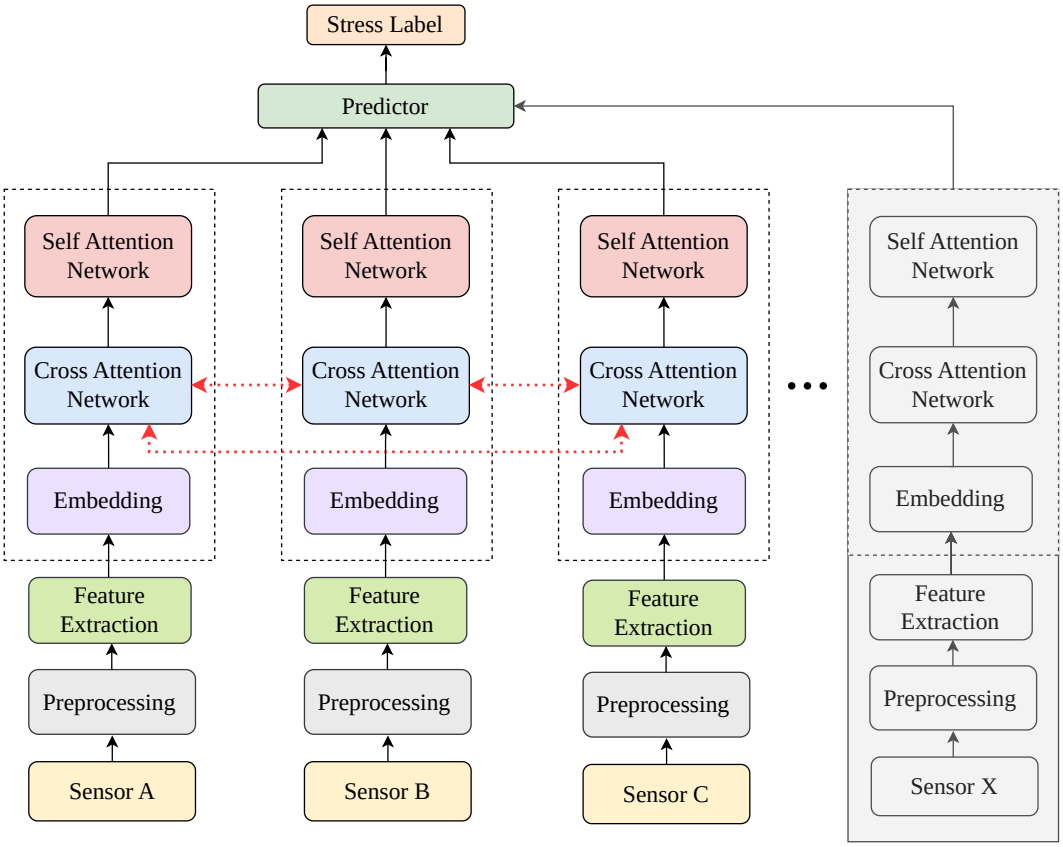


Figure 4.1: The proposed model architecture leverages cross-attention networks to temporally align manually extracted features and incorporates early fusion between each unique pair of sensors. Self-attention networks are used to reduce noise and redundant information, emphasize key details, and extract latent space feature representations of stress. The predictor then fuses these features from the separate branches into a single prediction using an ensemble learning approach. In unimodal mode, skip connections are used to bypass the cross-attention blocks; in other modes, the model adapts to any combination of modalities inputted by only inferring the cross-attention blocks that attend between the pairs of active sensors. The model also employs a novel caching mechanism in the attention networks, detailed in Section 4.4.2, which enables stress detection to minimise window size, and thus compute, without a loss of temporal context.

The ML pipeline consists of: a) signal pre-processing b) feature extraction c) model training and validation. Each process is explained below in detail, which utilize the source code developed to conduct the necessary steps for training and inference of the signals. Each step of the pipeline is shown in high-level on the Demo Notebook which uses the WESAD dataset; the same process was applied to the other datasets.

4.2 Signal Preprocessing

The raw multimodal signal is first pre-processed applied on a per-modal basis, shown in Table 4.1 and the source code can be viewed here.

Modality	Filter
Wrist ACC	FIR filter, length = 64, cut-off frequency = 0.4 Hz
Wrist BVP	Butterworth band-pass filter, order = 3, cut-off frequencies = 0.7 Hz and 3.7 Hz
Wrist EDA	Butterworth lowpass filter, order = 6, cut-off frequency = 1 Hz
Wrist TEMP	Savitzky–Golay filter, window size = 11, order = 3
Chest ACC	Savitzky–Golay filter, window size = 31, order = 5
Chest ECG	Savitzky–Golay filter, window size = 11, order = 3
Chest ECG	Butterworth band-pass filter, order = 3, cut-off frequencies = 0.7 Hz and 3.7 Hz
Chest EDA	Savitzky–Golay filter, window size = 11, order = 3
Chest EDA	Butterworth lowpass filter, order = 2, cut-off frequency = 5 Hz
Chest EMG	Savitzky–Golay filter, window size = 11, order = 3
Chest EMG	Butterworth lowpass filter, order = 3, cut-off frequency = 0.5 Hz
Chest RESP	Savitzky–Golay filter, window size = 11, order = 3
Chest RESP	Butterworth band-pass filter, order = 3, cut-off frequencies = 0.1 Hz and 0.35 Hz
Chest TEMP	Savitzky–Golay filter, window size = 11, order = 3
fNIRS	Butterworth low-pass, order = 2, cut-off frequency = 0.1 Hz
fNIRS derived HR	Butterworth band-pass filter, cutoff frequencies = 1 Hz and 1.9 Hz

Table 4.1: Filters applied to the various modalities used in the datasets examined in this study.

The filter cut-off frequencies were selected to reduce the likelihood of misclassifications of the components of the signal due to noise. For example, the BVP signal is processed by a Butterworth band-pass filter with cut-off frequencies f1 = 0.7 Hz and f2 = 3.7 Hz, which limits the heart-rate down to which takes into account the heart rate at rest, approximately 40 BPM and a high heart rate due to exercise scenarios or tachycardia approximately 220 BPM [110]. For the EDA signal, only a lowpass filter of 1 Hz is required due to the slow response of the sweat glands, detailed in Figure 2.2.

The butterworth filter is well-used in biosignal processing due to its maximally flat frequency response in the passband and provides a smooth and monotonic transition from the passband to the stopband [111]. However, it has been argued that for ECG signals, the butterworth filter does not accurately provide useful information for the frequency morphology as it is poorly adjusted to analysing the frequency-varying structure of the ECG signal. This inadequacy arises from the formation of image signals, which can manifest as additional noise and baseline drift [112].

For the fNIRS signal we simply employ a 0.1Hz low-pass filter, to remove Mayer waves, cardiac pulsations and muscle artefacts. In the fNIRS HR derivation process, we adhere by the filtering techniques of Hakimi et al. [14]: a band-pass filter is applied to eliminate noise caused by physiological sources such as respiratory fluctuations and motion artefacts, ensuring that only the

heartbeat-related components are preserved. Additionally, a low-pass filter is used to refine the signal further, aiding in the accurate identification of key features, such as the R-peak in the ECG signal. These filtering steps are crucial to maintain the integrity of the heart rate estimation from both fNIRS and ECG signals.

The Savitzky–Golay filter, a type of finite impulse filter which aims to maintain a constant signal power and only tries to decrease the noise power, has been shown to be an effective equivalent to the butterworth filter for signals such as ECG and fNIRS [112], [113]. It aims at smoothing the signal to improve algorithmic extraction of key locations in the signal such as the systolic and diastolic peaks of the BVP signal [114] or the PQRST complex of the ECG [110]. The filter has shown to "preserve the peak of the ECG signal and maintain the original morphology" [112]. For signal processing of BVP however, a combination of the two mentioned filters have shown to work effectively together [22], [115].

Though the Savitzky–Golay is an effective digital filter, for the purposes of a commercial device, this would likely have to be substituted to an appropriate analogue equivalent such as an RC moving average filter to reduce latency and lower the computational requirements of the onboard CPU.

The multimodal signal of the dataset is divided into non-overlapping temporal slices of length n. For each modality m (where $m \in \{1, \ldots, M\}$), the signal $x_m(t)$ is segmented into temporal slices as follows:

$$x_m(t) \rightarrow \{x_m^{(1)}, x_m^{(2)}, \ldots, x_m^{(K)}\} \tag{4.1}$$

where:

$$x_m^{(k)} = x_m(t) \text{ for } t \in [(k-1)n, kn) \tag{4.2}$$

and $K = \lfloor \frac{T}{n} \rfloor$, with T being the total duration of the signal.

While alternative segmentation methods, such as the inter- and intra-slice segmentation proposed by Xia et al. [91], could be employed, the datasets used in this study do not offer sufficient task durations to explore the potential benefits of inter-slice segmentation. Therefore, to ensure a fair comparison with other methods, inter-slice segmentation was not considered in this investigation. Additionally, varying the segmentation lengths for each biosignal was not explored due to the increased complexity it would introduce. Instead, we adopted the approach used by Zhang et al. [93], where despite the slower nature of the EDA signal, the segmentation process was identical for the BVP signal. This alignment avoids the need for handling different window sizes, which would complicate the model architecture.

To increase the training size and add further variation to the sequences, data augmentation was utilised in the form of a sliding window. Overlapping is a type of synthetic oversampling whereby more data can be artificially generated, which has been shown to improve accuracy, particularly in deep learning models [116]. There is a risk however of data leakage, particularly when utilising LOSO-CV, and to ensure this is mitigated, meta-data was added to the signal to identify whether the segment was augmented. Thus, when validating the data on a subject, only the non-augmented segments will be utilized.

4.3 Feature Extraction and Selection

We believe that the accuracies showcased by stress datasets, such as WESAD, demonstrate that even with a relatively small selection of manually extracted features from pre-processed biosignals, the reliability of stress is sufficient without computationally-expensive automated feature extraction. We have also seen through our review of state-of-the-art methods, available in Section 5.3.3, that manually extracted features still outperform some automated feature extraction. Particularly, the underwhelming results of Zhang et al., in comparison with simpler methods, may likely be due to the feature extraction stage, not the methodology stage.

Thus, the temporal slices created undergo manual feature extraction to obtain localised features of the signal in the time, frequency and non-linear domains. For all sensors utilized in this study, except for fNIRS, the selected features for each sensor are taken from Schmidt et al. during their analysis of the WESAD dataset [16]. For fNIRS, the features are extracted by the methodology from Hakimi et al. [14]. These features are summarized in Table 4.2, and are explained in more detailed below.

For fNIRS, the feature extraction process is as follows. Heart rate features are extracted by detecting peaks in oxyhaemoglobin signal using the algorithm developed by Hakimi et al. [14] that considers both signal magnitude and first differences. Peaks are validated if they exceed a threshold and have a first difference at least 10% of the maximum observed. A 0.25-second window post-processing step removes non-peak relative maxima. Interbeat Intervals (IBI)s are calculated between corrected peaks, with outliers adjusted using a window-based method that considers local and global signal variations.

It is important to note that for some modalities, namely respiration and BVP, features could not be reliably extracted with such a small temporal slice (6 seconds). Thus, we derived these features

from their respective batch, rather than their independent temporal slice. This is a limitation to utilizing such a small window size and is discussed in Chapter 6. Additionally, it is also noted that for use of this pipeline in real-time applications, the signal captured via the sensor will be likely be pre-processed through analogue filtering and then a single individual temporal slice will be filled to a buffer before feature extraction and model inference, to reduce computational load.

4

4.4 Model Architecture

The model architecture was written using PyTorch, and its collection of modules are available on the models directory of the repository. The components of the architecture presented in Figure 4.1, will be explained in the following order: first, the modular version of the BCSA mechanism; next, the proposed Sliding Attention Score Caching mechanism along with its novel bidirectional encoder-decoder attention; and finally, the different predictors designed for the model's late-fusion stage.

4.4.1 Modular BCSA Mechanism

The model architecture utilises the same bidirectional cross and self attention block mechanism, termed BCSA, as in Zhang et al. [93]. However, appropriate modifications have to be made to make the architecture scalable to new modalities and modular to allow for different combinations of modalities to be used with the same pre-trained model.

The BCSA module, introduced in Section 2.3.1, is designed to enhance the integration of multimodal data by leveraging both cross-modal and self-modal attention mechanisms. Each BCSA module contains one cross-modal attention and one self-modal attention mechanism, working together to refine and correlate information across different modalities. The bidirectional cross-modal attention aims to temporally align features, ensuring that information flows from each modality to every other modality, excluding itself. Then the self-attention block reduces redundancy in the feature set and attenuates noise. The mathematical representation of this process, where modality m attends to modality n, is given by the following formula:

$$x_i'' = \text{cross-attention}(x_{mi}^{'}, \{x_{n1}^{'}, x_{n2}^{'}, \ldots, x_{nE}^{'}\}) \tag{4.3}$$

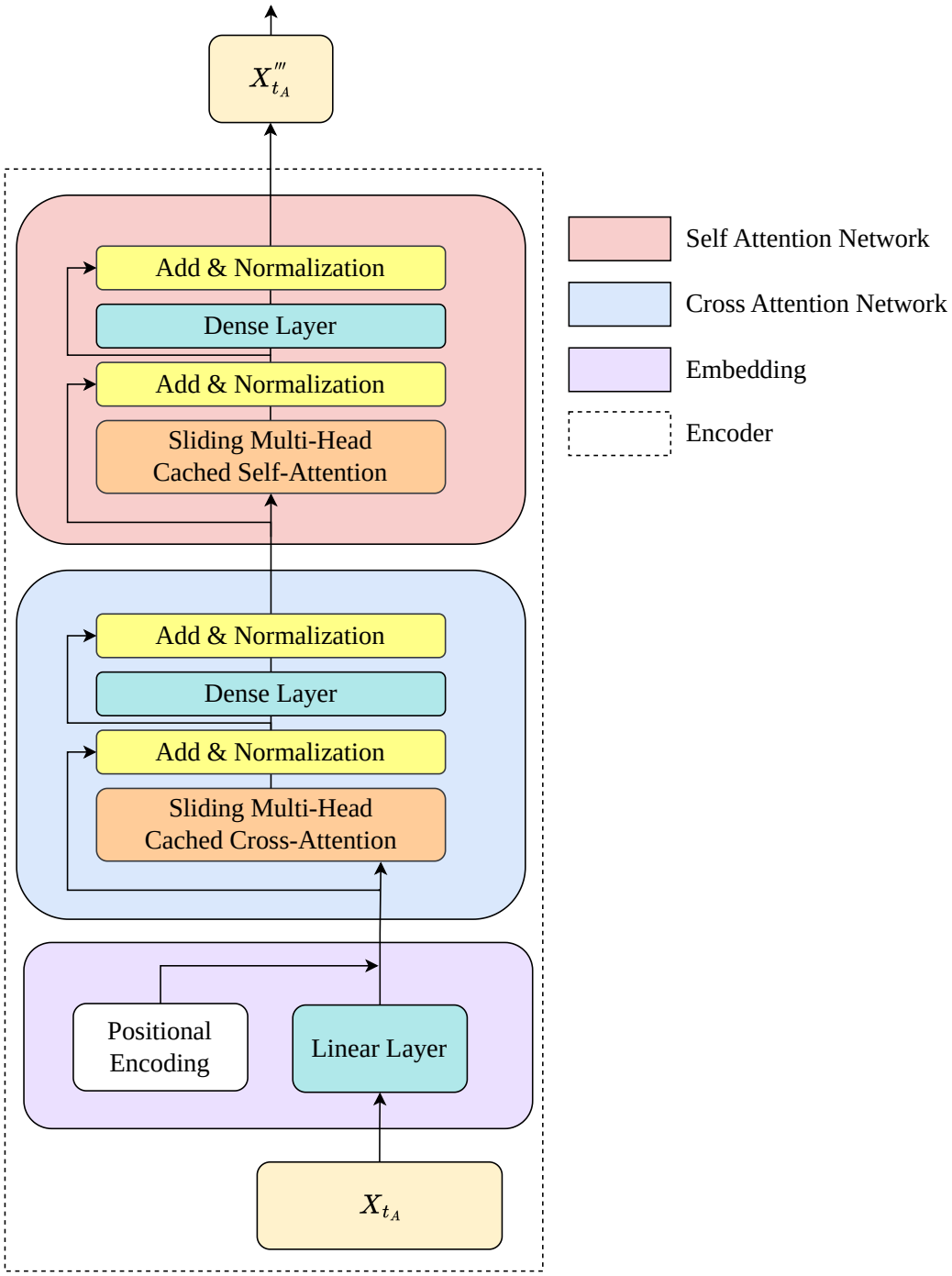


Figure 4.2: The encoder for one modality branch, Sensor A, depicting the embedding layer, to transform the sensors features X_t to a standard embedding dimension of $X_t \in \mathbb{R}^{B \times H \times L \times E}$, cross attention network to temporally align features from other modalities, improving context of the model, and self-attention networks to finalize the importance of the latent features.

Feature	Description
ACC	
$\mu_{ACC,i}, \sigma_{ACC,i} \quad i \in \{x, y, z, 3D\}$	Mean, STD for each axis separately and summed over all axes
$\left\|\int \mathbf{ACC}_i \, dt\right\| \quad i \in \{x, y, z, 3D\}$	Absolute integral for each/all axes
$f^{\text{peak}}_{ACC,j} \quad j \in \{x, y, z\}$	Peak frequency for each axis i
ECG and BVP	
μ_{HR}, σ_{HR}	Mean, STD of the HR
μ_{HRV}, σ_{HRV}	Mean, STD of the HRV
$NN50, pNN50$	# and percentage of HRV intervals differing more than 50 ms
$TINN$	Triangular interpolation index
rms_{HRV}	Root mean square of the HRV
$\sum^{f}_{x \in \{ULF,LF,HF,UHF\}}$	Energy in ultra low, low, high, and ultra high frequency component of the HRV
$f^{LF/HF}_{HRV}$	Ratio of LF and HF component $\sum_{x \in \{ULF,LF,HF,UHF\}} f_x$
$\sum$ the freq. components in ULF-HF	
$relf_x$	Relative power of freq. component
LF_{norm}, HF_{norm}	Normalised LF and HF component
EDA	
μ_{EDA}, σ_{EDA}	Mean, STD of the EDA signal
min_{EDA}, max_{EDA}	Min and max value
$\partial_{EDA}, range_{EDA}$	Slope and dynamic range
$\mu_{SCL}, \sigma_{SCL}, \sigma_{SCR}$	Mean, STD of the SCR/SCL
$corr(SCL, t)$	Correlation btw SCL and time
$\#SCR$	# identified SCR segments
$\sum Amp_{SCR}, \sum t_{SCR}$	$\sum$ SCR startle magnitudes and response durations
$\int scr$	Area under the identified SCRs
EMG	
μ_{EMG}, σ_{EMG}	Mean, STD of EMG signal
$range_{EMG}$	Dynamic range
$\left\|\int EMG \, dt\right\|$	Absolute integral
$\tilde{EMG}$	Median of the EMG signal
$P^{EMG}_{10}, P^{EMG}_{90}$	10th and 90th percentile
$f_{EMG}, \tilde{f}_{EMG}$	Mean, median and
f^{EMG}_{peak}	Peak frequency
$PSD(f_{EMG})$	Energy in seven bands
$\#peaks_{EMG}$	# peaks
$Amp_{EMG}, \sigma^{EMG}_{Amp}$	Mean, STD of peak amplitudes
$\sum Amp_{EMG}, \sum \tilde{Amp}_{EMG}$	$\sum$ and normalised $\sum$ of peak amplitudes
RESP and EDR	
$\mu_x, \sigma_x \quad x \in \{I, E\}$	Mean, STD of inhalation (I) and exhalation (E) duration
I/E	Inhalation/exhalation ratio
$range_{RESP}, vol_{insp}$	Stretch, Volume
$rate_{RESP}$	Breath rate
$\sum_{RESP}$	Respiration duration
TEMP	
$\mu_{TEMP}, \sigma_{TEMP}$	Mean, STD of the TEMP
min_{TEMP}, max_{TEMP}	Min, max TEMP
$range_{TEMP}$	Dynamic range
∂_{TEMP}	Slope
fNIRS	
$\mu_{fNIRS}, \sigma_{fNIRS}$	Mean, STD of the fNIRS signal
$S_{t,fNIRS}, K_{t,fNIRS}$	Skewness, Kurtosis of the fNIRS signal
$\mu_{f,fNIRS}, \sigma_{f,fNIRS}$	Mean frequency, STD of frequency for fNIRS signal
$S_{f,fNIRS}, K_{f,fNIRS}$	Skewness, Kurtosis of frequency for fNIRS signal
LF/HF_{fNIRS}	Ratio of low to high-frequency power in the fNIRS signal
$\mu_{FDHR}, \sigma_{FDHR}$	Mean, STD of the FDHR signal
$S_{t,FDHR}, K_{t,FDHR}$	Skewness, Kurtosis of the FDHR signal
$\mu_{f,FDHR}, \sigma_{f,FDHR}$	Mean frequency, STD of frequency for FDHR signal
$S_{f,FDHR}, K_{f,FDHR}$	Skewness, Kurtosis of frequency for FDHR signal
LF/HF_{FDHR}	Ratio of low to high-frequency power in the FDHR signal
rms_{FDHR}	Root mean square of the FDHR signal
IBI_{FDHR}	Inter-beat intervals derived from FDHR

Table 4.2: List of manually extracted features for all modalities which are included in the three datasets being investigated, taken from Schmidt et al. [16] and Hakimi et al. [14].

The resulting cross-modal attention aligns and fuses information from the different modalities, thus creating joint representations of the modalities. Given M modalities, the number of cross-attention blocks in required is:

$$\text{Number of Cross-Attention Blocks} = M \times (M - 1) \tag{4.4}$$

All possible modalities, and thus all cross-attention combinations are used to train the model. Once trained, the model is able to infer on any combination of modality in an automated manner by only inferring on the active branches. However, additional fine-tuning on common modalities are recommended to prevent a degradation in performance when the reliance on certain modality is removed, this will be discussed further in Section 4.5.2. The evaluation of modularity performance can be viewed in Chapter 5.

In addition to cross-modal attention, the BCSA module incorporates self-modal attention to refine the information within a single modality, and is crucial for reducing noise and redundancy, which is introduced by multimodal data and the cross-modal attention process [93]. The self-modal attention uses the output from the cross-modal attention as query terms, with all time slices within the same modality serving as key and value terms. This process enhances the significant information within the modality and diminishes noise. The mathematical representation of self-modal attention on modality m is as follows:

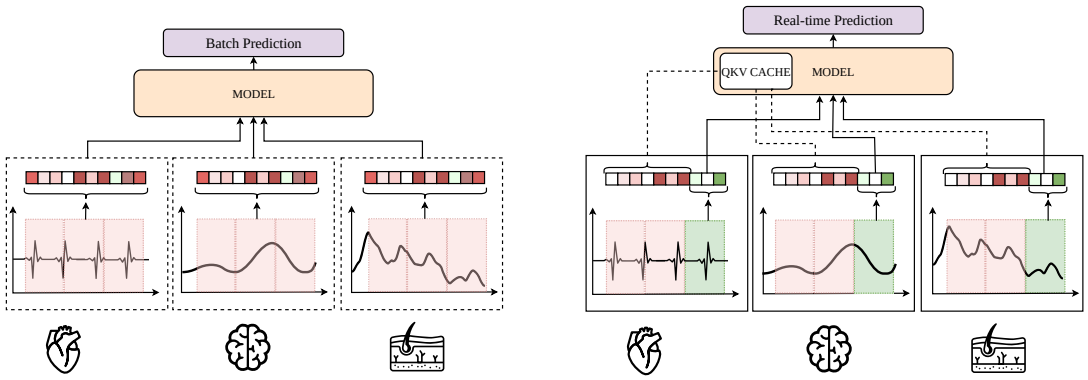
$$x'''_i = \text{self-attention}(x''_{mi}, \{x''_{m1}, x''_{m2}, \ldots, x''_{mE}\}) \tag{4.5}$$

In each BCSA module, self-modal attention follows cross-modal attention. This sequence is intentional, as the fusion process based on cross-modal attention can introduce additional noise and redundant information, which the self-modal attention then mitigates. The stacking of multiple BCSA modules allows the network to learn different levels of abstract information, thereby improving the performance and robustness of the model in integrating multimodal data. However, since this paper is focused on optimizing for latency, and the embedding representation is not large as a result of manual feature extraction, it was found that one single BCSA module was sufficient.

4.4.2 Sliding Attention Score Caching Mechanism

Further extending from the work of Zhang et al., a sliding mechanism is introduced into the encoder which holds memory in the form of projection and attention score cache, holding rich, latent space

features of previous timestamps in a sequence-to-sequence manner. Figure 4.3, conceptualizes this idea, showing how the model holds contextual memory which is held in cache, rather than a larger window size, reducing computational load during inference.



(a) (b)

Figure 4.3: a) Model training. b) Model inference. The model captures previous context by caching previous temporal slices. This allows the model to have a larger context of the signal, so that it can identify the sympathetic reaction and the parasympathetic rebound, illustrated in Figure 2.1, whilst maintaining a relatively small window size. The caching mechanism is disabled during training and the model learns to predict on all sequence lengths. During inference, only the current temporal slice, illustrated in green, undergoes projection and attention score computation. The previous temporal slices are stored as projections and attention scores in the proposed Sliding Attention Score Caching Mechanism.

The Sliding Attention Score Caching Mechanism, shown in Figure 4.5, is inspired by the decoder of the transformer used for text generation tasks, which employs KV caching with encoder-decoder attention to prevent the need for recomputing the previous tokens each time a new token is generated in its autoregressive process. The autoregressive nature of the decoder means it generates tokens sequentially, with each new token depending on all previously generated tokens. Without this tool, the k and v vectors for all previous tokens, given in Equation 2.1, would have to be re-computed (re-projected by W_K and W_V respectively) each time a token is generated. Instead, the previous projected keys and values are stored, only requiring the projection of the new token, before appending these new tokens key and value to their respective cache for scaled dot product attention.

This computational saving is drastic, especially when the context size and thus token length increases, therefore KV caching is employed in almost all commercial large language models. The KV-caching mechanism can be written formally as follows:

Let $K \in \mathbb{R}^{B\times H\times S\times D}$ be the existing key cache and $V \in \mathbb{R}^{B\times H\times S\times D}$ be the existing value cache. A new token is inputted, whereby $k_{new} \in \mathbb{R}^{B\times H\times 1\times D}$ is the new key already projected by

W_K, and likewise $v_{new} \in \mathbb{R}^{B\times H\times 1\times D}$ be the new value projected by W_V.

Where B is the batch size, H is the number of heads, S is the sequence length and D is the head dimension. The embedding dimension E is split into H parts, each of size $D = \frac{E}{H}$.

The updated key cache K' and value cache V' are given by $K' = \text{concat}(K, k_{new}, \text{axis} = 2)$ and $V' = \text{concat}(V, v_{new}, \text{axis} = 2)$ respectively. Where $\text{concat}(A, B, \text{axis} = 2)$ denotes concatenation of tensors A and B along the third dimension (sequence length).

Explicitly, the updated cache is $K' \in \mathbb{R}^{B\times H\times (S+1)\times D}$, and the updated value cache is $V' \in \mathbb{R}^{B\times H\times (S+1)\times D}$. As discussed in Section 2.3.1, the three types of attention implemented in the original Transformer are encoder self-attention, encoder-decoder attention and masked decoder attention. Since we do not require autoregressive prediction, the masked decoder attention is redundant for this application. For the encoder self-attention mechanism, seen in Figure 4.4, the computational complexity, including the projections, are presented below:

$$\mathcal{O}(L \cdot E^2 + L^2 \cdot E) \tag{4.6}$$

Where the target embedding length L is equal to the source embedding S; it is bound to the quadratic of the new embedding length and subsequently demands a large throughput. The encoder self-attention also experiences high latency because the batch of temporal slices must be fully populated before inference can occur. Likewise, the memory complexity also demands significant resources:

$$\mathcal{O}(L \cdot E) + \mathcal{O}(L^2) + \mathcal{O}(E^2) \tag{4.7}$$

Where $\mathcal{O}(L \cdot E)$ corresponds to projections and output storage, $\mathcal{O}(L^2)$ for attention scores and weights, and $\mathcal{O}(E^2)$ for parameters storage.

The encoder-decoder attention, as proposed in Vaswani et al. [92], refers to the attention mechanism used in the decoder to focus on relevant parts of the encoded input sequence while generating each output token. It operates by ensuring the query only attends to previous keys and values; there is no re-attendance to future keys and values. Thus, the new query tokens of target sequence length L, are multiplied by the keys (and subsequently values), which are of typically larger, of source sequence length S in the scaled dot-product attention. Typically, the target length is significantly larger than the source length, that is $S >> L$, and therefore the computational and memory costs are significantly reduced.

The decoder output is projected into the query, of complexity cost $\mathcal{O}(LE^2)$, whilst the encoder output is projected into the key and values, with a cost of $O(SE^2)$. Consequently, the final cost for the encoder-decoder mechanism is:

$$\mathcal{O}(LE^2 + SE^2 + LSE) \tag{4.8}$$

Since $S \gg L$, the dominant terms are $\mathcal{O}(SE^2)$ and $\mathcal{O}(LSE)$. Furthermore, the contribution of the past key and value projection terms, $\mathcal{O}(SE^2)$, can be removed in the decoding phase as they are retrieved from the KV cache, thus becoming:

$$\mathcal{O}(LE^2 + LSE) \tag{4.9}$$

The computational complexity is no longer bound to the square of the source length, resulting in significantly reduced computational resource requirements for real-time edge prediction. However, due to caching the memory complexity is impacted as follows:

$$\mathcal{O}(S \cdot E) + \mathcal{O}(L \cdot E) + \mathcal{O}(L \cdot S) + \mathcal{O}(E^2) \tag{4.10}$$

Given that $S \gg L$, the dominant terms in the memory complexity are the KV cache $\mathcal{O}(S \cdot E)$ and the attention scores and weights: $\mathcal{O}(L \cdot S)$.

KV cache is conventionally used in the decoder of a Transformer, i.e. during text generation, where each token is generated based on the tokens that have been generated before it. However, in biosignals, a previous token may represent one that was initially understood to have a certain meaning based off of its context, i.e. that biomarkers of stress were identified, but since the context has now been updated by the introduction of the latest token, it may take up a different meaning, i.e. this sequence should now be interpreted as noise since this current token suggests that no biomarkers of stress is present. This is comparable to BERT where the main goal was to understand contextual meaning between words and sentences through encoder self-attention, capturing context via past and future tokens during masked language modelling (MLM) rather than generating text in an autoregressive manner where only previous and present tokens are considered [117]. Putting it simply, the stress detection model should not only utilise the history of the signal to contextualize the current timestep, but also be able to re-evaluate the significance of past signals in light of this new information.

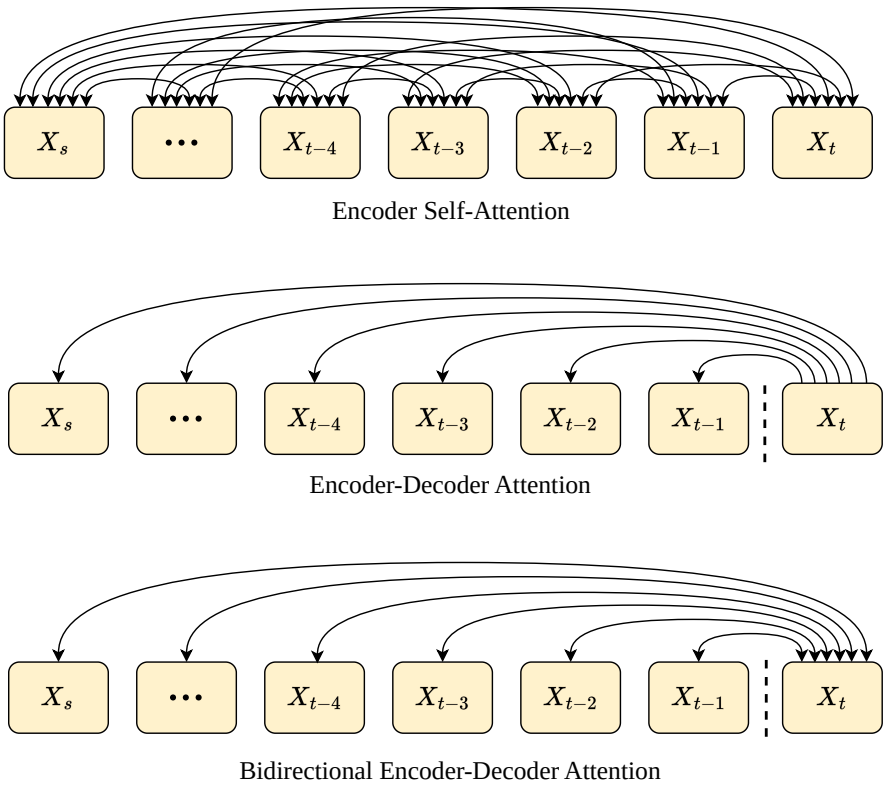


Figure 4.4: The three types of attention mechanisms suitable for biosignal classification are: encoder self-attention, encoder-decoder attention, and the proposed bidirectional encoder-decoder attention. Encoder self-attention, where all tokens attend to one another, is the most computationally expensive due to the inability to use KV-caching and the need for a complete batch of temporal slices before inference, leading to high latency. Encoder-decoder attention is more efficient as it computes attention scores on individual slices rather than batches and can leverage KV-caching. The bidirectional encoder-decoder attention shares these efficiencies and additionally allows for re-attending to previous embeddings alongside projection and attention score caching.

Thus, we propose a novel attention mechanism for biosignal classification called bidirectional encoder-decoder attention, see Figure 4.4, whereby the new input embedding attends to previous input embeddings and vice versa. With this, the caching mechanism not only stores the previous KV projections, but also the previous attention scores, which are usually discarded in Transformer based architectures.

Attention scores are not cached in Transformer architectures since the new token, taking the form of the query term, is only attending to previous cached key and value tokens; it must remain cohesive when generating text due to its autoregressive nature. For biosignal classification, it is often more desirable to re-evaluate previous predictions rather than ensuring cohesive or consistent predictions. In other words, the embeddings of the past signals can be re-evaluated in light of new information rather than being fixed as they are in generative tasks. This means that previously

generated attention scores now have a use and instead of being discarded at each new inference, they can be re-used during the re-evaluation process. The only requirement is that the attention scores must be shifted forward and the new attention scores written in red in Table 4.3 must be computed and appended.

	$\boldsymbol{Q_S}$	...	$\boldsymbol{Q_{t-4}}$	$\boldsymbol{Q_{t-3}}$	$\boldsymbol{Q_{t-2}}$	$\boldsymbol{Q_{t-1}}$	$\boldsymbol{Q_t}$
$\boldsymbol{K_S}$	$\frac{Q_S \cdot K_S}{\sqrt{d_k}}$	...	$\frac{Q_{t-4} \cdot K_S}{\sqrt{d_k}}$	$\frac{Q_{t-3} \cdot K_S}{\sqrt{d_k}}$	$\frac{Q_{t-2} \cdot K_S}{\sqrt{d_k}}$	$\frac{Q_{t-1} \cdot K_S}{\sqrt{d_k}}$	$\frac{Q_t \cdot K_S}{\sqrt{d_k}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\boldsymbol{K_{t-4}}$	$\frac{Q_S \cdot K_{t-4}}{\sqrt{d_k}}$	...	$\frac{Q_{t-4} \cdot K_{t-4}}{\sqrt{d_k}}$	$\frac{Q_{t-3} \cdot K_{t-4}}{\sqrt{d_k}}$	$\frac{Q_{t-2} \cdot K_{t-4}}{\sqrt{d_k}}$	$\frac{Q_{t-1} \cdot K_{t-4}}{\sqrt{d_k}}$	$\frac{Q_t \cdot K_{t-4}}{\sqrt{d_k}}$
$\boldsymbol{K_{t-3}}$	$\frac{Q_S \cdot K_{t-3}}{\sqrt{d_k}}$	...	$\frac{Q_{t-4} \cdot K_{t-3}}{\sqrt{d_k}}$	$\frac{Q_{t-3} \cdot K_{t-3}}{\sqrt{d_k}}$	$\frac{Q_{t-2} \cdot K_{t-3}}{\sqrt{d_k}}$	$\frac{Q_{t-1} \cdot K_{t-3}}{\sqrt{d_k}}$	$\frac{Q_t \cdot K_{t-3}}{\sqrt{d_k}}$
$\boldsymbol{K_{t-2}}$	$\frac{Q_S \cdot K_{t-2}}{\sqrt{d_k}}$	...	$\frac{Q_{t-4} \cdot K_{t-2}}{\sqrt{d_k}}$	$\frac{Q_{t-3} \cdot K_{t-2}}{\sqrt{d_k}}$	$\frac{Q_{t-2} \cdot K_{t-2}}{\sqrt{d_k}}$	$\frac{Q_{t-1} \cdot K_{t-2}}{\sqrt{d_k}}$	$\frac{Q_t \cdot K_{t-2}}{\sqrt{d_k}}$
$\boldsymbol{K_{t-1}}$	$\frac{Q_S \cdot K_{t-1}}{\sqrt{d_k}}$	...	$\frac{Q_{t-4} \cdot K_{t-1}}{\sqrt{d_k}}$	$\frac{Q_{t-3} \cdot K_{t-1}}{\sqrt{d_k}}$	$\frac{Q_{t-2} \cdot K_{t-1}}{\sqrt{d_k}}$	$\frac{Q_{t-1} \cdot K_{t-1}}{\sqrt{d_k}}$	$\frac{Q_t \cdot K_{t-1}}{\sqrt{d_k}}$
$\boldsymbol{K_t}$	$\frac{Q_S \cdot K_t}{\sqrt{d_k}}$	...	$\frac{Q_{t-4} \cdot K_t}{\sqrt{d_k}}$	$\frac{Q_{t-3} \cdot K_t}{\sqrt{d_k}}$	$\frac{Q_{t-2} \cdot K_t}{\sqrt{d_k}}$	$\frac{Q_{t-1} \cdot K_t}{\sqrt{d_k}}$	$\frac{Q_t \cdot K_t}{\sqrt{d_k}}$

Table 4.3: The scaled dot product grid, QK^T, of key-query pairs, termed the attention score, for a real-time sliding sequence. S denotes the maximum sequence length or context. Thus, reading left-to-right, K_s and Q_s represent the key and query pair of the oldest embedding or temporal slice in the sequence and K_t and Q_t are the key and query pair that correspond to the latest embedding. All components in this grid, except for the latest column and row marked in red, have been previously computed in earlier inferences and do not require re-computation due to the novel Sliding Cached Attention Mechanism.

In our bidirectional encoder-decoder attention, since the target sequence attends to the source sequence in a bidirectional manner, attendance is doubled to a new complexity of:

$$\mathcal{O}(LE^2 + 2LSE) \tag{4.11}$$

Though, the computational complexity still equates to:

$$\mathcal{O}(LE^2 + LSE) \tag{4.12}$$

However, it is important to note that the computation will be larger than encoder-decoder attention due to the added bidirectionality. The memory complexity will be the same as in Equation 4.10, however, with the linear increase of including query cache in addition to the key-value cache.

To obtain the improved computational performance discussed, the Scaled Dot-Product Attention must be modified to store previous attention scores before concatenating them with the new token's projected embedding. This is shown in Figure 4.5.

The sample embedding for a modality m is given by:

$$X'_{t_m} \in \mathbb{R}^{B\times H\times L\times D} \qquad \forall m \in \{A, B, C \ldots, M\} \tag{4.13}$$

During training, $B >> 1$ and $L = S$, where S represents the maximum sequence length. When the cache is used during inference, the batch size and the incoming temporal slice will both be one ($B = 1$ and $L = 1$), and it will utilise the cached attention score with its respective S embeddings.

As depicted in Figure 4.5, the new embedding is linearly projected by the weights W_Q, W_K, W_V. Then, for each head, h, the projected query and key, of dimensions $Q_t, K_t \in \mathbb{R}^{B\times H\times 1\times D}$, are multiplied and scaled using the cached projections of the previous and current embeddings to produce the attention scores for the new embedding. This represents the new bidirectional attendance, which is in the form of two vectors each of shape $S \times E$ representing the new column and row of the attention score matrix, written in red in Table 4.3. The remainder of the cache has already undergone these computations in previous iterations, and thus can be retrieved and immediately used in the scaled dot-product attention.

For each branch, the Sliding Cached Attention Mechanism will store the cache in four buffers - three tensors Q^{cache}, K^{cache} and V^{cache}, hold the projected embeddings of previous tokens and are required to compute the new attention score vectors, and one tensor, AS^{cache}, holds the attention scores of previous embeddings. Once the buffer has been saturated after S iterations, the cache will be in the form:

$$Q_m^{cache}, K_m^{cache}, V_m^{cache} \in \mathbb{R}^{B\times H\times S\times D} \qquad \forall m \in \{A, B, C \ldots, M\} \tag{4.14}$$

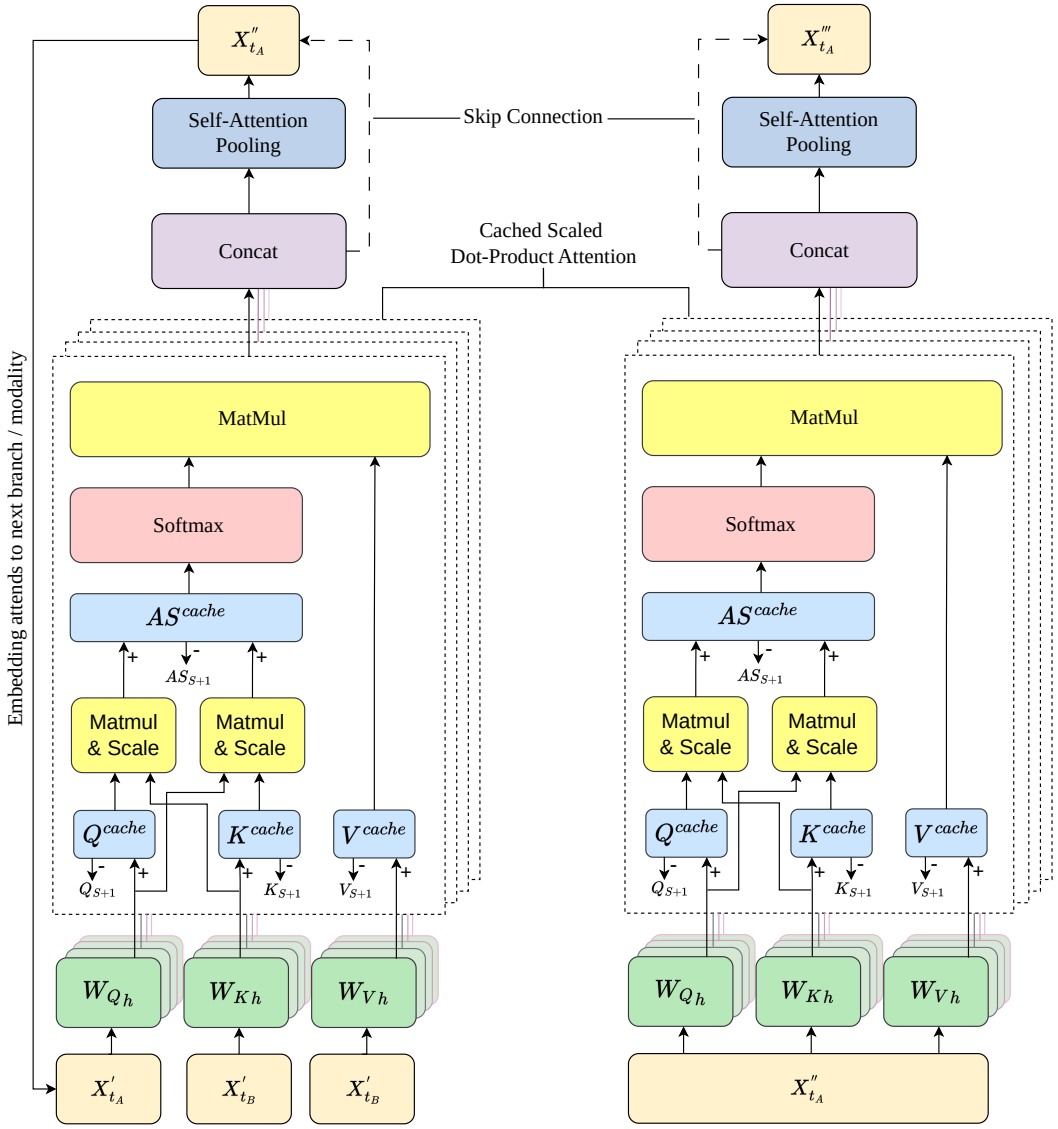


Figure 4.5: Sliding cached multi-head cross-attention (left) and self-attention (right) networks. The cross-attention network illustrated shows Sensor A attending to Sensor B. However, the embeddings for Sensor A are repeatedly transformed during every attendance to the other sensors. The other modalities also undergo this in their branch to ensure bidirectional information flow. During real-time classification, when the caching is active, the new sample embedding, X_t, is projected and attended to the previous tokens and vice versa, and is then appended to the cached attention scores. The cached attention score that represents the oldest sample, QK_{s+1}, is removed, to create a sliding buffer effect. Each time a generated token is added, this mechanism allows the model to utilise previous history of each signal to make an informed prediction, whilst significantly reducing both memory and compute, since the majority of the scaled dot-product attention does not need to be recomputed during this process.

$$AS_m^{cache} \in \mathbb{R}^{B\times H\times S\times S} \qquad \forall m \in \{A, B, C \ldots, M\} \tag{4.15}$$

To achieve this in the context of a sequence-to-sequence model, we must modify the convectional KV cache implementation to store both previous projections and attention scores, and add a sliding

mechanism, as shown in Figure 4.5. Once the cache exceeds a certain length, that is, when it holds queries, keys and value corresponding to $S+1$ embeddings, the projections Q_{S+1}, K_{s+1} and V_{s+1} and attention score AS_{s+1} that correlate to the oldest embeddings are removed.

PyTorch's multi-head attention mechanism, the `MultiHeadAttention` class [118], can be utilised for the attention encoder if caching was not required. However, to integrate QKV cache so that the new token query not only attends to the previous KV cache, but also the previous tokens attend to the new query, the entire mechanism must be rewritten. Likewise, whilst PyTorch's LLM Fine-tuning library, torchtune, has a `KVCache` method [119], it must be rewritten to implement attention score caching. The sliding caching mechanism class, named `SlidingAttnScoreCache`, is located in attention_mechanisms.py, in addition to the `SlidingCachedMultiHeadAttention` class which corresponds to the complete attention mechanism shown in Figure 4.5. This code is presented in algorithmic form, as shown in Algorithm 1.

Since the previously cached embeddings are used in the cached scaled dot-product attention, the output after concatenation will be of shape $X_t^{''} \in \mathbb{R}^{B \times S \times E}$. To obtain an embedding of the same shape as the input, that is of sequence length equal to one, the embedding must be downsampled back to a single embedding when caching is used. This is achieved through the self-attention pooling mechanism, shown in Figure 4.5 and expanded in 4.6

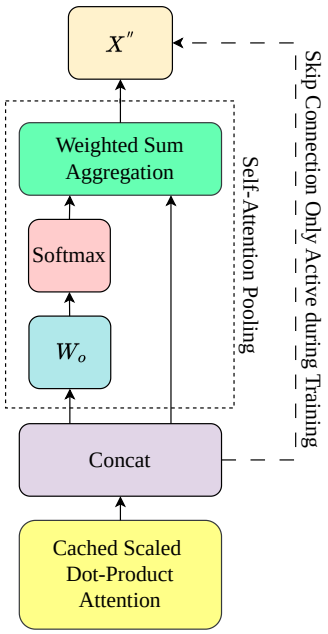


Figure 4.6: Self-Attention Pooling is employed to downsample the source sequence length, S, to the original target sequence length, L, when caching is utilized during inference. Since the model does not utilize caching during training, the skip connection is active instead of pooling.

Algorithm 1 Sliding Cached Multi-head Attention

Input:
q: Query matrix of size $[B, L, D]$ ▷ B: Batch size, L: Sequence length, D: Embedding dimension
k: Key matrix of size $[B, L, D]$
v: Value matrix of size $[B, L, D]$
num_heads: Number of attention heads
$head\_dim = D/num\_heads$: Dimension per head
Output: Output tensor of size $[B, L, D]$
function INITIALIZATION(D, num_heads, max_batch_size, max_seq_len)
$head\_dim \leftarrow D/num\_heads$
Initialize projection matrices for each head:
$W^Q \in \mathbb{R}^{D \times head\_dim}$, $W^K \in \mathbb{R}^{D \times head\_dim}$, $W^V \in \mathbb{R}^{D \times head\_dim}$
Initialize caches for attention:
$q\_cache \in \mathbb{R}^{B, num\_heads, max\_seq\_len, head\_dim}$,
$k\_cache \in \mathbb{R}^{B, num\_heads, max\_seq\_len, head\_dim}$,
$v\_cache \in \mathbb{R}^{B, num\_heads, max\_seq\_len, head\_dim}$
Initialize attention score cache:
$attn\_score\_cache \in \mathbb{R}^{B, num\_heads, max\_seq\_len, max\_seq\_len}$
end function
function FORWARD PASS(q, k, v)
Step 1: Linear Projections
Project query, key, and value into multiple heads:
$q_h = [qW^Q] \in \mathbb{R}^{B, num\_heads, L, head\_dim}$
$k_h = [kW^K] \in \mathbb{R}^{B, num\_heads, L, head\_dim}$
$v_h = [vW^V] \in \mathbb{R}^{B, num\_heads, L, head\_dim}$
Step 2: Append Projections to Cache
Update sliding caches for query, key, value:
$q\_cache[:, :, current\_seq\_len, :] \leftarrow q_h[:, :, 0, :]$
$k\_cache[:, :, current\_seq\_len, :] \leftarrow k_h[:, :, 0, :]$
$v\_cache[:, :, current\_seq\_len, :] \leftarrow v_h[:, :, 0, :]$
Step 3: Compute Scaled Dot Product Attention
Compute attention scores:
$q_t \leftarrow q_h \cdot k_h^\top / \sqrt{head\_dim}$
Update attention score cache:
$attn\_score\_cache[:, :, current\_seq\_len, :] \leftarrow q_t[:, :, 0, :]$
$attn\_score\_cache[:, :, :, current\_seq\_len] \leftarrow k_h[:, :, :, 0]$
Retrieve all cached attention scores:
$attn\_scores \leftarrow attn\_score\_cache$
Apply softmax:
$attn\_weights = softmax(attn\_scores)$
Step 4: Compute Attention Output
Multiply attention weights with value cache:
$output = attn\_weights \cdot v\_cache$
Step 5: Attention Pooling
Apply attention pooling to downsample:
$pooled\_output \leftarrow$ AttentionPooling($output$)
Step 6: Final Output
Return the final output:
$output \leftarrow dropout(pooled\_output)$
Increment sequence length:
$current\_seq\_len \leftarrow current\_seq\_len + 1$
end function

Whilst other downsampling methods can be used, self-attentive pooling has shown to efficiently capture complex relationships between features, outperforming the traditional pooling techniques, convolutional and max pooling, in terms of accuracy and memory efficiency [120]. This is particularly useful for resource-constrained applications, such as deploying the proposed stress detection model on edge devices and compensating for the increased memory usage due to caching. Another reason for selecting this method is that self-attention pooling does not rely on local context, unlike conventional methods, and can capture long-range dependencies. This will enable better scaling when a larger context window and thus sequence length, S, is chosen.

4.4.3 Predictor

The goal of the predictor is to decode and fuse the embeddings produced by each encoder from the ensemble modality branches. The form of the latent space embeddings are tensors in the same dimension as the feature embeddings, i.e. $X' \in \mathbb{R}^{B \times S \times E \times M}$. During the early fusion stage, the bidirectional cross-attention blocks, which aid in aligning the data from different sensors, will inherently put more emphasis on certain features than others. Nonetheless, a predictor is still required to decode and map joint representations into a final predictive label y_t. This mapping must be done in a modular way as before so that any different combination of sensors may be utilised.

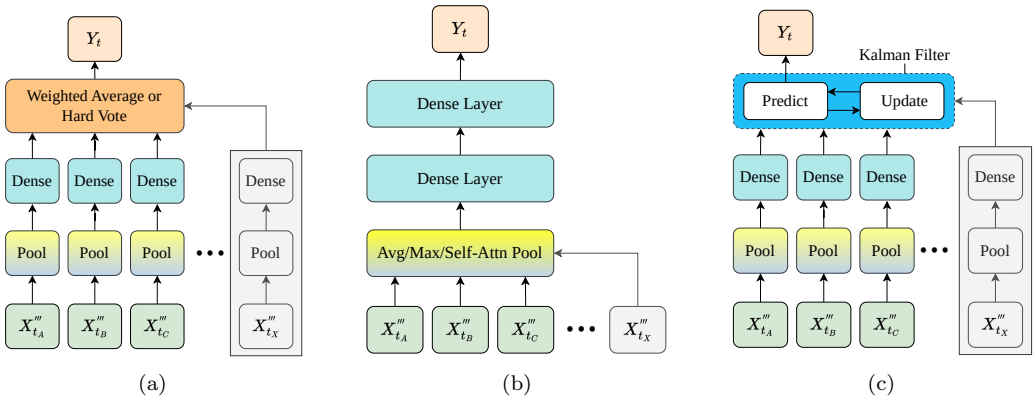


Figure 4.7: a) Soft or hard vote modular predictor. b) Pooling branch fusion predictor. c) Kalman filter predictor. Each of the fusion predictors are designed for modularity: to take any number of input branches (modalities) and output a reliable stress-label prediction. N.B. the pooling methods are applied to the sequence length dimension. Thus during inference, when only one temporal slice is inferred, there is no transformation since the target sequence length is one.

Conventional methods for feature in–decision out (FEI-DEO) fusion can be split into hard voting and soft-voting methods [121]. In hard voting, each branch provides a discrete prediction,

typically the class label with the highest probability. These predictions are then aggregated by a majority vote. The final prediction is the class that receives the most votes from the individual branches. This approach can be beneficial when the individual branches are highly confident in their predictions, for example when one sensor has a significantly more reliable prediction than another, perhaps as it is less susceptible to motion artefacts or one subject's biomarkers is more prominent on the heart than the sweat glands, for example. However, this will not always capture nuanced variations in the confidence levels of different branches. Therefore, it is expected that this will not produce optimal results, nonetheless the `ModularHardVoting` module is employed by the script hard_voting.py to test this.

Soft voting methods aggregate the predicted probabilities from each branch instead of the hard class labels. In this approach, each branch outputs a probability distribution over all possible classes, reflecting its confidence in each prediction. These probability distributions are then averaged (or weighted averaged), and the class with the highest aggregated probability is chosen as the final prediction. The weighted average variant aims at decoding how reliable each branch is by a learned weight parameters, which can later be personalized to adapt to a specific user's stress response.

The pooling branch fusion predictor utilizes average, max or self-attention pooling followed by linear transformations and activation functions to process the features from each branch. The "ModularWeightedPool" class which can be found in soft_voting.py introduces learnable weights for each branch, allowing the model to dynamically adjust the importance of each branch based on the data, prioritizing more informative branches but not to the extent of the hard voting model.

A larger average pooling predictor, whereby the features of each branch are first fused and then pass through two fully-connected layers, has proven effective in other studies utilizing attention for feature convergence [93], [102]. Zhang et al. tested several late fusion approaches, finding that for binary classification on the UBFC-Phys dataset, average and maximum pooling predictors, termed AvgPool and MaxPool respectively, significantly outperformed other architectures. In three-class classification tasks, predictors without BiLSTM consistently exhibited superior performance. Similarly, on the WESAD dataset, predictors without BiLSTM showed better accuracy in both binary and three-class scenarios, with AvgPool achieving the highest results. While BiLSTM can capture long-term dependencies, its complexity often leads to overfitting, especially with the small sample sizes in these datasets. This results in diminished performance. In contrast, simpler structures like AvgPool and MaxPool, which aggregate features by averaging or selecting maximum values across the temporal slices, proved more robust and effective.

An alternative fusion technique to BiLSTM, which may show to be more robust in capturing the temporal context of the signal is the Kalman filter. Pakrashi et al. [122] treats the ensemble model as an unknown state that can be estimated by the Kalman filter, just like the aim of the memory cell in BiLSTM. Through this, they achieved state-of-the-art performance in 30 datasets in several domains, including ones with physiological data. The filter that has since further been exemplified specifically for multimodal stress classification tasks through the SELF-CARE system developed in Rashid et Al. [22]. Whilst achieving the highest accuracies in the WESAD dataset through an ensemble of different modality combinations, it demonstrated that the Kalman filter showed to excel in capturing the temporal dynamics and updating its estimates in real-time in an ensemble system, crucial for applications where physiological states change dynamically. For its well-documented handling of noisy measurements by explicitly accounting for noise via its modelling of noise measurement to make it more robust to the inherent variability in physiological sensor data.

Additionally, the Kalman filter enhances the energy efficiency of model inference, which is critical for wearable devices. By optimizing sensor and classifier usage based on the current temporal context, it reduces unnecessary computations. In contrast, BiLSTM models, although capable of capturing temporal dependencies, require more computational power and memory, making them less suitable for resource-constrained edge computing. The adaptive sensor fusion capability of the Kalman filter ensures the most relevant data is used without overloading the system, significantly improving both accuracy and energy efficiency. The SELF-CARE system demonstrated to be an extremely energy-efficient compared to other traditional methods due to the lightweight nature of the Kalman filter. These attributes collectively highlight the superiority of the Kalman filter in real-time, low-power stress detection applications compared to BiLSTM models. Thus, for this study, a Kalman filter was chosen instead of the BiLSTM to investigate the effectiveness of temporal dynamics monitoring of the system.

This implementation enhances the SELF-CARE Kalman filter for multi-class classification by introducing learnable parameters, no longer requiring tuning, which would be tedious for personalized models. The state transition matrix F, measurement matrix H, process noise covariance Q, initial state x_0, and initial covariance P_0 are learned during training, unlike SELF-CARE's fixed approach. The filter retains SELF-CARE's dynamic measurement noise covariance calculation: $R = ((1 - z) \cdot 2 \cdot I)^2$ for 3-class and $R = ((1 - z)/2 \cdot I)^2$ for 2-class problems, where z is the measurement. Notably, the epsilon threshold (ε) for measurement selection and gamma ($\boldsymbol{\gamma}$) scaling factor for class imbalance correction are implemented as learnable parameters, initialized to

the same values as in SELF-CARE. The implementation of this filter can be viewed on the script Kalman.py of the repository.

4.5 Model Training and Validating

The proposed model architecture and its variants were validated using the publicly available WESAD dataset [16], as described in Chapter 2, in addition to the UBFC-Phys dataset, which facilitates comparison with models from previous research. Additionally, the model was validated on the MUSED dataset, collected specifically for this study, for which no prior benchmarks exist. A summary of the datasets are provided in Table 3.2.

N.B. For the WESAD dataset, S1 and S13 were removed as explained in the original paper [16]. For the UBFC-Phys dataset, we must note that due to the reasons expressed in the paper [17], we must neglect 14 subjects from the rest condition, 30 subjects from the speech task, and a further 23 subjects due to sensor malfunction. Therefore, we are left with 15 subjects (s2, s7, s15, s16, s18, s20, s23, s24, s29, s34, s36, s43, s44, s46, s51). For the MUSED dataset, we neglect S2 and S3 due to compromised fNIRS signal quality.

4.5.1 Model Pre-Training

To evaluate how effectively the model generalizes to new individuals, we employed LOSO-CV when training on participant data. This approach was used for the WESAD, UBFC-Phys, and MUSED datasets to ensure consistency with performance comparisons from other studies.

Similar to other sequence-to-sequence models, the proposed model must learn the temporal patterns in sequential data samples. However, due to the protocols used in these datasets, training a model sequentially on one class followed by another class from the same subject could introduce bias and result in unstable learning. To address this, the dataloader was designed to randomly shuffle the batched training segments by permuting the indices of all samples. This shuffling occurs at the batch level, meaning the order of training sequences is randomized, rather than shuffling individual temporal slices within each sequence. This approach improves learning stability and reduces the risk of overfitting.

During training, the sliding caching mechanism was not employed, which means the model learned to make predictions based on entire batched sequences rather than a single new token

informed by previous sequences. It is worth noting that alternative training methods may be more suitable for larger datasets. For instance, in masked language modelling used by BERT, the model is trained by randomly masking certain words in a sentence and predicting them [117]. This approach could be more efficient for learning contextual relationships between embeddings, although it was not considered in our current methodology.

As demonstrated in Table 3.2, the stress labels are not uniformly distributed across any of the datasets used in this study. To address the issues caused by data imbalance, where the classifier may become biased toward the majority class, we employed focal loss as the loss function for our model. Focal loss, proposed by Lin et al. [123] and demonstrated to be effective for attention-based models in Yu et al. [23], helps mitigate class imbalance by dynamically adjusting the learning focus towards harder-to-classify examples. The focal loss function is defined as follows:

$$FL(p_t) = -(1 - p_t)^{\gamma} \log p_t \tag{4.16}$$

p_t represents the logit output from the model after the softmax is applied, and is given by:

$$p_t = y \cdot \hat{y} + (1 - y) \cdot (1 - \hat{y})$$

Where y is the predicted probability as a vector, $\hat{y}$ is the one-hot encoded ground truth label and γ is the focusing parameter. The primary aim of using this loss function is to direct the model's attention towards hard-to-classify samples during training by decreasing the weights assigned to easy-to-classify samples.

To optimize model training, a range of hyperparameters, as listed in Table

Hyperparameter	**Search Space**
embed_dim	[8, 16, 32]
hidden_dim	[16, 32, 62, 64, 128, 256]
n_head_gen	[2, 4, 8]
dropout	[0.3, 0.5, 0.7]
attention_dropout	[0.3, 0.5, 0.7]
learning_rate	[0.0001, 0.001, 0.01]
batch_size	[8, 16, 32]
epochs	[5, 7, 10]
fine_tune_epochs	[1, 3, 5]
fine_tune_learning_rate	[0.001, 0.0001, 0.00005]
early_stopping_patience	[5, 7, 10, 20]
early_stopping_metric	[loss, accuracy]

Table 4.4: Hyperparameter Grid Search Space. For the optimal hyperparameters selected for each model tested, see the config files directory.

4.5.2 Modular Fine Tuning

To ensure the model is able to perform well under modular circumstances, i.e. different sensor combinations, the branches corresponding to the sensors are randomly removed during training and thus the model relies on other combinations to make the prediction. For UBFC-Phys, this was simply a matter of removing one-by-one the BVP and ECG branches individually during fine-tuning. However, for the WESAD dataset, the all-modality model that was trained on ten sensors, see Table 3.2, it would result in 1023 different combinations: $\sum_{k=1}^{10} \binom{10}{k} = 2^{10} - 1 = 1023$. Likewise, the MUSED dataset, which employs nine sensors, would result in 511 combinations. Since this is not pragmatically feasible, the training and fine-tuning for both datasets were conducted in the order as follows:

Training Type	Modality	Reason
Pre-train	All	General knowledge and high co-dependence between modalities. Optimal for all modalities.
Fine-Tune	Unimodal	Strengthen individual branches to improve self-reliance, reduce co-dependence and ensure it can interpret embeddings that have not undergone transformation via the cross-attention network.
Fine-Tune	Physiological	Remove reliance on motion signals and learns to detect and disregard noise/artefacts from signal.
Fine-Tune	Wrist	Strengthen for wrist only stress detection.
Fine-Tune	Chest	Strengthen for chest only stress detection.

Table 4.5: Reasoning behind each training stage for modular assurance of the WESAD dataset. For MUSED dataset, a similar approach is employed, whereby the fine-tuning occurs for each modality (each of the four devices used) and on a per-modality basis.

4.5.3 Non-Batched Fine-Tuning

As described in Section 4.4.2, the prediction task during training differs slightly from the prediction task during real-time inference. During training, the model learns to predict a label for the entire input sequence, whereas during real-time inference, it predicts the label corresponding to the final embedding, using the context of the entire sequence. This difference implies that the pre-trained model may not sufficiently focus on the most recent embedding. To address this issue, after completing the modular fine-tuning process described earlier, the model undergoes additional tuning on a per-token basis to increase the attention given to the latest embedding during the attention mechanism.

To facilitate this tuning, a new dataloader, termed `SeqToSeqDataloader`, was developed. This dataloader segments the data into randomized portions while preserving the sequential relationship

between temporal slices. After a specified interval, defined by the parameter `fine_tune_sequence_length`, the model resets its cache and continues tuning on a new sequence of data drawn from a different part of the dataset. This approach ensures smoother learning transitions. The fine-tuning process takes a value for `fine_tune_sequence_length` that is divisible by S, the number of temporal slices stored in memory. For instance, if the model uses a memory of 6 temporal slices, a sequence length of 30 is chosen, corresponding to 5 sequential batches before a new sequence is selected at random and the cache is reset. Although this fine-tuning approach is less efficient, the pre-trained model only requires minimal adjustment to accommodate per-token prediction.

4.5.4 Computational Performance Evaluation

As discussed in Chapter 6, the performance gains from caching are closely tied to RAM read and write speeds. To optimize the dataloader's efficiency, we utilize the HDF5 file format for efficient data loading [124]. To reliably measure model inference times, we include a warm-up phase to stabilize performance, which is particularly crucial when using GPUs. Inference times are then measured over multiple iterations of the forward pass, after the data has been loaded into memory. To ensure accurate timing on GPUs, synchronization operations are employed. The elapsed times, measured in milliseconds, are calculated along with their corresponding averages and standard deviations to assess typical performance and variability. These measurements are repeated for each number of modalities in the WESAD dataset to evaluate the impact of modularity on computational time.

4.5.5 Predictor Evaluation

To fairly evaluate the chosen predictors detailed in Section 4.4.3, the checkpoints from the best performing model for each dataset and subsequent classification task were stored. The model then employed transfer learning by substituting the original predictor, which utilized the Average Pooling variant of the Pooling Branch Fusion Predictor shown in Figure (c)), with the tested predictor. All other layers were frozen since late-fusion should not depend on any internal workings of the branches, just their final prediction, thus the cross-attention fusion is unaltered in this evaluation.

4.5.6 Personalized Model Training

As emphasized by Yu et al. [23], while fine-tuning a generalized model on an individual dataset is effective, it does not combat the redundancy of parameters which were required to learn inter-participant stress responses. Therefore, for our model, we train the personalized model from scratch with randomly initialized weights. While the model architecture remained the same as the generalized version, the number of parameters was reduced to minimize redundancy. Specifically, the hidden dimension size was reduced from 32 to 16, decreasing the complexity of the linear layers that likely contained redundant information related to inter-individual variability. The embedding dimension was kept at 16, as reducing it further led to performance degradation, likely due to underfitting at the embedding layer. This underfitting occurs when a smaller feature size results in insufficient capacity to capture the necessary representational information.

5

Results

Contents

First, the raw signals captured during the study are analysed to validate the integrity of the data and identify any key findings. Subsequently, we present the model's performance results on two well-established datasets, WESAD and UBFC-Phys, as well as on the newly curated MUSED dataset. These results evaluate the model's overall performance across the different datasets and under various modality combinations, demonstrating its ability to adapt and modularize. To further evaluate the influence of different components within the model architecture, an ablation study is performed. Finally, the model's performance is benchmarked against other state-of-the-art models from recent literature.

5.1 Signal Validation

5.1.1 fNIRS

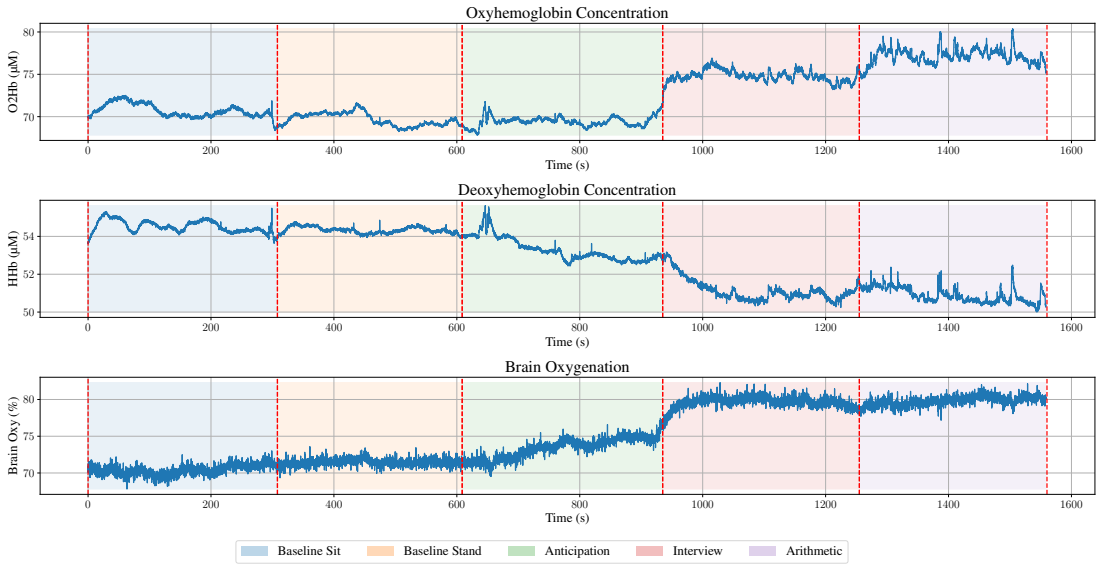
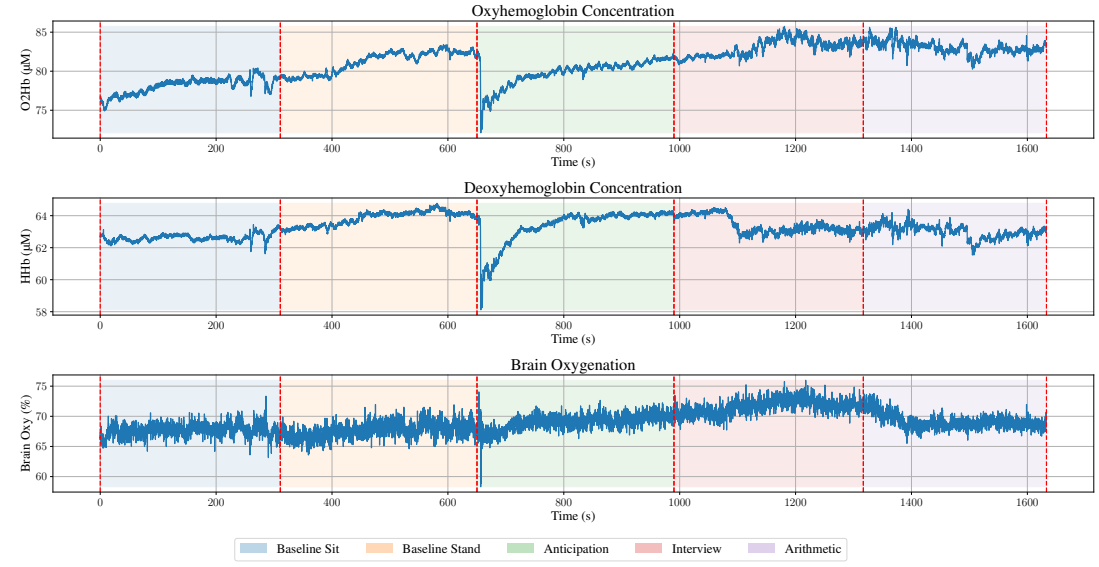


Figure 5.1: fNIRS captured from Subject 13 shows an increase in oxyhaemoglobin concentration and brain oxygenation and a respective decrease in deoxyhaemoglobin during the stress conditions, indicating an increase in brain activity.

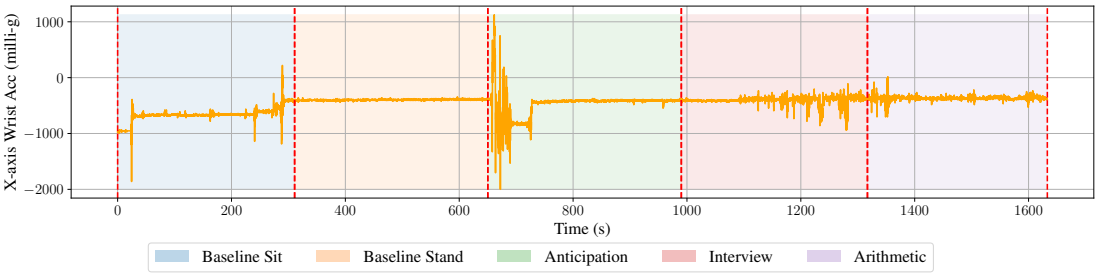
Figure 5.1 indicates good signal quality, allowing for the identification of a clear shift in brain activity of the prefrontal cortex in both interview and arithmetic conditions, which indicates an increased cognitive load. Other subjects' data follow this trend, whilst a few show no noticeable increase in brain oxygenation levels, named non-responders, which can be viewed in Figure 5.8. These individuals may not have experienced the same mental effort during these tasks, or their biomarkers may not show a blood oxygenation increase to the same degree as others.

Some fNIRS data was not accurately captured, particularly for Subjects 2 and 3, resulting in their exclusion from model training. Additionally, as shown in Figure 5.2, motion artefacts can cause a prolonged recovery period of the hemodynamic features. Furthermore, Figure B.2 illustrates that the fNIRS data for Subject 16 exhibited an extended recovery period following a motion artefact that caused a shift in blood volume; however, the brain oxygenation measurements remain stable, indicating resilience to such artefacts. In contrast, the fNIRS data from Subject 17 display a period of compromised signal quality in the brain oxygenation channel, potentially due to poor sensor contact or excessive sweat on the sensor.

The extraction of heart rate from fNIRS, as outlined in Section 4.3, is presented in Figure B.1.



(a)

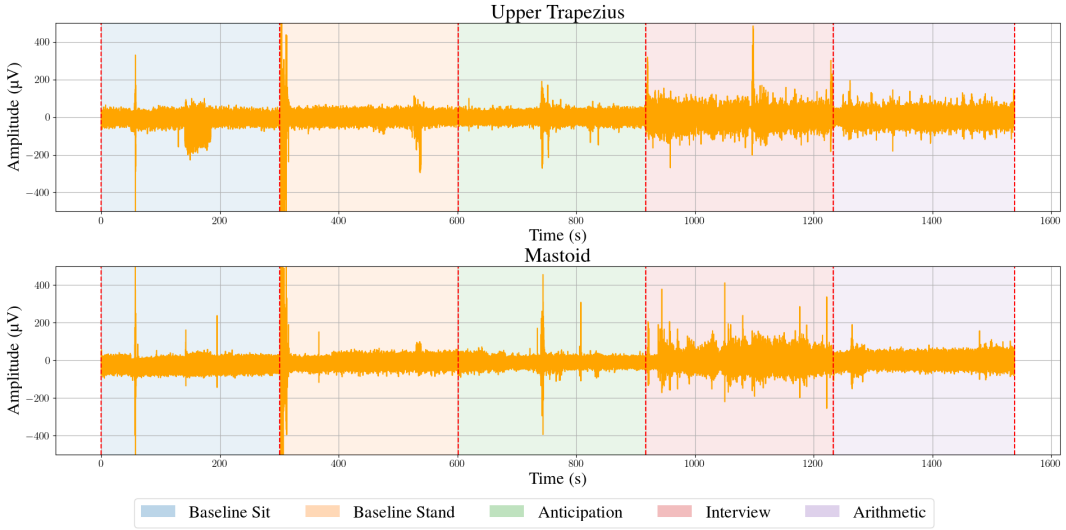


(b)

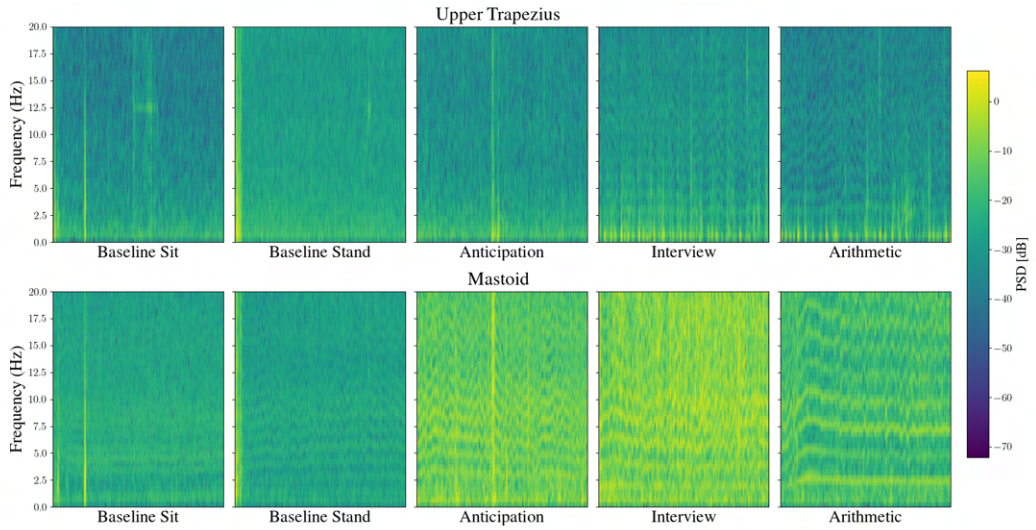
Figure 5.2: Motion artefacts in Subject 12 fNIRS data: (a) Slow recovery of oxyhaemoglobin and deoxyhaemoglobin levels following motion, and (b) corresponding wrist accelerometer data confirming the introduction of the motion artefact at 620 seconds.

5.1.2 sEMG

The sEMG recordings, at both the upper trapezius and mastoid region, reveal a signal with higher power spectral density during the interview condition, shown in Figure 5.3 (b). The upper trapezius sEMG shows clearer distinction between frequency bands, most particularly during the arithmetic task, which aids in the distinction of muscular contractions, whilst the mastoid has surrounding noise, lowering the visibility of these pulses. Vertical pulses across all frequency bands, as seen at the start of the baseline stand (while getting up from the chair), and during the anticipation task (other movement) are due to motion artefact and should be discarded as noise. As expected, sEMG is particularly susceptible to surrounding noise in the environment, which can be observed as an increase in power spectral density across different frequencies and a reduced disparity be-



(a)



(b)

Figure 5.3: a) sEMG measurement from Subject 7. b) Corresponding low frequency spectrogram reveals elevated power spectral density (PSD) during anticipation, interview and arithmetic tasks with some fluctuations rhythmic fluctuations can be observed in the mastoid, indicating the presence of sEMG during the stressor.

tween frequency bands. This noise may obscure the detection of distinct muscle activity pulses, making it harder to discern specific muscular signals from overlapping noise. The high-frequency spectrogram, displayed in Figure B.3 shows minimal activity in Subject 7, with other subjects demonstrating similar results, suggesting that the muscular activity did not reach the same amplitude levels observed during a shoulder contraction, as demonstrated in the pilot, see Figure 1.2 (b).

5.1.3 Cardiovascular & Hemodynamic Measurements

Whilst PPG obtained through the Empatica E4 has shown to be relatively robust and effective in determining stress in numerous studies, including the two other databases being investigated [16], [17], we do observe that several recordings demonstrate severe noise disturbances. Figure B.5 illustrates two distinct sources of induced noise. In Subject 1, the noise is attributed to poor sensor connection, as no corresponding wrist acceleration movements were detected. In contrast, Subject 18 shows noise associated with clear hand gestures.

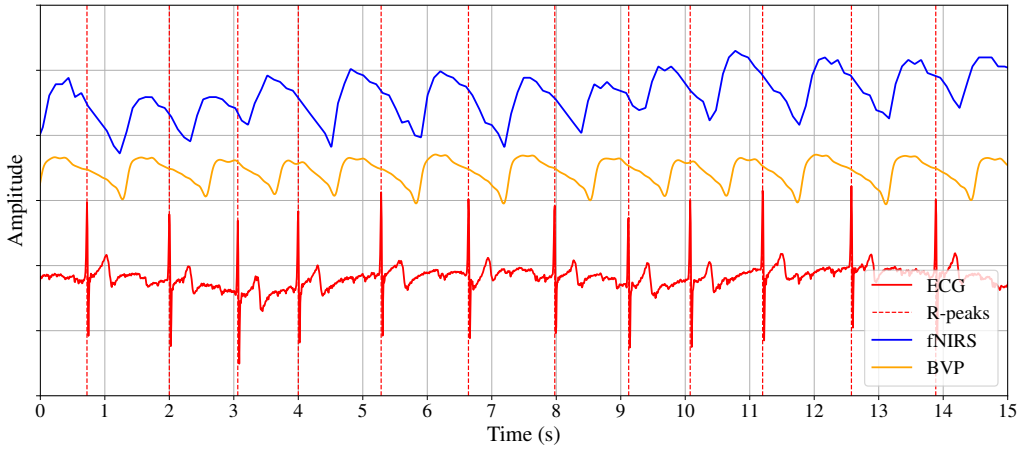
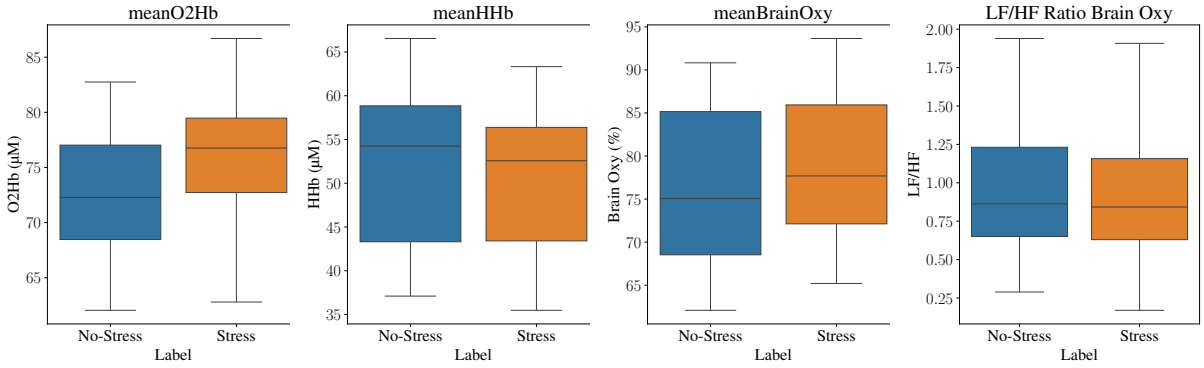


Figure 5.4: Cardiovascular (ECG) and hemodynamic measurements (BVP and fNIRS) during a subjects' baseline sit period. R-Peaks were obtained from the polar device, and due to its highest sampling rate of 1kHz is taken as ground truth for heart rate. The synchronization of the three devices is demonstrated, however as explained in Section 3.0.2, modalities were not precisely aligned in some recordings.

The R-peaks signal generated by the Polar device can serve as a ground truth and provides a more accurate method for interpreting key features, such as HRV compared to the other measurements shown in Figure 5.4. HRV, has demonstrated to be a strong indicator of stress and is widely discussed in the literature [20], [36], [104]. Nonetheless, the additional information from other signals should not be dismissed. For example, the fluctuation in R-peak amplitudes are visible in the ECG recording which can be used to derive respiration, which is also influenced by stress, as detailed in section 2.1. Additionally, more disguised features such as EDR, which may be derived from this using efficient algorithms and shows to be as effective as a respiration belt [125]. We present the extraction of EDR, peaks, and onsets and offsets of PQRST complex derived from the ECG in Figure B.4. Further signal processing and corresponding observations can be viewed in Section B.2.

5.2 Feature Evaluation



(b)

Figure 5.5: Box plots illustrating various features extracted from the single-channel fNIRS device. Statistical analysis using a T-test revealed that mean oxygenated haemoglobin (meanO2Hb), mean deoxygenated haemoglobin (meanHHb), and mean brain oxygenation (meanBrainOxy) all show statistically significant differences ($p < 0.05$).

Adhering to the qualitative findings obtained during signal inspection, shown in Figure 5.5, we identify a slight increase in oxygenated haemoglobin (O2Hb) and absolute brain oxygenation (brain oxy) levels under stress conditions. Conversely, the mean of deoxygenated haemoglobin (HHb) decreases. Whilst, the net difference between mean O2Hb is greater than brain oxygenation, due to its more sensitive recording, it can be observed that the stress condition has a wider variance.

The LF/HF ratio feature is presented since it has been thought to substitute the heart rate derived fNIRS, and has shown to provide a more reliable especially in stress conditions, compared to using LF and HF metrics individually [126]. However, a recent study have made counter-arguments to this claim [127]. For the MUSED dataset, no statistical significance between the two labels are observed. Nonetheless, we do observe statistical significance for two heart-rate derived features, namely mean O2Hb_HR and max O2Hb_HR ($p > 0.05$), demonstrating the effectiveness of the algorithm proposed by [14] even for a single channel device.

While the variance of alpha and beta power increases significantly during the stress condition, it is not statistically significant. Utilising the relative power between these two EEG frequency bands yields does show a visible slight decrease in both modalities. This ratio does however adhere to the expectation, where alpha waves, associated with calmness, decrease relative to beta waves, which infer logical reasoning, decrease [91]. However, both mastoid and upper trapezius have comparable features, which suggests that this deviation could be attributed to a condition such as noise, which is apparent in both sensors, and thus we cannot conclude that EEG activity is present.

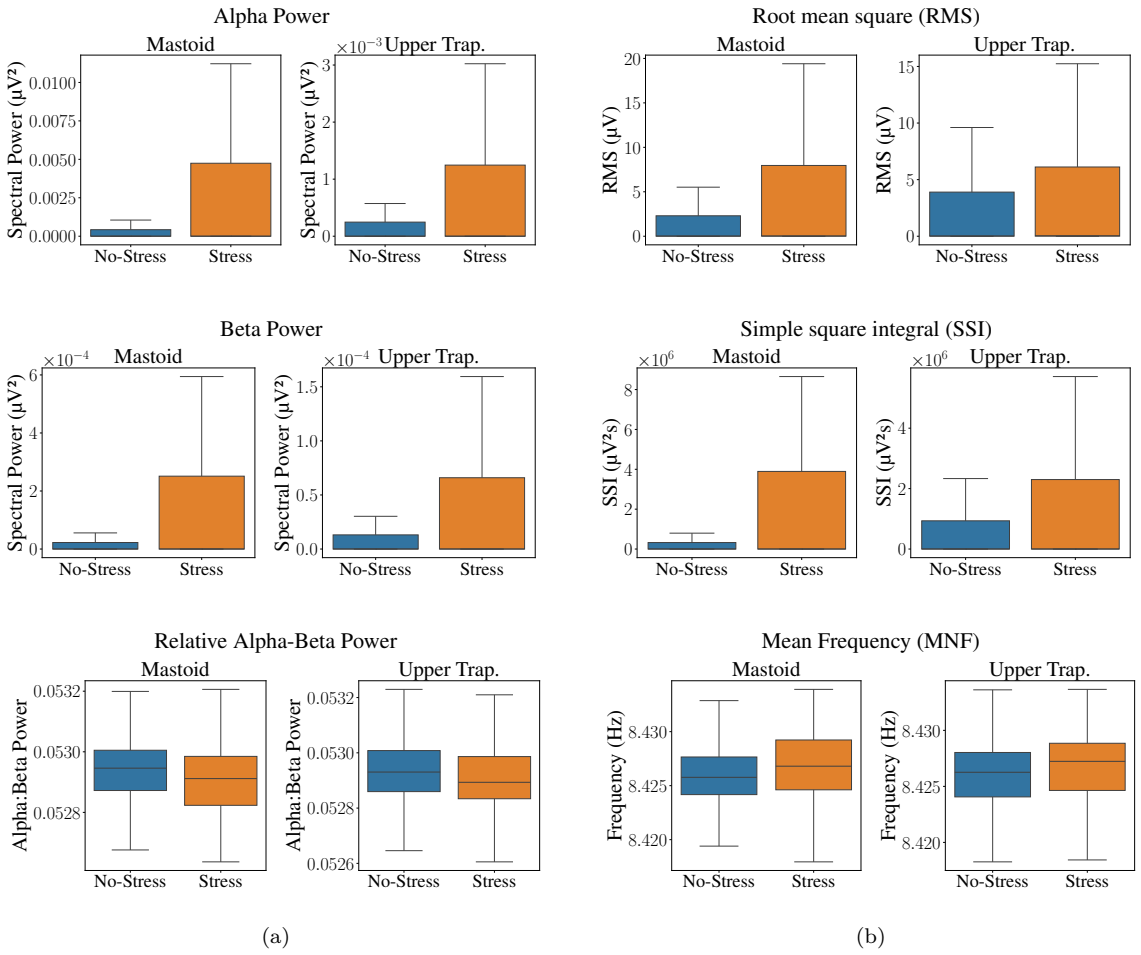
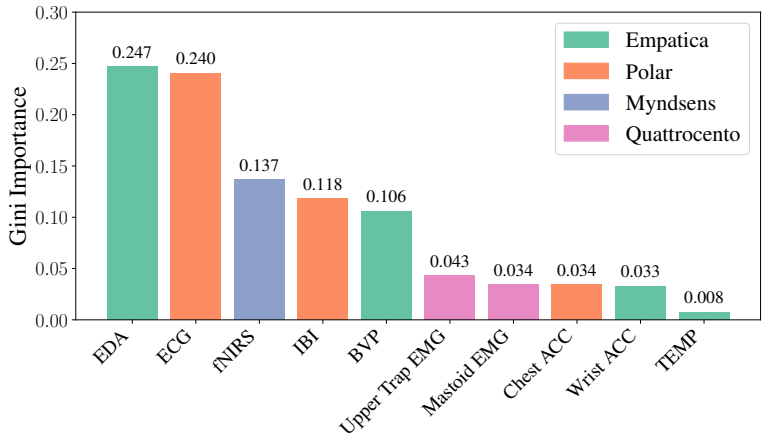


Figure 5.6: Box plots illustrating the extracted features from the EEG and sEMG signals recorded at the mastoid and upper trapezius. (a) The relative beta-alpha ratio is not significantly different ($p = 0.078$) in both the mastoid and upper trapezius regions. (b) Root Mean Square (RMS) values show significant differences ($p < 0.05$) at both the mastoid and upper trapezius. Additionally, Mean Frequency (MNF) demonstrates significant differences ($p < 0.01$) across both conditions. All other comparisons yielded non-significant differences ($p > 0.05$).

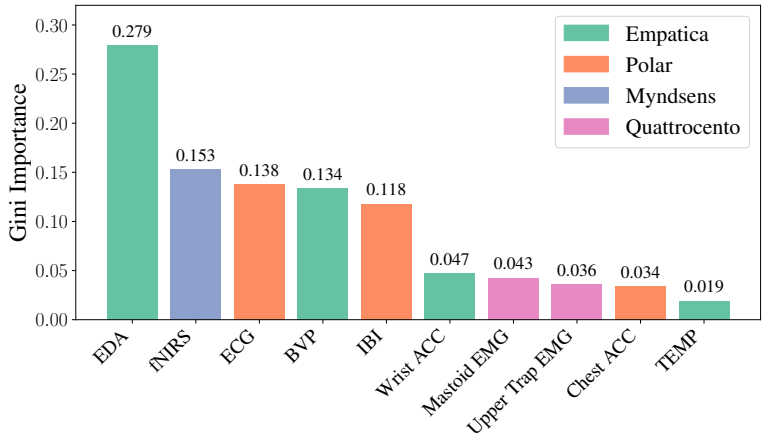
The sEMG-based features, SSI and MNF, were selected for their demonstrated effectiveness in two studies utilizing the trapezius muscle for stress classification [54], [128]. The RMS feature also proved to be a valuable addition, as highlighted in [54], although it was computed relative to a reference contraction specific to each participant. Notably, a substantial difference between stress and non-stress conditions was observed for SSI and RMS; however, similar to EEG features, the mean difference between conditions was not statistically significant. This may suggest that while some participants exhibited greater muscle activation under stress, the majority did not show notable changes between the conditions.

The MNF feature emerged as the most effective, showing a significant increase during the stress condition. A disparity between the mastoid and upper trapezius muscles was also observed, with

the mastoid demonstrating lower variance in the non-stress condition, potentially indicating less deviation between participants. However, this finding requires cautious interpretation, as noise may have contributed to this variance, especially considering the class imbalance, where the stress condition lasted for a longer duration, see Figure 3.2. Further investigation is needed to assess the sEMG features with greater rigour.



(a)



(b)

Figure 5.7: Ranked Gini importance scores of all modalities of the MUSED dataset for a) binary classification and b) four-level classification. The importance represents the reduction in the Gini criterion contributed by the feature in question.

The Gini importance scores shown in Figure 5.7 highlight that EDA emerged as the most significant feature as a single predictor for stress. However, it is essential to consider the study's protocol when interpreting these results. Given that perspiration, and thus skin conductance, is likely to increase as a result of wearing the Empatica E4, irrespective of stress levels, thus there is a potential risk of data snooping. We discuss this possibility in depth in Chapter 6.

The second most influential feature was ECG, underscoring its reliability compared to other hemodynamic measurements, such as those derived from heart-based signals. Two factors likely contribute to this result. First, the Polar chest-strap ECG device proves to be more reliable and less susceptible to motion artefacts compared to other devices. This is corroborated by the importance of Polar IBI data, which shows that reliable heart rate intervals alone are as effective as the more data-dense BVP signal. The BVP signal, as shown in Figure B.5, was found to be more prone to poor connection and motion artefacts. Second, ECG offers enriched features, such as those derived from EDR, which further enhance its predictive importance over simpler IBI measurements.

The ranked Gini importance scores for each feature, as shown in Figure B.4, provide notable insights. First, features related to heart rate variability (HRV) dominate the binary classification tasks, supporting the extensive evidence of HRV's effectiveness as a robust analytical tool [20], [104], [129]. These features are primarily attributed to the Polar device, rather than the other hemodynamic sensors from Empatica and Myndsens, further emphasising the Polar's reliability. While the Empatica device also demonstrates strong independent features, these are limited to EDA metrics. Additionally, the fNIRS device shows that oxygenated haemoglobin signals offer substantial classification power in both binary and four-level classification tasks, highlighting its potential as a commercially viable option for future applications.

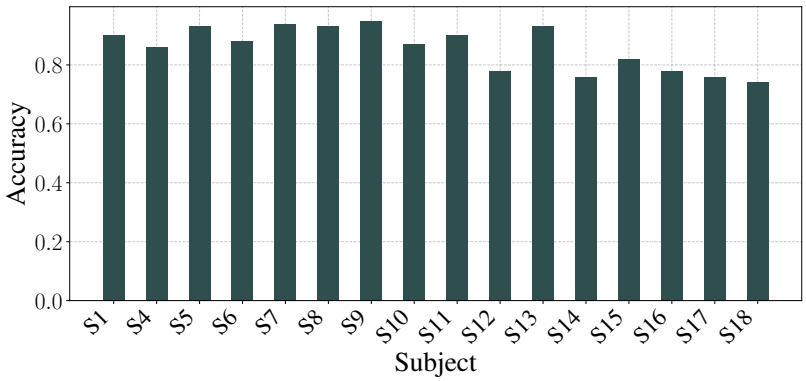


Figure 5.8: fNIRS modality performance on a per-subject basis, using the fixed model and LOSO-CV, illustrates large variance between subjects.

It is important to note that while the top features exhibit relatively high Gini importance, their individual contributions are lower than the overall importance of their respective sensor. This suggests that a combination of features is necessary to optimize separability and achieve reliable stress classification. Dimensionality reduction techniques, such as Principal Component Analysis (PCA), can be employed to identify a limited set of features that maximize variance,

providing a more efficient approach to feature selection compared to utilizing an embedding layer, which is computationally more expensive, as described in Section 4.1.

To assess the substantial inter-participant variability in fNIRS signals, identified both visually during individual signal inspections and quantitatively in the box plots of feature distributions shown in Figure 5.5, a comparison of modality performance was conducted. As shown in Figure 5.8, the classification performance of this sensor varies significantly, with binary classification accuracy ranging from 0.743 to 0.954.

Several factors may contribute to this high variance: the stress protocol may not have consistently induced a high cognitive load in all participants, suboptimal signal quality could have compromised classification performance, or some participants may be non-responders. Variability in fNIRS measurements can arise from inter-participant anatomical differences, such as variations in scalp-to-cortex distance and the volume of gray matter penetrated by the near-infrared light [130]. Furthermore, systemic physiological factors, such as extracranial blood flow, can confound the measurement of neural activation, further contributing to inter-individual differences. This variability underscores the importance of employing fNIRS in a multimodal setting, where more robust biomarkers can be integrated, or selecting sensors that are best suited to the specific needs of individual users through modular predictors.

5.3 Modality and Modularity Evaluation

5.3.1 Predictive Performance

To more clearly demonstrate the impact of the bidirectional cross-modal attention mechanism, we perform a visual analysis of the learned cross-modal attention weights. Illustrated in Figure 5.9, we confirm that the temporal slices of the target sequence attend to the source sequence in both directions for two temporally misaligned features, BVP and EDA, thus demonstrating bidirectional information flow, adhering to the expectation of the cross-modal attention introduced by Zhang et al. [93].

The visualization of ECG attending to ACC and sEMG are shown, whereby ECG gain contexts about noise from motion and muscular activity. The high variance between attention weights in Figure 5.9 (c) demonstrate that the ECG modality pays particular attention to some temporal slices in ACC such as target sequence 0, which may show noise, whilst neglecting others, such as

target sequence 2, which may indicate little movement. Similarly, in Figure 5.9 (d), we see that less attention is given to temporal slices 3 and 4 relative to others.

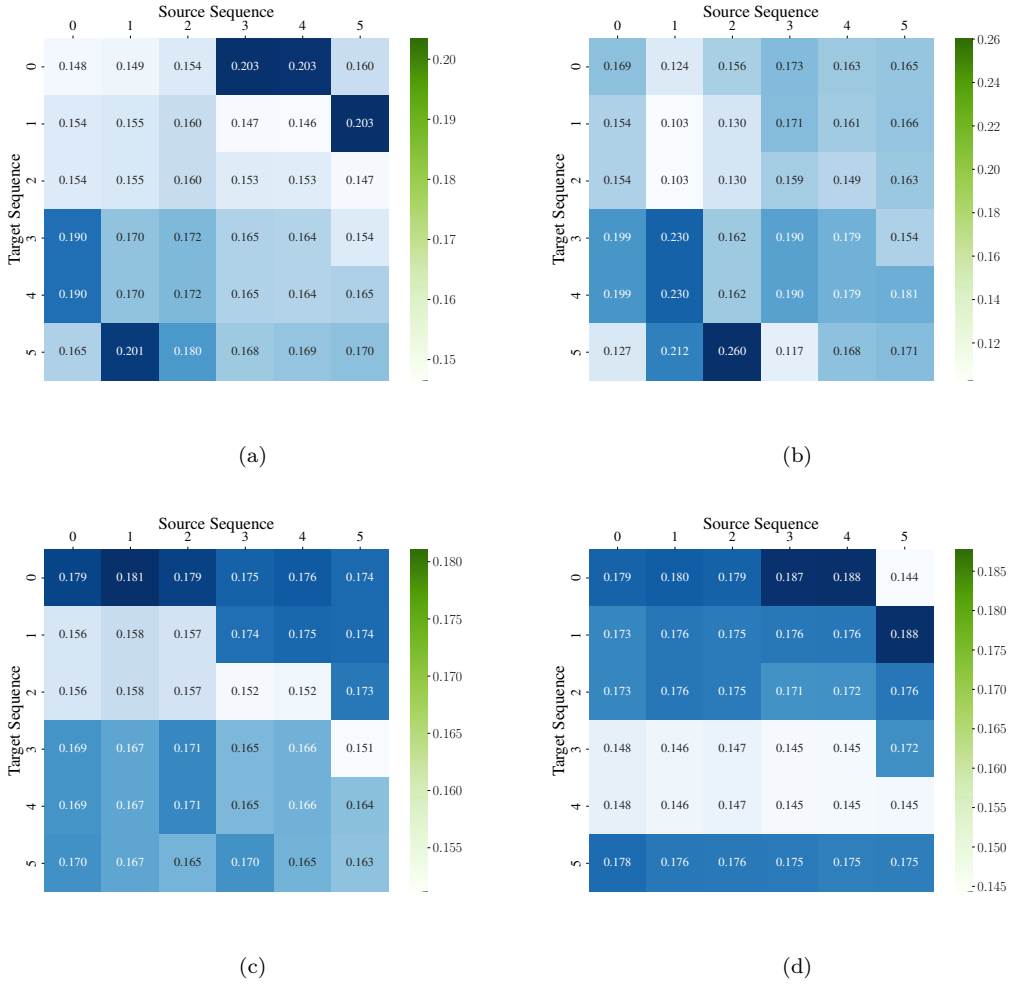


Figure 5.9: Visualization of independent heads of cross-modal attention weights where a) BVP attends to EDA b) EDA attends to BVP c) ECG attends to ACC d) ECG attends to sEMG using the bidirectional encoder-decoder attention and sliding attention score caching mechanism. The attention scores are retrieved from the cache, concatenated with the new single-sequence token attention score, then softmax is applied to normalize each source sequence, producing the attention weights presented.

In Table 5.1, we present the performances of our proposed model on WESAD in the same manner as the original paper [16]. The accuracies achieved through this model are superior to the original paper for all modality combinations, for both the fixed model and generalised modular equivalent. Performance related to more sophisticated methods, deemed state-of-the-art, are shown in Table 5.5. All F1 scores reported in this study are macro-averaged to ensure a balanced evaluation across all classes.

The higher performance in chest based features are shown, and show comparable performance

	Binary Classification				Three-Level Classification			
	Fixed Model		Generalized Modular Model		Fixed Model		Generalized Modular Model	
Modality	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
				Wrist				
BVP	0.863 ± 0.061	0.832 ± 0.058	0.837 ± 0.061	0.806 ± 0.058	0.740 ± 0.081	0.626 ± 0.078	0.716 ± 0.081	0.602 ± 0.078
EDA	0.794 ± 0.086	0.764 ± 0.084	0.770 ± 0.086	0.740 ± 0.084	0.694 ± 0.086	0.578 ± 0.083	0.666 ± 0.086	0.550 ± 0.083
TEMP	0.703 ± 0.031	0.663 ± 0.033	0.679 ± 0.031	0.639 ± 0.033	0.552 ± 0.020	0.437 ± 0.022	0.526 ± 0.020	0.411 ± 0.022
ACC	0.742 ± 0.104	0.711 ± 0.101	0.718 ± 0.104	0.687 ± 0.101	0.560 ± 0.134	0.449 ± 0.131	0.534 ± 0.134	0.423 ± 0.131
Wrist Physiological	0.895 ± 0.073	0.865 ± 0.071	0.869 ± 0.073	0.839 ± 0.071	0.767 ± 0.084	0.657 ± 0.082	0.747 ± 0.101	0.637 ± 0.099
Wrist All	0.903 ± 0.096	0.872 ± 0.094	0.877 ± 0.096	0.846 ± 0.094	0.787 ± 0.095	0.676 ± 0.093	0.761 ± 0.092	0.650 ± 0.090
				Chest				
ECG	0.862 ± 0.085	0.831 ± 0.083	0.836 ± 0.085	0.805 ± 0.083	0.751 ± 0.082	0.640 ± 0.080	0.725 ± 0.082	0.614 ± 0.080
EMG	0.710 ± 0.101	0.679 ± 0.099	0.684 ± 0.101	0.653 ± 0.099	0.532 ± 0.107	0.421 ± 0.105	0.506 ± 0.107	0.395 ± 0.105
EDA	0.821 ± 0.080	0.790 ± 0.078	0.795 ± 0.080	0.764 ± 0.078	0.671 ± 0.040	0.561 ± 0.038	0.645 ± 0.040	0.535 ± 0.038
RESP	0.841 ± 0.036	0.801 ± 0.034	0.815 ± 0.036	0.775 ± 0.034	0.668 ± 0.035	0.558 ± 0.033	0.642 ± 0.035	0.532 ± 0.033
TEMP	0.698 ± 0.024	0.668 ± 0.022	0.672 ± 0.024	0.642 ± 0.022	0.480 ± 0.081	0.370 ± 0.079	0.454 ± 0.081	0.344 ± 0.079
ACC	0.802 ± 0.132	0.771 ± 0.130	0.776 ± 0.132	0.745 ± 0.130	0.699 ± 0.134	0.588 ± 0.132	0.673 ± 0.134	0.562 ± 0.132
Chest Physiological	0.934 ± 0.055	0.903 ± 0.057	0.928 ± 0.035	0.897 ± 0.033	0.824 ± 0.095	0.713 ± 0.093	0.798 ± 0.095	0.687 ± 0.093
Chest All	0.942 ± 0.048	0.919 ± 0.047	0.938 ± 0.0520	0.910 ± 0.051	0.837 ± 0.068	0.726 ± 0.067	0.811 ± 0.065	0.700 ± 0.062
				Wrist + Chest				
All Physio	0.934 ± 0.045	0.903 ± 0.047	0.931 ± 0.045	0.900 ± 0.043	0.861 ± 0.065	0.750 ± 0.063	0.859 ± 0.065	0.750 ± 0.063
All Modalities	0.941 ± 0.047	0.921 ± 0.048	0.941 ± 0.047	0.921 ± 0.048	0.863 ± 0.045	0.752 ± 0.042	0.863 ± 0.045	0.752 ± 0.042

Table 5.1: Performance evaluation of classifiers using LOSO-CV on the WESAD dataset for both binary classification (baseline vs. stress) and three-level classification (baseline vs. stress vs. amusement) across various modality combinations. The fixed model was inferred using the same sensors as in training, while the generalized modular model, trained on all ten modalities, leveraged its modular capabilities to infer with different modality combinations, as explained in Section 4.4.1.

to when all modalities are used, neglecting the need for wrist-based wearables in combination with chest wearables. Nevertheless, BVP demonstrates to be as effective as ECG for a single based feature, undermining the large discrepancy found in the Gini importance scores of Figure 5.7).

The difference between the physiological-based results and the complete results for wrist, chest and all modalities, are due to the accelerometer. In several papers, acceleration data has mistakenly been used to classify stress due to its "ability to capture physical changes such as trembling sensations and shaking" [131]. It is more plausible that the wrist accelerometer readings reflect the gestures made during the speech of the TSST, rendering this modality a data snooping tool. Nevertheless, accelerometer readings may be utilised by the model as a noise contextualisation signal to disregard other signals such as BVP, which was previously shown in Figure B.5 to cause motion artefacts, and could be misleading. This approach was also adopted in two other studies: Rashid et al. employed a data fusion gating process that incorporates the acceleration signal [132], while the SELF-CARE study used acceleration data alongside other sensors in an ensemble branch method [22].

	Binary Classification	**Three-Level Classification**
Average Accuracy Downgrade	0.0199	0.0225
Average F1 Score Downgrade	0.0202	0.0223

Table 5.2: Expected compromise in performance for utilizing the generalized modular model rather than the fixed model for the WESAD dataset combinations tested in Table 5.1.

A minimal decrease in model performance is observed in unimodal classification tasks when using the generalized modular model compared to the fixed model, primarily due to the application of unimodal fine-tuning. This reveals that the modular model does not lose significant information or context when the network is further fine-tuned on other modalities, which could potentially re-encode the output of the cross-attention network and thus the weights of the self-attention network would adapt to this new output. Instead, the cross-attention network successfully learns to maintain an intermediate representation that will translate well no matter which combination of modality is used, validating the fine-tuning method described in Section 4.5.2. The compromises that are associate with selecting the generalised modular model for prediction are summarised in Table 5.2.

	Binary Classification				**Three-Level Classification**			
	Fixed Model		**Generalized Model**		**Fixed Model**		**Generalized Model**	
Modality	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
BVP	0.830 ± 0.068	0.821 ± 0.067	0.821 ± 0.061	0.810 ± 0.067	0.690 ± 0.082	0.683 ± 0.080	0.684 ± 0.075	0.673 ± 0.072
EDA	0.842 ± 0.065	0.830 ± 0.064	0.835 ± 0.055	0.827 ± 0.054	0.704 ± 0.048	0.697 ± 0.049	0.673 ± 0.058	0.697 ± 0.054
BVP + EDA	0.862 ± 0.059	0.849 ± 0.061	0.862 ± 0.059	0.849 ± 0.061	0.748 ± 0.085	0.735 ± 0.089	0.748 ± 0.085	0.735 ± 0.089

Table 5.3: Performance evaluation of classifiers using LOSO-CV on the UBFC-Phys dataset for both binary classification (baseline vs. stress) and three-level classification (baseline vs. social stress vs. cognitive stress) across various modality combinations.

The UBFC-Phys dataset demonstrates comparable performance between EDA and BVP when inferred independently. Symbiosis of the two modalities are effective, further demonstrating the fusion capabilities of the cross-attention network in the model. For binary classification, the generalized modular model shows a 2.7% improvement, with an even larger 7.5% increase when distinguishing between different the stress conditions in the three-level classification task.

	Binary Classification				**Four-Level Classification**			
	Fixed Model		**Generalized Modular Model**		**Fixed Model**		**Generalized Modular Model**	
Modality	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
Wrist-based								
BVP	0.820 ± 0.214	0.768 ± 0.162	0.788 ± 0.182	0.746 ± 0.140	0.536 ± 0.123	0.510 ± 0.144	0.524 ± 0.111	0.491 ± 0.125
EDA	0.806 ± 0.107	0.767 ± 0.115	0.790 ± 0.091	0.729 ± 0.077	0.604 ± 0.158	0.592 ± 0.130	0.588 ± 0.142	0.570 ± 0.108
ACC	0.811 ± 0.143	0.749 ± 0.117	0.784 ± 0.116	0.714 ± 0.082	0.691 ± 0.133	0.675 ± 0.141	0.676 ± 0.118	0.657 ± 0.123
Wrist All	0.866 ± 0.221	0.847 ± 0.143	0.837 ± 0.192	0.828 ± 0.124	0.682 ± 0.124	0.667 ± 0.127	0.646 ± 0.088	0.653 ± 0.113
Chest-based								
ECG	0.880 ± 0.094	0.851 ± 0.116	0.865 ± 0.079	0.825 ± 0.090	0.576 ± 0.131	0.559 ± 0.136	0.538 ± 0.093	0.534 ± 0.111
IBI	0.799 ± 0.058	0.752 ± 0.071	0.777 ± 0.036	0.724 ± 0.043	0.447 ± 0.058	0.442 ± 0.061	0.408 ± 0.019	0.427 ± 0.046
ACC	0.703 ± 0.063	0.650 ± 0.068	0.674 ± 0.034	0.625 ± 0.043	0.505 ± 0.093	0.481 ± 0.087	0.479 ± 0.067	0.451 ± 0.057
Chest All	0.896 ± 0.095	0.872 ± 0.105	0.880 ± 0.079	0.847 ± 0.080	0.612 ± 0.118	0.591 ± 0.126	0.588 ± 0.094	0.561 ± 0.096
Muscular-based								
Upper Trapezius EMG	0.743 ± 0.077	0.652 ± 0.101	0.730 ± 0.064	0.620 ± 0.069	0.453 ± 0.110	0.408 ± 0.108	0.414 ± 0.071	0.373 ± 0.073
Mastoid EMG	0.755 ± 0.105	0.684 ± 0.134	0.727 ± 0.077	0.644 ± 0.094	0.482 ± 0.197	0.438 ± 0.183	0.465 ± 0.180	0.400 ± 0.145
EMG All	0.753 ± 0.079	0.666 ± 0.102	0.717 ± 0.043	0.641 ± 0.077	0.455 ± 0.105	0.406 ± 0.101	0.436 ± 0.086	0.393 ± 0.088
Brain-based								
FNIRS	0.858 ± 0.071	0.824 ± 0.090	0.820 ± 0.031	0.807 ± 0.073	0.612 ± 0.106	0.557 ± 0.114	0.580 ± 0.074	0.542 ± 0.099
All Sensors								
All physiological (w/o ACC)	0.872 ± 0.292	0.852 ± 0.194	0.838 ± 0.258	0.833 ± 0.175	0.618 ± 0.144	0.606 ± 0.168	0.584 ± 0.110	0.565 ± 0.127
All non-muscular (w/o EMG)	0.847 ± 0.287	0.836 ± 0.181	0.816 ± 0.256	0.807 ± 0.152	0.614 ± 0.135	0.604 ± 0.146	0.589 ± 0.110	0.586 ± 0.128
All non-motion-artefacts (w/o ACC + EMG)	0.835 ± 0.310	0.827 ± 0.209	0.797 ± 0.272	0.793 ± 0.175	0.622 ± 0.152	0.603 ± 0.164	0.609 ± 0.139	0.587 ± 0.148
All modalities	0.883 ± 0.088	0.883 ± 0.087	0.883 ± 0.088	0.883 ± 0.087	0.697 ± 0.091	0.691 ± 0.095	0.697 ± 0.091	0.691 ± 0.095

Table 5.4: Performance evaluation of classifiers using LOSO-CV on the MUSED dataset for both binary classification (baseline vs. stress) and four-level classification (baseline vs. anticipatory stress vs. social stress vs. mental stress) across various modality combinations.

In contrast to the evaluation of WESAD, the ECG signal holds significantly larger prediction capabilities to BVP for this study, which leads to two possibilities: either this is due to poor contact amongst multiple subjects in the MUSED study, or the WESAD ECG features are more difficult to derive further features, i.e. PQRST identification.

Nevertheless, the WESAD dataset is comparable to the MUSED dataset in terms of accelerometer recordings; both datasets exhibit a significant impact of acceleration signals on classification tasks when these signals are included or excluded. For the same reasoning discussed previously for acceleration as a data snooping tool, we also exclude non-muscular modalities (mastoid and upper trapezius sEMG), since particularly during the interview task, significantly more motion and thus muscular activity is observed. Excluding sEMG is analogous to excluding the accelerometer signal, raising a similar question regarding the distinction between data snooping and noise contextualisation. Further investigation is needed to fully assess this distinction.

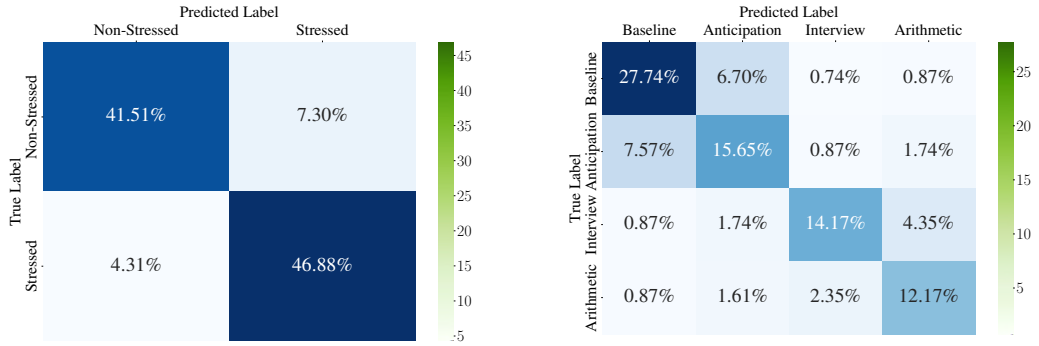


Figure 5.10: Distribution of classification labels for the MUSED dataset for a) binary b) four-level classification.

To gain further insights into how the protocol induced different stress responses, and how difficult it is to distinguish these differences, the confusion matrix is plotted in Figure 5.10. For binary classification, we observe that the model is biased toward predicting the stress condition, the majority class. This is likely attributed to the class imbalance, and could suggest that the γ value is too small in the focal loss function described in Equation 4.16. Additionally, it may indicate that some participants demonstrated stress symptoms from the virtual reality experience, although this was not reflected in the self-questionnaires. For four-level classification, the distribution shows that the anticipation condition was often misclassified as the baseline condition, rather than the stress condition. Further investigation could determine accuracies for stress detection without the anticipation task. Similarly, it was often difficult to distinguish the arithmetic from the interview condition, and thus determine the nuance of social versus cognitive stress.

5.3.2 Latency Performance

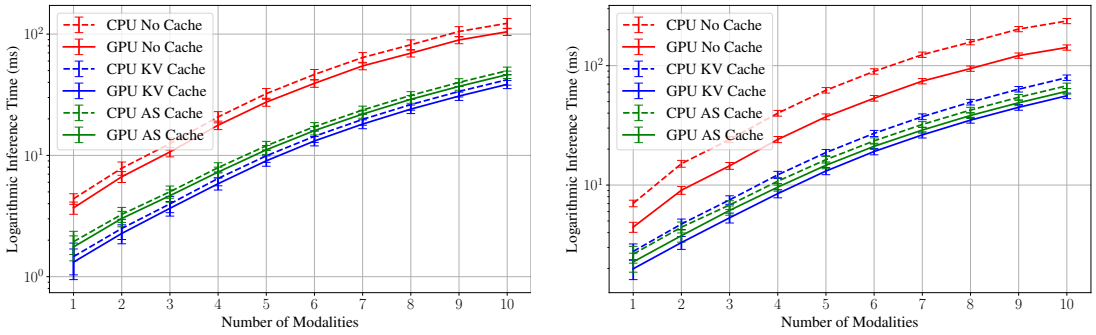


Figure 5.11: Computational performance comparison across different caching methods for the generalized model on two hardware configurations: (a) Machine A: Intel® Core™ i9-10900 CPU @ 2.80GHz × 20 with NVIDIA GeForce RTX 3090, 64GiB DDR4 Synchronous 2133MHz RAM, and (b) Machine B: Intel® Core™ i7-10510U CPU @ 1.80GHz × 8 with NVIDIA GeForce MX250, 12GiB DDR4 2400 Synchronous 1200Hz RAM.

The computational performance of the generalized modular model was assessed on two devices with different computational and memory capabilities. For Machine A, the KV cache consistently achieves the lowest inference time across all modality configurations, with its CPU implementation ranking second lowest. The minimal discrepancy between CPU and GPU performance suggests that the GPU is not fully optimized for parallel tasks. This is likely due to the low-level operations described in Section 4.4.2 which are not well-optimized for parallelization with CUDA optimizations, unlike the linear layers (embedding and hidden layers) which are fully optimized through PyTorch. Additionally, the inference process may be more constrained by memory speeds than by computational power. The KV cache proves to be more efficient than the Attention Score (AS) cache, likely due to the higher memory demands associated with caching and then retrieving the projections of keys, values, and queries, as well as attention score caching from RAM.

For Machine B, similar trends are observed regarding CPU and GPU performance, though with a slightly larger gap between them. Notably, attention score caching on both CPU and GPU outperforms KV caching on the CPU. This indicates that for this device, the limiting factor is nearly equal between memory retrieval and compute. It also suggests that with higher-performing memory read/write speeds relative to its computational capabilities, further reductions in inference time might be achievable. However, additional research is necessary to explore this hypothesis and determine the extent of potential improvements.

5.3.3 Comparison with Relevant Literature

We compare our results against other methods illustrated in the literature for both the established datasets tested - WESAD and UBFC-Phys. We aim to showcase as many combinations of modalities as possible to evaluate the effectiveness of the generalizability performance of our modular model.

Paper	Modalities	Accuracy	F1 Score	Window Size (s)	Modular
		Wrist			
[133]	EDA	0.875	-	60	✗
[134]	EDA	0.929	0.890	60	✗
[135]	EDA	0.958	0.951	-	✗
OURS (Fixed)	EDA	0.794	0.764	6	✗
OURS (Generalized)	EDA	0.770	0.740	6	✓
[136]	BVP	0.993	0.99	210	✗
[137]	BVP	0.908	0.86	15	✗
[137]	BVP	0.951	0.928	30	✗
OURS (Fixed)	BVP	0.863	0.832	6	✗
OURS (Generalized)	BVP	0.837	0.806	6	✓
[93]	BVP, EDA	0.721	0.67	30	✗
[137]	BVP, EDA	0.977	0.967	30	✗
OURS (Fixed)	BVP, EDA	0.933	0.921	6	✗
OURS (Generalized)	BVP, EDA	0.925	0.901	6	✓
[138]	Wrist All	0.868	0.970	-	✗
[98]	Wrist All	0.931	-	60	✗
[22]	Wrist All	0.941	0.929	60	✗
OURS (Fixed)	Wrist All	0.903	0.872	6	✗
OURS (Generalized)	Wrist All	0.877	0.846	6	✓
		Chest			
[16]	ECG	0.854	0.813	60	✗
[86]	ECG	0.918	0.905	60	✗
[137]	ECG	0.978	0.972	15	✗
[137]	ECG	0.983	0.976	30	✗
OURS (Fixed)	ECG	0.862	0.831	6	✗
OURS (Generalized)	ECG	0.836	0.805	6	✓
[97]	ECG, EDA	0.921	0.911	10	✗
[137]	ECG, EDA	0.994	0.992	30	✗
OURS (Fixed)	ECG, EDA	0.912	0.897	6	✗
OURS (Generalized)	ECG, EDA	0.868	0.852	6	✓
[16]	Chest All	0.928	0.911	60	✗
OURS (Fixed)	Chest All	0.942	0.919	6	✗
OURS (Generalized)	Chest All	0.938	0.910	6	✓
		Wrist and Chest			
[139]	Wrist + Trans. Chest	0.921	0.897	60	✗
[86]	All modalities	0.918	0.977	60	✗
[87]	All modalities	0.948	0.950	-	✗
OURS (Fixed/Generalized)	All modalities	0.941	0.921	6	✓

Table 5.5: Performance comparison of our proposed fixed and modular methods with existing methods from the literature on the WESAD database.

In the wrist modality, our proposed fixed and generalized models achieve competitive accuracy and F1 scores, particularly when using a combination of EDA and BVP signals, demonstrating substantial improvements in modularity. However, methods such as the SELF-CARE method outperform both models using ensemble late fusion method and a Kalman filter to take advantage

of the temporal dynamics in the stress. However, all branches were trained and this is not feasible for scaling to ten modalities, which our modular approach can. Our model consistently achieves high performance despite the shorter window sizes (6 seconds) compared to many other methods, which rely on larger windows, (typically between 30-60 seconds), reflecting the model's desirability for compute limited detection.

The chest modality also exhibits strong results, with both fixed and generalized versions achieving close to or exceeding 0.9 accuracy and F1 scores, however, not competing to sophisticated methods tailored for ECG such as [137]. Nonetheless, our generalized model achieves comparable results to that of [16], whilst maintaining modularity and surpasses this when the fixed model is used.

When integrating wrist and chest modalities, our method performs well with high accuracy and F1 score, however even the fixed model cannot match the results obtained by [86] which utilizes a residual-Temporal Convolutional Network (Res-TCN) and employs multi-level fusion (sensor-level, feature-level, and decision-level) with a Bayesian Network. This is testament to their method, and further showcases the power of multi-level fusion, to leverage the benefits from combining features early and late. Nonetheless, this method is very computationally expensive since all features are automatically extracted by the Res-TCN. Another high performing model for all modality fusion is [87] which instead undergoes early fusion through model fuses multivariate time-series data from multiple physiological sensors by converting them into a unified Gramian Angular Field (GAF) image, capturing temporal correlations between sensor modalities, and then this image is inferred by a convolutional neural network. However, like the other method discussed, automatic feature extraction is an expensive and inefficient approach. Thus, overall, the findings suggest that the proposed modular approach delivers competitive and efficient performance while maintaining flexibility across sensor types, which no other method offers.

5.4 Ablation Study

To evaluate the impact of each component in our proposed model architecture, an ablation experiment was conducted using the WESAD dataset, since it includes the most modularities out of the datasets tested. The model variants are described below:

1. *Self-Attention Only:* Only the self-attention network is used, and the cross-attention network is removed. Therefore, the model relies on late fusion only to fuse the modalities together. The proposed bidirectional encoder-decoder attention mechanism is used.

Paper	Modalities	Accuracy	F1 Score	Window Size (s)	Modular
[17]	BVP	0.855	-	30	✗
[140]	BVP	0.820	-	-	✗
[137]	BVP	0.922	0.943	15	✗
[137]	BVP	0.922	0.945	30	✗
Ours (Fixed)	BVP	0.830	0.821	6	✗
Ours (Generalized)	BVP	0.821	0.81	6	✓
[102]**	BVP, EDA	0.758	0.756	30	✗
[23]**	BVP EDA	0.769	0.768	30	✗
[93]*	BVP, EDA	0.818	0.817	30	✗
Ours (Generalized)	BVP, EDA	0.862	0.849	6	✓

Table 5.6: Performance comparison of our proposed fixed and modular methods with existing methods from the literature on the UBFC-Phys database. * indicates that the binary classification was on the baseline vs the speech task and not the arithmetic task. ** indicates that these results were also for baseline vs speech task condition and the results were conducted by [93].

2. *Cross-Attention Only:* Only the cross-attention network is used, and the self-attention network is removed. Therefore, the model relies on attendance between modalities to obtain a latent stress representation. The proposed bidirectional encoder-decoder attention mechanism is used.

3. *Encoder Self-Attention:* This utilizes the complete model architecture proposed in Figure 4.1 but with encoder self-attention. This means that the five 6 second temporal slices are gathered into a batch (30s in total) and the model infers on the batch, as in [93].

4. *Encoder-Decoder Attention:* The complete model architecture proposed is used but employs encoder-decoder attention from the one new temporal slice to the four cached temporal slices each of 6 seconds; the model infers on one temporal slice.

5. *Bidirectional Encoder-Decoder Attention:* The complete model architecture proposed is used with the bidirectional encoder-decoder mechanism. Each new temporal slice attends to the four cached temporal slices each of 6 seconds, and each cached temporal slice attends to the new slice; the model infers on one temporal slice.

6. *Fine Tune:* This is the same model as the Bidirectional Encoder-Decoder Attention, however as explained in Section 4.5.3, due to the nuanced difference between the training and testing prediction tasks, the model was fine-tuned sequentially on a per-temporal slice basis.

As shown in Table 5.7, removing the cross-attention module and relying only on the predictor for late fusion between modalities causes the accuracy to suffer, observing a loss between 3.5%-8.9% in accuracy and 1.1%-7.0% for F1-score for the two classification tasks. Adhering to the findings of [86], this demonstrates the drawbacks from selecting a late fusion approach, and shows that multi-level fusion is able to leverage the benefits from combining features at an earlier stage in addition to a later stage by the use of branches acting as an ensemble for more robust classification.

The removal of the self-attention module, also degrades performance in comparison to the complete architecture models tested, ranging from 3.5%-8.9% in accuracy and 4.4-11.0% F1 score. The worsened performance suggests that the predictor used, in this case Average Pooling Branch Fusion, cannot sufficiently interpret the representation output from the cross-attention block; this

Model	Binary Classification		Three-Level Classification	
	Accuracy	F1 Score	Three_Level_Accuracy	F1 Score
Wrist All				
Self-Attention Only	0.832 ± 0.086	0.802 ± 0.082	0.724 ± 0.092	0.639 ± 0.069
Cross-Attention Only	0.788 ± 0.102	0.758 ± 0.103	0.671 ± 0.058	0.586 ± 0.099
Encoder Self-Attention	**0.879 ± 0.078**	**0.849 ± 0.107**	**0.766 ± 0.052**	**0.681 ± 0.090**
Encoder-Decoder Attention	0.858 ± 0.072	0.832 ± 0.079	0.743 ± 0.041	0.658 ± 0.048
Bidirectional Encoder-Decoder Attention	0.851 ± 0.081	0.821 ± 0.082	0.751 ± 0.060	0.666 ± 0.096
Fine Tune	0.867 ± 0.096	0.846 ± 0.094	0.761 ± 0.092	0.650 ± 0.090
Chest All				
Self-Attention Only	0.862 ± 0.082	0.832 ± 0.068	0.741 ± 0.044	0.656 ± 0.103
Cross-Attention Only	0.841 ± 0.122	0.791 ± 0.097	0.682 ± 0.090	0.597 ± 0.062
Encoder Self-Attention	0.930 ± 0.049	**0.907 ± 0.052**	**0.820 ± 0.070**	**0.735 ± 0.043**
Encoder-Decoder Attention	0.921 ± 0.059	0.891 ± 0.053	0.805 ± 0.040	0.720 ± 0.102
Bidirectional Encoder-Decoder Attention	0.932 ± 0.049	0.902 ± 0.063	0.807 ± 0.062	0.722 ± 0.044
Fine Tune	**0.938 ± 0.079**	0.902 ± 0.077	0.811 ± 0.103	0.700 ± 0.101
All modalities				
Self-Attention Only	0.891 ± 0.064	0.861 ± 0.107	0.781 ± 0.045	0.756 ± 0.076
Cross-Attention Only	0.852 ± 0.157	0.821 ± 0.086	0.752 ± 0.108	0.727 ± 0.106
Encoder Self-Attention	**0.950 ± 0.049**	0.920 ± 0.044	**0.866 ± 0.103**	**0.781 ± 0.076**
Encoder-Decoder Attention	0.933 ± 0.079	0.893 ± 0.102	0.850 ± 0.092	0.765 ± 0.099
Bidirectional Encoder-Decoder Attention	0.948 ± 0.092	0.900 ± 0.073	0.859 ± 0.095	0.750 ± 0.093
Fine Tune	0.941 ± 0.0478	**0.921 ± 0.079**	0.863 ± 0.105	0.752 ± 0.103

Table 5.7: Ablation study for the generalized modular model architecture using the WESAD dataset.

is particularly apparent when predicting on all modalities where the predictor fails to interpret the diverse feature representations from respective branches. The self-attention module therefore still establishes its purpose for highlighting the most important inter-modal feature representations and reducing the impact of noise, both in the features of the signal, and the noise introduced during the cross-attention process.

Overall, the model that employs encoder self-attention performs the best in most classification tasks, outperforming the other methods in terms of F1 scores for all three modality tests in both classification tasks. The reasoning why there is a discrepancy between the utilising local-bidirectional encoder-decoder attention is unclear, since the cached attention score should be able to reconstruct the attending from embeddings as if it all tokens had attended to one another collectively as in the encoder-self attention form. One hypothesis is that due to the individual temporal slices generated by the `SeqToSeqDataloader`, there will be samples, where the cached segments do not correspond to the same label during transitional tasks. This was only considered retrospectively, and further testing could efficiently test this hypothesis.

The deviation between the encoder-decoder attention and its bidirectional equivalent demonstrates the benefits of utilizing an attention mechanism which has the capability of re-evaluating, by means of attending, previous embeddings in light of new information. As we discussed in Section 4.4.2, the bidirectionality dismisses the need for consistent interpretations of the signal through forward-only attendance. The model is now able to re-evaluate, along the backward temporal

direction, previous samples which may take up new meaning due to the added context. Due to this, we observe an average improvement of 0.84% in accuracy and 0.79% in F1 score. In one case, the all wrist binary classification task, we identify a minor inverse effect - a degradation in performance due to the bidirectional component. This is likely due to adding too much emphasis to the new sample, leading to overfitting, which is compensated for by the encoder-decoder attention, acting as a regularizer to not re-evaluate. The fine-tuning of the model in a sequence-to-sequence manner results in an average improvement of 1.0% for accuracy and 0.6% for F1 score, confirming its purpose to learn the nuance between the train and test tasks.

5

5.5 Predictor Comparison

	Wrist Physiological		**Chest Physiological**		**All Physiological**	
Predictor	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
Hard Vote	0.858 ± 0.031	0.801 ± 0.031	0.918 ± 0.025	0.877 ± 0.026	0.908 ± 0.052	0.831 ± 0.052
Soft Vote	**0.869 ± 0.073**	0.839 ± 0.071	0.928 ± 0.035	0.897 ± 0.033	0.931 ± 0.075	0.900 ± 0.073
Avg Pooling BF	0.823 ± 0.191	0.801 ± 0.191	0.851 ± 0.193	0.842 ± 0.143	0.929 ± 0.105	**0.912 ± 0.103**
Max Pooling BF	0.795 ± 0.181	0.768 ± 0.181	0.853 ± 0.180	0.838 ± 0.180	0.920 ± 0.101	0.902 ± 0.0998
Self-Attention Pooling BF	0.861± 0.101	0.829 ± 0.101	0.861 ± 0.169	0.849 ± 0.167	**0.941 ± 0.124**	0.901 ± 0.125
Kalman Filter	0.861 ± 0.094	**0.842 ± 0.093**	**0.931 ± 0.035**	**0.899 ± 0.034**	0.927 ± 0.085	0.907 ± 0.083

Table 5.8: Performance evaluation of predictors on the generalized modular model for the WESAD dataset, using LOSO-CV. BF stands for branch fusion and corresponds to Figure 4.7 (c).

The chosen predictors were evaluated across three physiological data categories: Wrist, Chest, and All Physiological signals. Table 5.8 reveals clear patterns of how different fusion methods handle the task of multimodal stress detection, with the adaptive methods, that fuse later, generally outperforming the early branch fusion predictors.

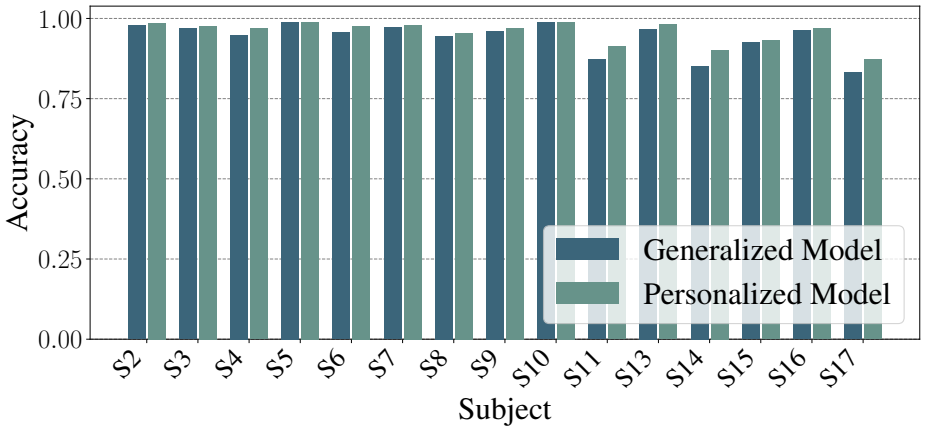
Hard voting, shows consistent but unremarkable performance, with an average accuracy of 0.895 and F1 score of 0.836; it is clear that the rigidity of treating each sensor equally and aggregating discrete class predictions limits its ability to capture the nuanced confidence levels inherent in different modalities. This is particularly relevant in physiological data, where certain sensors may provide more reliable information depending on the noise context. Soft Voting emerged as one of the top-performing predictors, with an average accuracy of 0.909 and F1 score of 0.879. This supports the findings of Zhang et al. who demonstrated performances that surpassed more complex predictors such as without BiLSTM, which was thought to overfit [93]. By aggregating the predicted probabilities rather than class labels, soft voting dynamically adjusts the importance of each sensor modality. This flexibility allows it to better handle variability across sensor streams, which is particularly beneficial in the personalized setting, to combat the “non-responders” discussed earlier.

Branch fusion pooling methods, specifically average and max pooling, show comparatively weaker results. Average pooling achieves an average accuracy of 0.868 and F1 score of 0.852, while max pooling performed the worst overall, with an average accuracy of 0.856 and F1 score of 0.836. The branch fusion predictor, which aggregate features across temporal slices via early pooling, leads to too small dimensionality reduction. However, the self-attention pooling variant demonstrated strong results, particularly when applied to the fused physiological data, achieving an accuracy of 0.941 in the All Physiological category. This method outperforms the other pooling techniques since the attention mechanism is able to attend to features more widely, prioritizing important features effectively, particularly for numerous modalities.

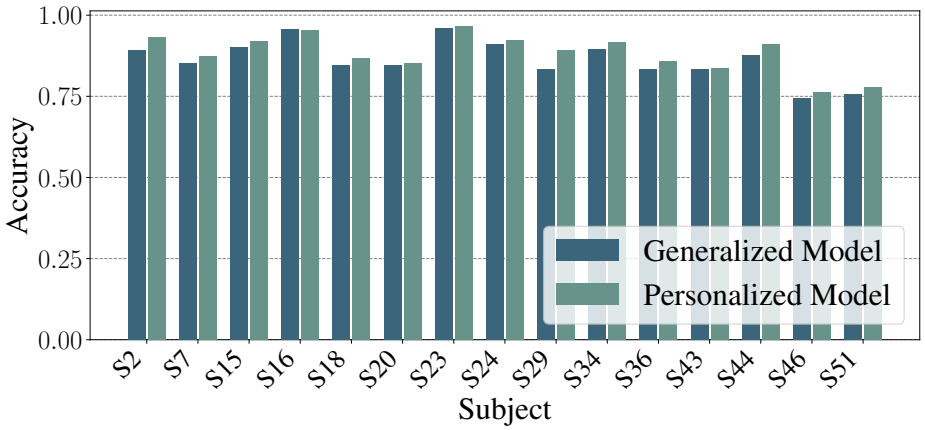
The Kalman filter achieves similarly high performance, with an average accuracy of 0.906 and the highest average F1 score of 0.883. The Kalman filter allows for temporal dynamics to be considered and handle noise between modalities robustly. However, as demonstrated in the ablation study, since the cross-attention modality successfully integrates dynamic alignment, the Kalman filter does not exhibit the same level of effectiveness as in the SELF-CARE study [22]. Nonetheless, it serves as a lightweight alternative to the branch fusion techniques, which have a larger parameter count and energy usage. Additionally, the Kalman filter could adapt well under personalized model conditions, where variables like the state transition matrix, measurement matrix, and covariance parameters are learned during training, though this was not tested in the study.

5.6 Personalization Evaluation

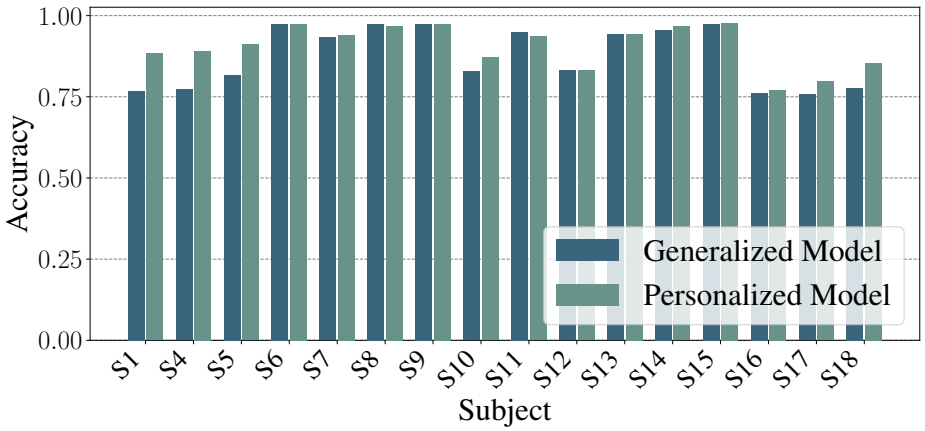
From Figure 5.12, our personalized model, evaluated across three datasets (WESAD, UBFC-Phys, and MUSED), demonstrates moderate but varying degrees of improvement. On average, personalized models outperformed generalized ones by 1.59% for WESAD, 1.99% for UBFC-Phys, and 3.14% for MUSED, with standard deviations of 1.48%, 1.50%, and 4.33%, respectively. MUSED showed the greatest potential for personalization but also exhibited the highest variability, as reflected in its larger standard deviation. This suggests either significant inter-participant variability, potentially due to subject differences or the reliability of the stress-inducing protocol, or an overfitting issue with the generalized model, limiting its ability to generalize effectively. In contrast, WESAD and UBFC-Phys displayed more consistent yet smaller gains. These findings indicate that while personalization generally enhances performance, the extent and consistency of improvement are influenced by dataset-specific factors.



(a)



(b)



(c)

Figure 5.12: Comparison of generalized and personalized model performance per subject, evaluated using LOSO-CV across three datasets: (a) WESAD, (b) UBFC-Phys, and (c) MUSED.

	Generalized Modular Model	Personalized Modular Model
Number of Parameters	193,300	140,500
CPU Inference Time (ms)	49.884 ± 3.555	41.354 ± 3.157
GPU Inference Time (ms)	46.283 ± 3.223	38.646 ± 2.980

Table 5.9: Model parameter reduction and computational improvements for the personalized model. Tested on Machine A (NVIDIA GeForce RTX 3090).

Table 5.9 presents the computational gains achieved by reducing the hidden dimension of the personalized model. The parameter size was reduced by 27.3% from the generalized model, resulting in a 17.1% and 16.5% decrease in inference time. Although this reduction in model size does not reach the 70% achieved by Yu et al. [23] for their attention-based model, the combination of improved accuracy and reduced computational demands demonstrates the value of employing personalized models.

5

6

Discussion

Contents

6.1 MUSED Dataset for Multimodal Stress Detection

The MUSED dataset, offers novel insights into the effectiveness of monitoring mastoid sEMG and EEG activity, in addition to the effectiveness of a commercial single-channel fNIRS device placed on the pre-frontal cortex.

We collect data from 18 participants, which should allow for a dataset that produces significant inter-participant variability and response to the stressors in the study. However, Vos et al. [141] examine whether models trained on datasets with fewer than 50 participants can generalize effectively to new data, highlighting the importance of adequate sample size for reliable research. While smaller samples are common in stress datasets [16], [142], [143], achieving 80% statistical power often requires over 50 participants when an effect size of 0.4 is used. While this was not

feasible in this study, if more statistical power is required, due to its numerous modalities, it may be concatenated with other datasets like previous stress detection research [141], [144].

Through the statistical evaluation of the questionnaires, we can conclude that stress was reliably induced in most participants, with the SSSQ scores indicating that the subjects felt more engaged and worried than distressed during the TSST task. This is the intended response, and believe that it further demonstrates the effectiveness of a VR-based TSST. However, we did observe significant variance between subjects, and ones that were higher than in the other datasets used, this was supported by the increased standard deviations shown in Table 5.4 and the personalization performance results shown in Figure 5.12. Through qualitative feedback and via the dataset prediction, we observe that the anticipation task was difficult to distinguish and may not have had the intended effect. Additionally, it was often difficult to distinguish between the social and cognitive test, and more research is required to identify if these two types of stress can be differentiated.

A difficult element of the findings from the MUSED dataset, and one that has not resolved, is the distinction of the accelerometer and sEMG signals, which may act in combination as a data snooping and noise contextualisation tool. We observe that in other datasets and subsequent benchmarking papers, acceleration data has been used to classify stress [16], [22], [131] without the consideration that it could be used as a data snooping tool. Since we frequently observed wrist gestures made during the speech task of the TSST, we believe some there is the possibility for data snooping, though it could also be argued that motion during these tasks propagated to other sensor artefacts such as BVP. However, the accelerometer signal could be utilised for noise contextualization, as shown in Rashid et Al. and SELF-CARE [22], [132]. This is why throughout our extensive experiments, we ensured that we tested the model underwhich no acceleration data (i.e. all physiological) and no muscular data was used. However, as in the WESAD or UBFC-Phys study, we did not control the possibility that sweat and thus EDA skin conductance could increase throughout the study irrespective of the stress condition. Such attributes may be part of the reason why existing stress detection models often fail to generalize on new, unseen stress datasets, even between protocol-based datasets [141].

6.2 Consumer Mastoid-based sEMG and fNIRS Feasibility

6.2.1 Mastoid Feasibility

From the analysis of the pilot study, and the MUSED dataset, it is evident that the feasibility of using consumer-grade mastoid sEMG and single-channel fNIRS devices for stress detection presents both opportunities and challenges. From a technological standpoint, the integration of these modalities into a consumer-friendly device, such as headphones, offers a promising avenue for continuous, non-invasive monitoring. However, there remain significant uncertainties regarding their capabilities, particularly when compared to more established, multichannel laboratory systems for recording sEMG, EEG and fNIRS.

Regarding mastoid sEMG, our study suggests that muscle activity can be detected from the mastoid region, though the amplitude of the signal is notably lower than that recorded from the upper trapezius muscle. While some features, such as MNF (Mean Frequency) and RMS (Root Mean Square), have shown promise in distinguishing between stress and non-stress conditions, the lack of statistical significance in these results suggests high variability across participants, which could be attributed to placement, inter-participant variability and noise conditions. Some individuals exhibited greater muscular activation under stress, though the majority did not display a notable response, often due to noise limiting the reliability of sEMG for stress detection in a broad user base. One technique to distinguish muscular activity from motion artefacts and other noise is done by monitoring the asymmetry between left and right upper trapezius muscles [128]. Nevertheless, as noise is a prevalent issue in these measurements, further refinement of the signal processing techniques and more rigorous testing are necessary to validate the robustness of this modality in consumer devices.

For EEG monitoring via the mastoid region, pilot data indicates that alpha wave activity, often associated with relaxation, can be detected at this site. However, despite visible trends in alpha and beta power changes under stress conditions, the changes are not statistically significant. This raises questions about the consistency and reliability of EEG-derived stress biomarkers when measured from the mastoids, and favours the reliability but poor practicality of multichannel, EEG demonstrated to be robust in numerous stress monitoring studies [15], [145] or devices such as the one developed by Microsoft which use in-ear EEG electrodes [10]. Additionally, the overlap in features observed between the mastoid and upper trapezius regions suggests that non-EEG factors,

such as noise and muscular artefacts, may be influencing the readings. Neurable's 20-channel EEG-based headset claims that its signal processing techniques and support vector machine model, which is used to detect alpha waves for focus estimation, effectively eliminate sEMG artefacts, according to the findings reported by Enten et al. [11]. Nonetheless, while feasible near the mastoid region has been previously demonstrated, the accuracy and reliability of bipolar mastoid EEG as a standalone stress detection tool remain uncertain.

6.2.2 fNIRS Feasibility

The inclusion of single-channel fNIRS in this system is an exciting development, especially given the underexplored potential of this modality for stress detection. However, prior research has relied on multichannel systems for stress detection [14], [146], making it difficult to predict the efficacy of a single-channel device in capturing stress-related changes with sufficient resolution. While the commercial fNIRS device from SilverLine Research provides a convenient and user-friendly form factor, further validation studies are essential to determine its sensitivity and accuracy in detecting stress, particularly in real-world environments.

As observed by several subjects' fNIRS signals, the single-channel fNIRS device demonstrates reliable signal quality for detecting changes in brain activity within the prefrontal cortex, especially during cognitively demanding tasks such as interviews and arithmetic. This finding underscores the device's ability to measure hemodynamic responses associated with increased cognitive load. However, as observed in the data from Subjects 16 and 17, motion artefacts introduce prolonged recovery periods in the hemodynamic features, compromising signal quality and excessive sweat also degrades reliability. This reinforces the argument made by Nozawa et al. [146] during their investigation into prefrontal activity for detecting different neural activities for every-day use. They argue that through extensive signal preprocessing methods, particularly emphasizing the removal of physiological noise, it is possible to capture individual difference in cognitive tasks [146].

Additionally, some participants, referred to as non-responders, exhibit no significant changes in brain oxygenation during the stress conditions. This variation could be due to individual differences in how stress manifests in terms of cognitive effort or brain haemodynamics such as physiological factors like extracranial blood flow [130], highlighting a limitation in the universality of fNIRS as a standalone stress biomarker. Though this is contested by Nozawa et al. who demonstrate they can capture functionally relevant individual differences in prefrontal neural activity during cognitive tasks.

Oxygenated hemoglobin (O2Hb) and deoxygenated hemoglobin (HHb) are critical features for stress detection using fNIRS. As shown in qualitative signal inspections, stress conditions generally result in a slight increase in O2Hb and brain oxygenation levels, while HHb decreases. Despite the larger net difference in O2Hb, the stress condition is associated with wider variance in the data. This variance highlights both the sensitivity of the O2Hb signal and the challenges of detecting stress consistently across participants. However, heart-rate derived features such as mean and max O2Hb_HR do demonstrate statistical significance and the Gini importance analysis on a sensor and feature level basis showing promise for fNIRS in stress detection tasks. The high inter-participant variability in fNIRS signals, demonstrated in both visual and quantitative analyses, significantly impacts classification performance. Binary classification accuracy varies from 0.743 to 0.954, underscoring the challenges posed by individual differences.

fNIRS shows promise for stress detection, however its effectiveness is limited by signal reliability issues, motion artefacts, and inter-participant variability, leading to inconsistent classification accuracy. To overcome these challenges, integrating fNIRS with other modalities in a multimodal system can enhance the detection of robust stress biomarkers, compensating for individual limitations. Modular predictors that adjust sensor selection based on participant characteristics and noise environments should significantly improve system reliability, and move one step closer to a consumer product.

6.3 Modular and Personalized Biosignal Architecture

6.3.1 Modularity

The architecture presented in this study leverages a modular approach, crucial for multimodal stress detection. Modularity is vital for addressing the limitations of traditional fusion techniques such as data imputation and decision-fusion. We confirm the argument made by Naegelin et al. [18] that multimodal approaches are critical for robust stress detection. Yet, multimodal methods still struggle with missing or noisy data [19], [22], relying on all sensors to be present to make a prediction. The modular design of the proposed architecture aims to overcome these challenges by decoupling dependencies between sensors, ensuring flexibility across different data modalities. We demonstrate that the key features of the BCSA introduced by Zhang et al. [93], that is the temporal alignment of biosignal features across modalities which act symbiotically, can be repurposed for a modular multimodal fusion model.

The modular aspect of the BCSA mechanism demonstrates easy integration or removal of modalities. This adaptability is crucial for stress detection, where sensors may malfunction or data from specific modalities might be unavailable. For instance, if a wrist-based PPG device is not worn by a user, the model can still perform accurately using data from other modalities, such as if they were using headphones with mastoid sEMG integration. This flexibility is demonstrated by the minimal decrease in performance when classifying stress from a single modality compared to the full multimodal configuration, which shows the effectiveness of the single-modal fine-tuning. The compromise of selecting the generalized modular model over the fixed model that is trained exclusively on the modalities it infers on, is a 2% sacrifice in performance.

The BCSA approach is comparable to Han et al.'s Cascading Modular Multimodal Cross-Attention Network (CMMCN) [107], in that it refines multimodal interactions across deeper layers to enhance joint representations. We observe the same, where the cross-attention network successfully learns to construct an intermediate representation that translates between any combination of the ensemble modality branches. This temporal alignment of information ensures that stress detection from different modalities is handled robustly between combinations, and learns the alignment that is required to deal with the asynchronicity between biomarkers of stress, reducing the reliance on one particular signal and increasing robustness across a wide range of conditions. Similarly, the cross-attention mechanism, as used in Shi et al.'s co-attention network [106], which facilitates the interaction between modalities, particularly in cases where one modality serves as a central information source. In the same way, our cross-attention network relies more on some features than others.

The late-fusion predictors employed in the architecture further enhance its adaptability. By handling embeddings produced by cross-attention networks from multiple sensors, these predictors offer a modular decoding process that allows the system to flexibly integrate inputs from various sources without requiring retraining of the entire network. This ensures that stress detection is accurate even when faced with incomplete or missing sensor data, though further validation is still required, particularly in real-world conditions.

6.3.2 Personalization

The model architecture designed for stress detection incorporates personalization techniques to address the inter-participant variability widely documented in stress monitoring research. This variability, underscores the importance of personalizing models to capture individual nuances in

stress responses, which generalized models fail to detect. By employing personalization mechanisms at multiple stages of the model, we enhance the detection accuracy while reducing computational overhead.

The performance upgrades from the personalized model demonstrated to improve the performances across the three datasets tested, particularly for some subjects in the MUSED dataset, who were difficult to classify using the generalized model. We achieved comparable personalized performance to that of , who focused their research on addressing personalization. The reduced computational demands demonstrates the value of employing personalized models, though we could not match the parameter saving achieved by the personalized attention network of Yu et al. [23], since further reduction would lead to underfitting, likely due to the added cross-attention network which increases model complexity. These improvements, though less dramatic in parameter reduction, effectively balance performance and efficiency, emphasizing their use case for real-life applications.

6.4 Early-Onset Stress Detection Capabilities

The proposed Sliding Attention Score Caching Mechanism, integrated with a novel bidirectional encoder-decoder attention, demonstrates a promising advancement in the real-time detection of stress from multimodal biosignals. Unlike traditional stress detection models, which rely heavily on static representations of past signals, this approach dynamically reevaluates past inputs in light of new information. This bidirectional attention mechanism enables the model to reinterpret previous signals when new biosignal data becomes available, thus improving its adaptability and accuracy in identifying stress biomarkers, especially under fluctuating conditions.

The caching mechanism draws inspiration from the transformer model used in text generation, particularly its key-value (KV) caching methodology. In text generation, KV caching prevents the recomputation of tokens during the autoregressive process, yielding substantial computational savings. We apply this concept to biosignals and, to the best of our knowledge, no other research paper implements or discusses the computational savings of KV cache for real-time biosignal classification tasks. We take this one step further, by implementing query projections, enabling attention score caching, and reducing the computational complexity significantly. The Sliding Attention Score Caching Mechanism retains the history of projected keys, values, queries and attention scores (AS), allowing the model to continually recontextualize past signals as new data streams in.

This architecture addresses the two key challenges in early-onset stress detection identified by Alberdi et al. [24]: the need for high temporal resolution and the ability to capture subtle variations between features. By utilizing KV caching in the encoder, rather than in the decoder as traditionally implemented in Transformer architecture, the model can handle the dynamic, time-sensitive nature of biosignals more effectively. The bidirectional attention mechanism further allows for a more holistic understanding of the evolving stress response, reinterpreting historical data in light of newly detected patterns. This is verified by our ablation study, where we see an improved accuracy relative to the encoder-decoder attention used in the KV cache implementation.

In evaluating the effectiveness of the Sliding Attention Score Caching Mechanism, two key metrics were considered: inference speed and classification accuracy. Rashid et al. [132], emphasize the importance of computational efficiency in real-time stress detection systems. To evaluate its suitability for real-time classification, since few papers have published results on inference times performances on these datasets, and since it varies considerably between devices, we are unable to compare and only evaluate relative improvements on our own devices. The model was tested on two devices with varying computational capabilities, and results indicate significant improvements in latency reduction across both platforms. Whilst AS caching performs well, and significantly reduces inference times relative to no-caching (encoder self-attention), KV caching consistently produces the lowest inference times on both CPU and GPU configurations for Machine A and the best GPU inference time for Machine B. We speculate that the superior performance of KV caching over AS caching, which contradicts its theoretical higher computational complexity, is due to the added storage and retrieval of the query and AS cache. Additionally, it may suggest that further optimization may be necessary. The GPU's underperformance relative to its potential could be linked to the inherent low-level operations involved in attention score caching, which may not be fully optimized for parallel processing using CUDA. Similar patterns were observed in Machine B, where CPU performance lagged slightly behind GPU, but the difference was larger than in Machine A. This suggests that memory access speeds, rather than raw computational power, could be the limiting factor in these devices.

Further investigation is needed to optimize memory access times, which could yield even greater reductions in inference time. Nonetheless, by enabling either caching technique we reduce the window size six times, while still achieving competitive accuracies in both WESAD and UBFC-Phys, and our attention score caching still demonstrates to significantly lower the latency than encoder self-attention which were used in the previous studies investigating attention on stress detection [23], [93].

6.5 Limitations & Future Work

- *Generalizability:* The possibility for data snooping in the datasets used in this study, further extends the need for a non-protocol based dataset to improve generalizability, and avoid the possibility of such an effect. The inclusion criterion for a new dataset collection study could be broadened to those who have done sport, are pregnant and so on, to address the lack of intra-participant variability in protocolized studies, which pose challenges for this technology to be generalized to real-world use [147]. For the same reasoning discussed earlier, the problem of generalizability has led to other datasets being collected, such as the collection of biometric data of nurses during the COVID-19 outbreak [148], or the collection of participants during daily activities and receiving notifications to report their momentary stress at random times of the day [23]. A secondary dataset focusing on generalizability to real-life scenarios would be both feasible and beneficial, as the experimental apparatus in the MUSED dataset is appropriate for long-term monitoring studies, with all but the sEMG recording device being portable. This would allow for further evaluation of the high-performing methods across the protocolized datasets mentioned, ensuring better generalization to real-world cases.

- *Modularity:* The results from this study highlight the need for further research into using modular cross-attention mechanisms as a replacement for traditional data imputation techniques in stress detection. As noted by Yu et al. [23], traditional approaches often fail to generalize effectively to new datasets, particularly when dealing with missing data. By contrast, the modular cross-attention network presented here offers a promising alternative, capable of adapting to varying conditions without a loss in generalizability. However, we believe that matching state-of-the-art accuracies on several datasets, does not justly evaluate the generalizability performance, and we therefore re-iterate the need for testing on non-protocolized, real-world datasets.

- *Personalization:* To continue addressing the requirement for personalization, further work should aim at implementing the personalisation mechanisms developed in this paper into a user-ready device. By using active reinforcement learning, whereby the pre-trained model is fine-tuned to adjust for the new user's stress response based off the user's labelling, personalized learning will take place and could adapt the model. For example, the framework showcased in Tazaraz et al. [78] optimized label collection efficiency, reducing inconvenience for users whilst creating user-specific datasets for fine-tuning the model through online learning, demonstrating to be a pragmatic approach for real-life personalized deployment.

Additionally, while Kalman filters have shown potential in adaptive modelling scenarios, particularly when applied to personalized systems, we did not implement this within our architecture. However, future work could explore the adaptation of Kalman filters to dynamically update variables such as the state transition matrix and covariance parameters based on individual-specific data. This could provide further robustness to the personalization framework, especially in scenarios where data from real-world stress monitoring is subject to noise and irregularities.

- *Early-Onset Detection:* While we demonstrate that our bidirectional encoder-decoder attention successfully competes with encoder self-attention and allows for a significant drop in window size, there is still a slight discrepancy between the two attention mechanisms. The reason behind this is still unclear, since the cached attention score should be able to reconstruct the attending from embeddings as if it all tokens had attended to one another collectively as in the encoder-self attention form. We hypothesize that due to the individual temporal slices generated by the `SeqToSeqDataloader`, there will be samples where the cached segments do not correspond to the same label during a transitional task. This was only considered retrospectively, and further testing could efficiently test this hypothesis. Nonetheless, transition periods, especially in the context of early onset detection, are arguably most important to predict. Further demonstrating the need for a more generalizable dataset to include several transitional periods between stress states.

- *Real-time Classification:* Firstly, further investigation should be made to identify bottlenecks in the module and optimizations can be made to the SlidingAttnScoreCache module to ensure parallelization is occurring at all locations, especially on a per-attention head basis. It may be required to utilize CUDA kernels to directly control GPU execution and memory [149]. The results also suggests that with higher-performing memory relative to its computational capabilities, further reductions in inference time might be achievable. However, additional research is necessary to explore this hypothesis and determine the extent of potential improvements.

  Secondly, while alternative segmentation methods, such as the inter- and intra-slice segmentation proposed by Xia et al. [91], could be employed, the datasets used in this study do not offer sufficient task durations to explore the potential benefits of inter-slice segmentation, further emphasising the need to evaluate the model on real-world datasets.

  Thirdly, varying the segmentation lengths for each biosignal was not explored due to the increased complexity it would introduce. Instead, we adopted the approach used by Zhang

et al. [93], where despite the slower nature of the EDA signal, the segmentation process was identical for the BVP signal. This alignment avoids the need for handling different window sizes, which would complicate the model architecture, but such high temporal resolution is not required for modalities such as EDA. A similar approach can be taken for embedding dimensions, which could be different for each modality branch to cut computational costs.

6

Conclusions

In this paper, we focus our research on creating a robust, modular method for the early-onset detection of stress using several innovative, consumer friendly devices. Firstly, we investigate the feasibility of two novel modalities untested in the field of stress detection: sEMG and EEG based mastoid monitoring and single-channel consumer-grade fNIRS for pre-frontal cortex measurements. The reasoning for this is the mastoid offers an excellent location for discrete and convenient monitoring through headphones developed by BrainPatch, a neurotechnology company, whilst consumer fNIRS serve as a portable yet underexplored modality for stress monitoring. The headphones serve as a familiar and unobtrusive medium for users to wear, and this project seeks to test the feasibility of this idea by exploring the potential of a novel multi-modal system that would eventually use the mastoids as a strategic site for monitoring the sympathetic nervous system.

After validating the presence of sEMG and EEG based features in pilot testing, confirming that both muscular signals from the upper trapezius, and alpha waves from the brain are visible at the area of the mastoid, we conducted a new study to serve as a public dataset for these new modalities. To support the research community, we introduce the MUSED dataset, which captures simultaneous monitoring of mastoid sEMG, EEG, and single-channel fNIRS activity. This dataset complements existing ones like WESAD and UBFC-Phys by employing a well-documented stress induction protocol and broadening the scope of biosignal monitoring. We used a virtual reality based TSST, that records baseline conditions, and three types of stress: anticipation stress through means of mentally preparing for an interview, social stress via the simulated interview condition, and cognitive stress via an arithmetic task. We collected data from 18 participants (10 males, 8 females), aged between 22–41 years old, with a mean age of 26.33 $\pm$ 4.77. In addition to the ones tested, we collect data from other modalities: namely chest-strap ECG and wrist-based BVP, EDA, temperature and accelerometer readings.

Our study addresses the limitations of traditional multimodal fusion techniques such as data imputation and decision-fusion by proposing a truly modular architecture, whereby the model adapts to the number and combination of devices that a user has available. We achieve this by incorporating a modular ensemble of bidirectional cross-attention and self-attention blocks, which were introduced by Zhang et al. [93]. This approach aligns temporal information ensuring that

biosignals from different modalities are handled robustly between any different modality combination, and adapts to the inherent asynchronicity between biosignals. The architecture is designed to ensure seamless communication between different sensory inputs, enabling it to adapt to varying combinations of modalities without significant loss in accuracy. We observe that the cross-attention network successfully learns to maintain an intermediate representation that will translate well no matter which combination of modality is used. This adaptability, resulted in a generalized modular model that only suffered an average of 2% performance sacrifice, compared to a fixed model trained on the same number and combination of modalities as tested on.

The model architecture incorporates personalization techniques to address the inter-participant variability widely documented in stress monitoring research. This variability, underscores the importance of personalizing models to capture individual nuances in stress responses, which generalized models fail to detect. By employing personalization mechanisms at multiple stages of the model, we enhance the detection accuracy while reducing computational overhead.

To address early-onset detection, we must recognise that inference time is not the same as latency. In typical methods, a buffer of biosignal data, of a set window size, is collected before real-time inference can take place. Typical window sizes for stress monitoring are between 30 and 60 seconds, however we employ a sequence-to-sequence attention model and caching mechanism to reduce the required window size to 6 seconds without degrading performance. This is accomplished by storing previous temporal slices, as key and value projections cache (KV cache). To the best of our knowledge, no other research paper implements or discusses the computational savings of KV cache for real-time biosignal classification tasks. We take this one step further, by implementing query projections, enabling attention score caching, and reducing the computational complexity significantly. A key innovation of our system is the Sliding Attention Score Caching Mechanism, which integrates a novel bidirectional encoder-decoder attention inspired by transformer models. This mechanism dynamically re-evaluates past biosignal inputs in light of new data, improving adaptability and accuracy in detecting stress biomarkers. By leveraging KV caching and attention score caching (AS caching), the system achieves significant computational savings - a crucial advantage for real-time stress detection on resource-constrained devices.

While the study demonstrates the feasibility of using mastoid sEMG and consumer-grade single-channel fNIRS for stress detection, certain challenges related to signal robustness and noise remain. We found statistically significant ($p < 0.05$) features from, fNIRS and these features yielded high Gini importance against other features. However, key sEMG and EEG based biomarkers related to stress detection were not found to be statistically significant, raising concerns about signal

robustness and the influence of noise. To increase confidence in their capabilities, more rigorous research is needed to explore signal robustness, enhance feature extraction techniques, and validate the system in larger, more diverse populations. Nonetheless, the convenience and unobtrusiveness of these devices offer a compelling case for their continued development as part of a multimodal stress detection system.

The proposed Sliding Attention Score Caching Mechanism, integrated with a novel bidirectional encoder-decoder attention, demonstrates a promising advancement in the real-time detection of stress from multimodal biosignals. Unlike traditional stress detection models, which rely heavily on static representations of past signals, this approach dynamically reevaluates past inputs in light of new information. This bidirectional attention mechanism enables the model to reinterpret previous signals when new biosignal data becomes available, thus improving its adaptability and accuracy in identifying stress biomarkers, especially under fluctuating conditions. We conducted extensive testing on two well-established datasets and confirmed the capabilities of the generalized modular model, achieving comparable results to state-of-the-art whilst utilizing a fraction of the window size and enabling true modularity between modalities.

A

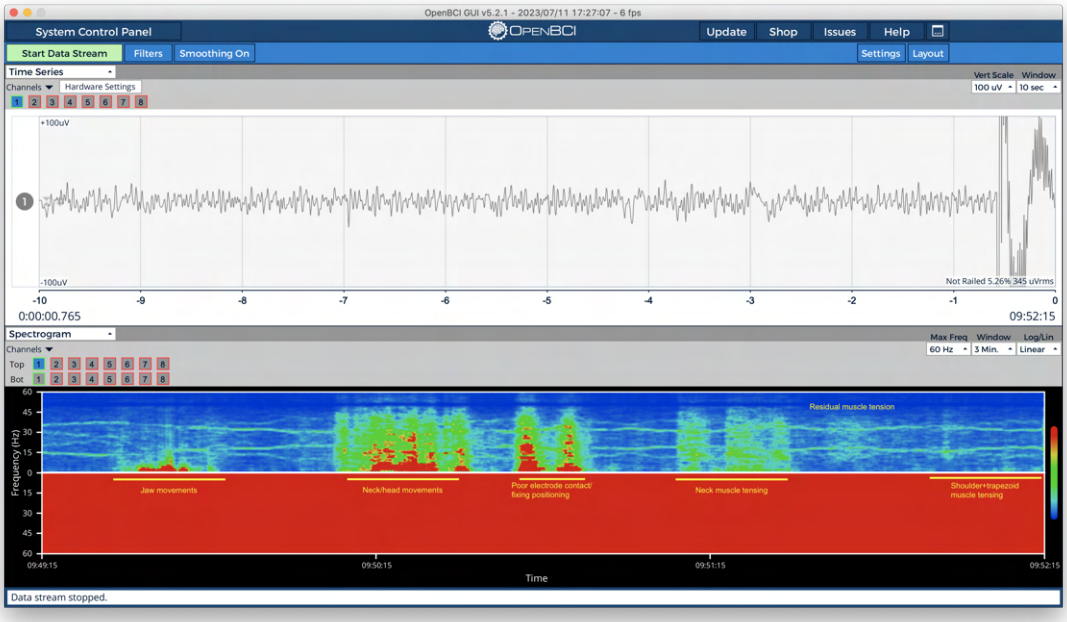
Foundations of Stress Analysis and Data Collection

A.1 Intra-Individual Factors

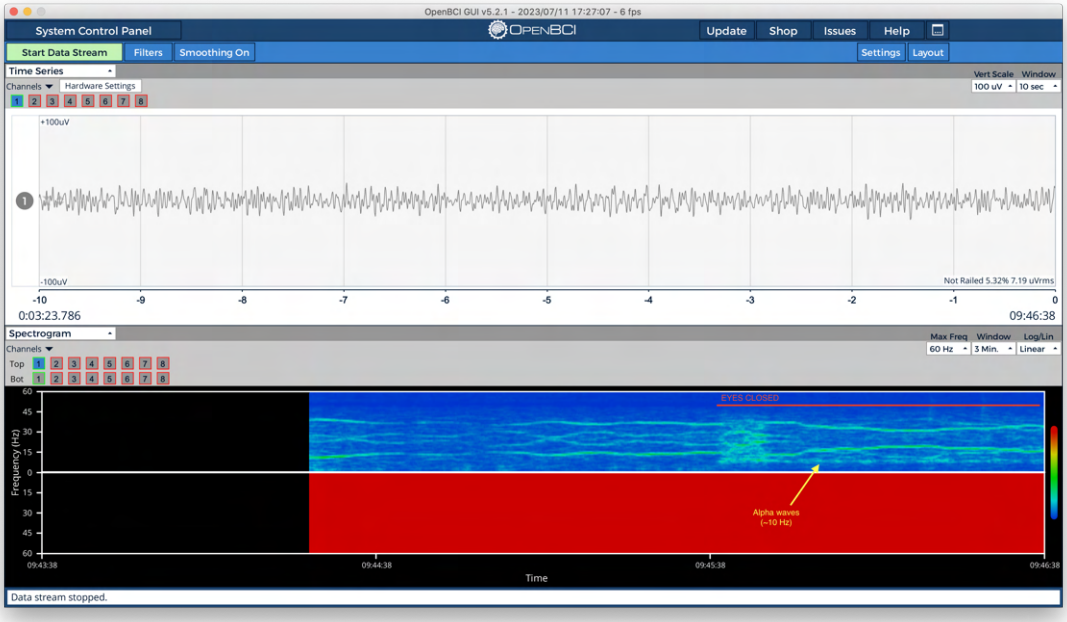
Sensor	Biomarker Of Stress	Intra-individual Examples
Salivary / Blood / Urine Sampling	Cortisol	Higher cortisol levels after caffeine [150]
EDA	Electrodermal Activity	Hypo-responsive periods [53]
ECG/PPG/fNIRS	Heart Rate	Quicker response to stressor in morning [151]
	Heart Rate Variability	10 ms decrease in RMSSD following caffeine consumption [152]
RESP / ECG/ fNIRS	Respiration Rate	Respiration rate and depth is dependent on the participant's reaction to different types of stressors [153]
fNIRS / EEG	Prefrontal Activity	Sleep deprivation causes lowering of prefrontal activity [154]

Table A.1: Examples of intra-individual factors that have shown to affect each biosignal discussed.

A.2 Further Pilot Studies



(a)



(b)

Figure A.1: Initial testing using the electrodes mounted on the headset show a) promising signs of muscular tension, detected via sEMG, particularly in the upper trapezius, and, b) alpha waves captured through EEG.

A.3 MUSED Dataset Documentation

Figure A.2: Read me attached to the MUSED dataset, which summarizes the devices, the signals collected, and how the data is structured. This will aid with further investigation of stress detection using our proposed dataset.

MUSED Dataset: Modular MUltimodal Stress Early-onset Detection

1 General Information

Contact details: William Powell, willpowell@gmx.co.uk.

If you publish material based on this dataset, please reference the publication [1].

2 Dataset Structure

The dataset is organised so that each subject has a folder (SX, where X = subject ID). Each subject folder contains the following files and directories:

- **empatica**: contains all data collected from the Empatica E4 wrist wearable; see Section 4.1
- **myndsens**: contains all data collected from the fNIRS wearable; see Section 4.2
- polar: contains all data collected from the Polar H10 chest wearable; see Section 4.3
- **quattrocento**: contains all data collected from the OT Bioelettronica Quattrocento EMG recording device; see Section 4.4
- **questionnaires**: contains the ground truth collected from the participant; see Section 6
- **Actions.csv**: contains the actions performed and their respective timestamps during the Open TSST-VR protocol.
- **Phases.csv:** contains the phases and their respective timestamps during the Open TSST-VR protocol.
- **readme.txt**: contains information about the subject (SX) and information about data collection and data quality (if applicable).
- **SX.pkl**: contains synchronised data and labels - use this concatenated file of all sensors and labels if you would like to avoid manual data preprocessing; see Section 5.

3 Subjects

18 subjects participated in the study. However, due to sensor malfunction, fNIRS data of two subjects (S2 and S3) are compromised.

4 Data Format

4.1 Emapatica

The "empatica" directory contains all sensory activity from the Epatica E4 wristband. The details for each signal is explained more detail on their website: https://support.empatica.com/hc/en-us/sections/200582445-E4-wristband-data. But to summarize:

- ACC.csv: sampled at 32 Hz. The 3 data columns refer to the 3 accelerometer channels. Data is provided in units of 1/64g.
- BVP.csv: sampled at 64 Hz. Data from Photoplethysmography (PPG) is unitless.
- EDA.csv: sampled at 4 Hz. Data is provided in µS.
- TEMP.csv: sampled at 4 Hz. Data is provided in °C.

4.2 Myndsens

The "myndsens" directory contains the file FNIRS.csv, which has already been interpolated to a sampling frequency of 10Hz. In the csv it contains headers: "O2Hb" (oxygenated hemoglobin), "HHb" (Deoxygenated Hemoglobin) and 'Brain Oxy' (Brain Oxygenation). O2Hb and HHb are both measured in micromolar (µM), which represents micromoles per liter (µmol/L). Brain Oxygenation levels are expressed as a percentage, providing an absolute measure of oxygen saturation.

4.3 Polar

The "polar" directory contains three separate CSVs:

- ACC.csv (which similar to the Epatica device, measures x, y, z accelerometer values, but on the chest rather than wrist). ACC is measured in milliG at 25 Hz with 16 bit precision. The accelerometer is configured to measure acceleration in the range [-2g, 2g], its units are 1/64g.
- ECG.csv measures at a sampling frequency of 130Hz at 14 bit precision - it is unitless.
- IBI.csv measures inter-beat intervals of the heart in seconds, at a sampling frequency of 1Hz. This is derived from the R peaks of the ECG at a sampling frequency of 1kHz, therefore it is assumed to be the ground truth for heart rate.

4.4 Quattrocento

The "quattrocento" directory contains one csv. This is the data captured by OT Bioelettronica, see https://otbioelettronica.it/en/quattrocento/. The CSV's first column, labeled "Upper Trapezius", and the second column, labeled "Mastoid", contain the bipolar EMG data recorded from their respective channels, recorded in g. All data is captured at 2048Hz with a precision of 16-bit resolution and is measured in microvolts. The onboard filtering was set to have a high pass of 0.7Hz and a low pass of 500Hz.

5 Data Synchronization and Labelling

The double-tap signal pattern was used to manually synchronise the devices' raw data. The synchronised data has been prepared in the files SX.pkl (pickle file), one file per subject. This file is a dictionary, structured as follows:

- Source Sensor (key): i.e. the recording device (empatica, myndsens, quattrocento or polar).
- Sensor Measurement (key) i.e. for polar this would contain: "ecg", "ibi" and "acc".
- Measurement Data Frame (i.e. for pkl_file["polar"]["acc"] this would return the pandas dataframe with columns "x", "y", and "z" and "Label".
- The "Label" column is present in all dataframe files and contains the ID for the protocol condition as follows:
  - Each label corresponds to the timestamps located in the 'Phases.csv' file, in order. For example 0 = Set Marker (should ignore), 1 = Baseline Sit, 2 = Baseline Stand, 3 = Anticipation 4 = Interview, 5 = Arithmetic Task.
  - In the study for the binary classification task, we assume Labels 1 and 2 to be our non-stress condition and 3, 4 and 5 to be our stress condition.

6 Ground Truth

Within the study protocol, immediately before and immediately after the virtual reality (post arithmetic task), the subjects were asked to fill in several self-reports. The self-reports consist of the Short-Form PANAS (PANAS-SF) and Short-Form STAI. Additionally an SSSQ was asked only immediately after the virtual reality protocol. Answers are provided in their respective files marked "pre" and "post" protocol in the questionnaires directory in csv format.

PANAS questionnaire items (1 = Not at all, 2 = A little bit, 3 = Somewhat, 4 = Very much, 5 = Extremely)

- Active
- Distressed
- Interested
- Inspired
- Annoyed
- Strong
- Guilty
- Scared
- Hostile
- Excited
- Proud
- Irritable
- Enthusiastic
- Ashamed
- Alert
- Nervous
- Determined
- Attentive
- Jittery
- Afraid
- Stressed
- Frustrated
- Happy
- (Angry)
- (Irritated)
- Sad

STAI questionnaire items (1 = Not at all, 2 = Somewhat, 3 = Moderately so, 4 = Very much so)

- I feel at ease
- I feel nervous
- I am jittery
- I am relaxed
- I am worried
- I feel pleasant

SAM questionnaire items (scale 1-9)

- Valence (1 = low valence, 9 = high valence)
- Arousal (1 = low arousal, 9 = high arousal)

SSSQ questionnaire items (1 = Not at all, 2 = A little bit, 3 = Somewhat, 4 = Very much, 5 = Extremely)

- I was committed to attaining my performance goals
- I wanted to succeed on the task
- I was motivated to do the task
- I reflected about myself
- I was worried about what other people think of me
- I felt concerned about the impression I was making

7 Disclaimer

References

[1] W. Powell, D. Farina, and A. Spiers, "Mused: Modular multimodal stress early-onset detection." 2024.

B

Additional Results

We present further results and corresponding observations related to validating the protocol, signal processing, feature analysis, protocol validation and raw data corresponding to plots reported in Chapter 5.

B.1 Protocol Validation

Variable	T-Statistic	T-Test P-Value
I am jittery	-4.68	0.000
I am relaxed	2.53	0.022
I am worried	-2.96	0.009
I feel at ease	2.60	0.019
I feel nervous	-3.00	0.008

Table B.1: Statistical significance of pre- and post- protocol conditions on STAI questionnaire using T-Test.

Group	T-Statistic	T-Test P-Value
Positive	0.5859	0.561925
Negative	-5.2263	0.000026

Table B.2: Statistical significance of pre- and post-conditions for positive and negative attributes of the PANAS questionnaire using T-Test.

Engagement	Distress	Worry
4.28 ± 0.787	3.61 ± 0.850	3.56 ± 1.10

Table B.3: Evaluation of the post protocol SSSQ indicates that subjects were strongly engaged during the interview and arithmetic tasks and had a comparable amount of moderate worry and distress.

B.2 Signal Validation

B.2.1 fNIRS

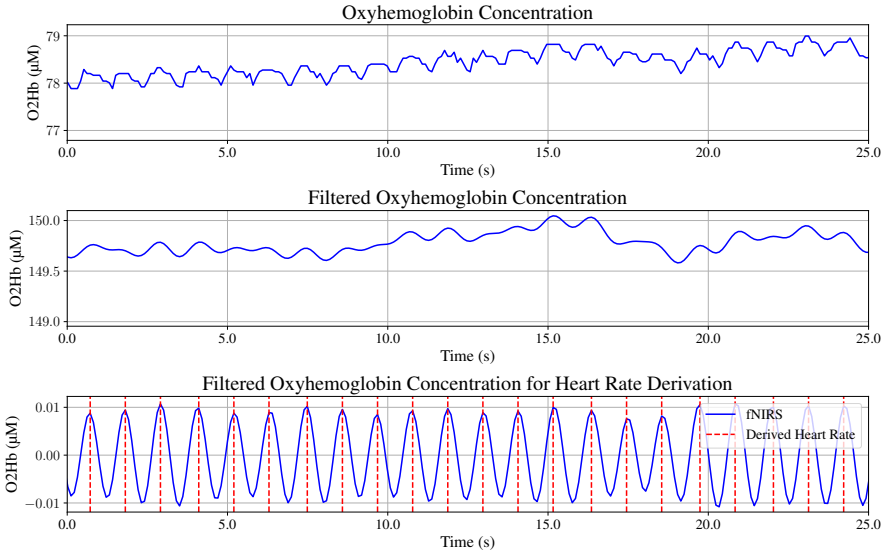
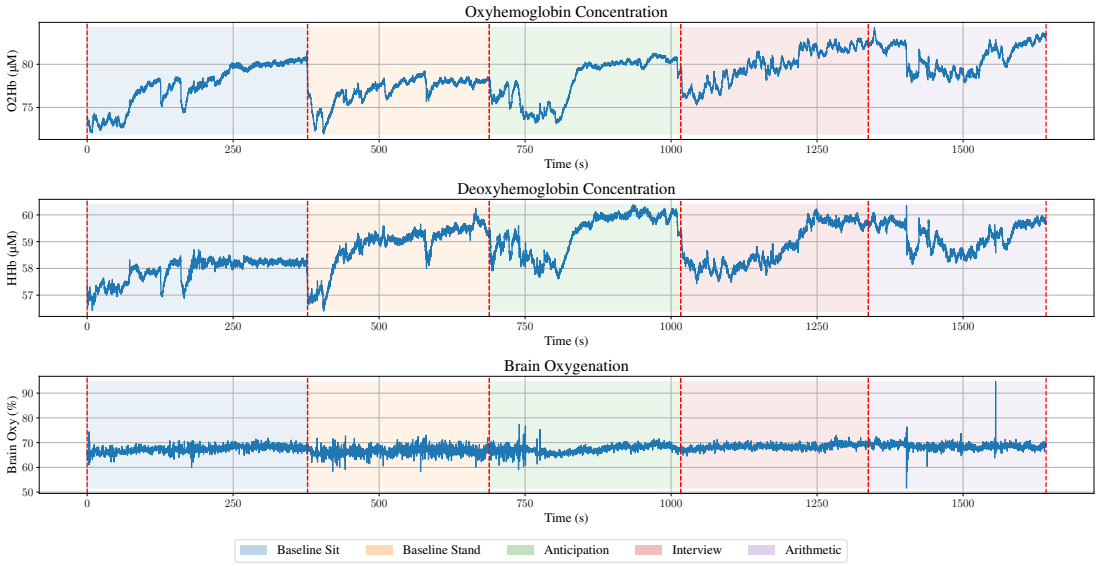
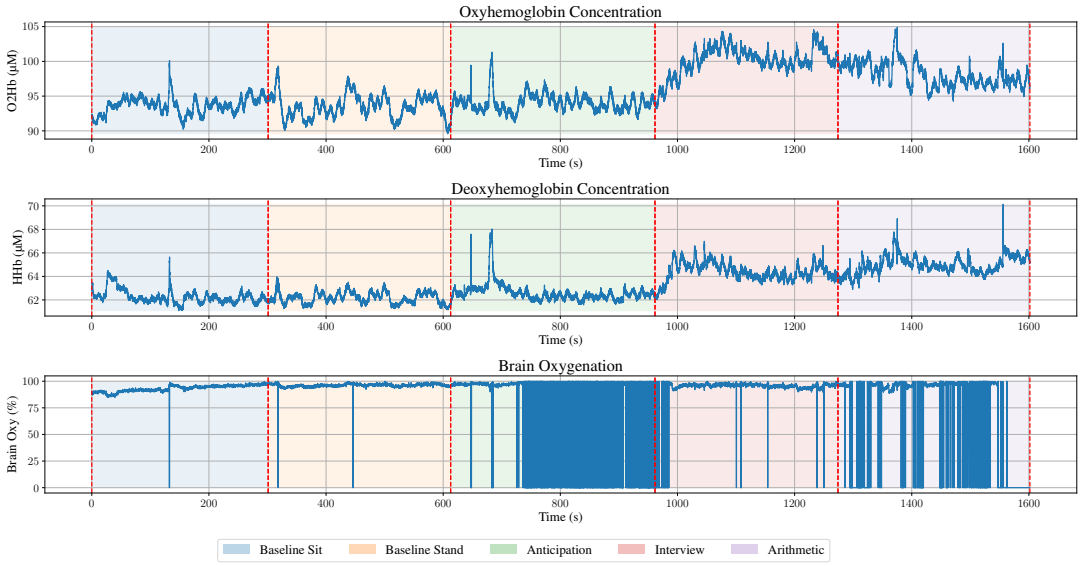


Figure B.1: fNIRS derived heart rate using the filtering and extraction algorithm detailed in Section 4.2 and 4.3 respectively. Note that the Mayer waves are attenuated in the filtered fNIRS signal.



(c)



(d)

Figure B.2: a) fNIRS from Subject 16 shows a prolonged recovery period after a motion artefact induces a blood volume shift. Nonetheless, the brain oxygenation measurement shows to be unaffected and demonstrates its robustness to motion artefacts. b) fNIRS from Subject 17 shows a period of poor signal quality on the brain oxygenation channel, which may be due to poor contact, or excess sweat on the sensor.

B.2.2 sEMG

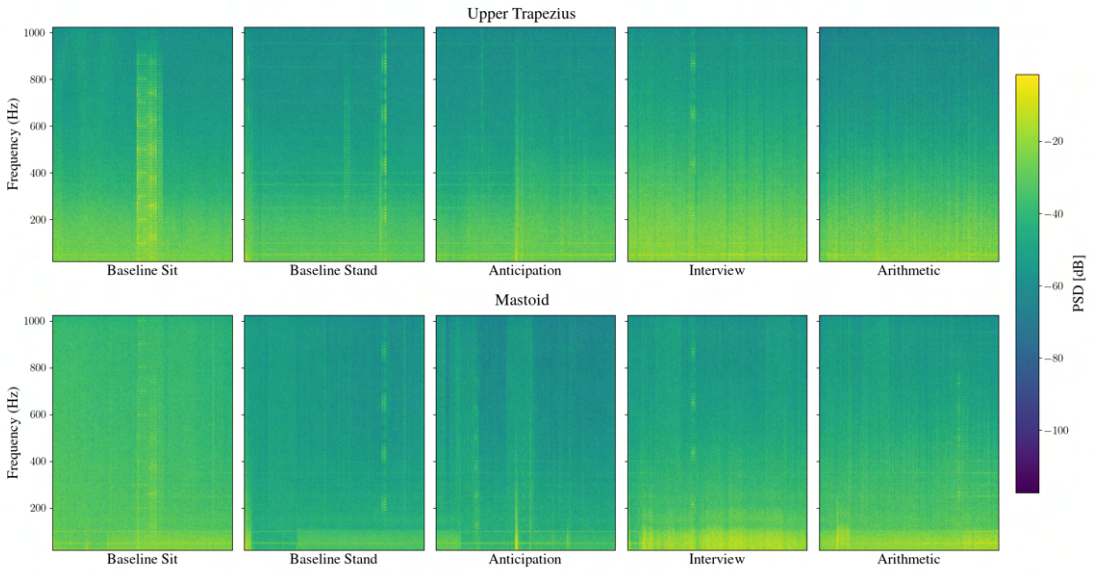


Figure B.3: High frequency sEMG spectrogram for Subject 7 shows minimal activity, with other subjects demonstrating similar results, suggesting that high frequency muscular activity did not reach the same amplitude levels observed during a shoulder contraction, as demonstrated in the pilot.

B.2.3 ECG

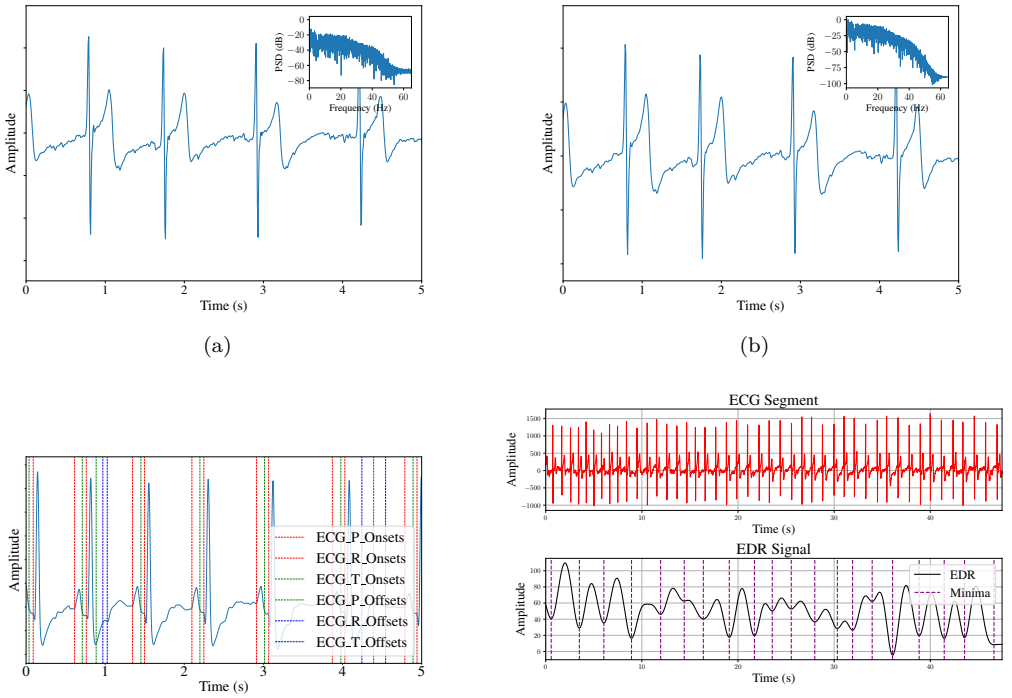
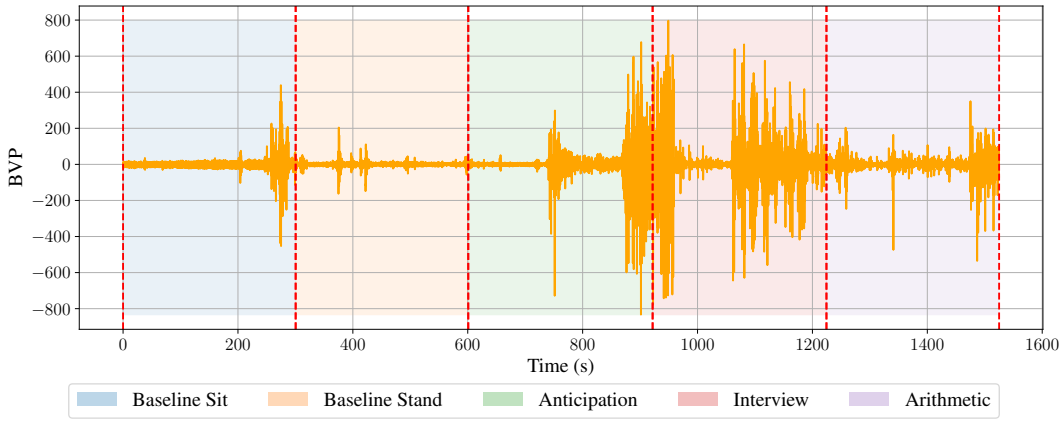
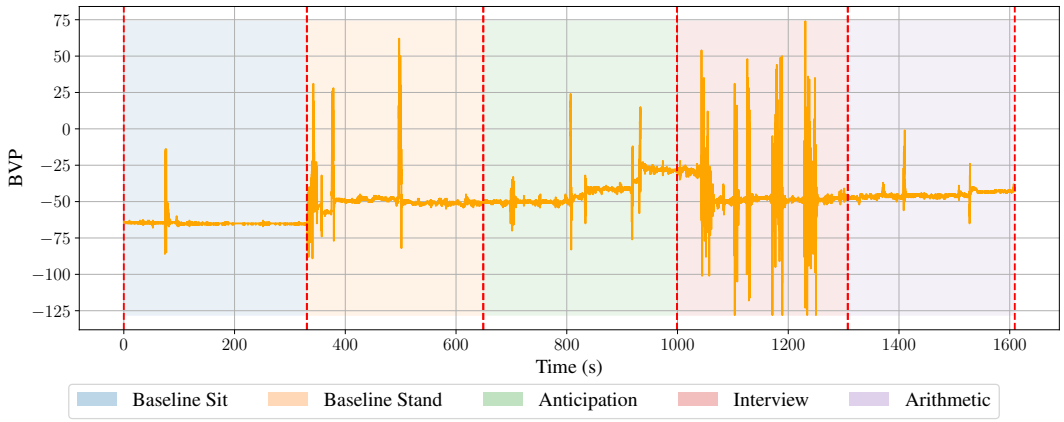


Figure B.4: ECG signal processing and feature extraction process of the Polar H10 ECG signal. a) Raw ECG signal. b) Cleaned ECG signal showing attenuation of frequencies above 50Hz. c) Onset and offset detection of ECG shows poor reliability for R-offsets, and T onsets and offsets due to the formation of the ECG. d) ECG segment and corresponding derived respiration signal (EDR).

B.2.4 PPG



(a)



(b)

Figure B.5: a) Subject 1 PPG signal from Empatica E4 shows large artefacts significantly above typical amplitudes, due to poor contact between sensor and wrist. b) Subject 18 PPG signal shows motion artefacts due to hand gestures made during the interview task.

B.3 Feature Validation

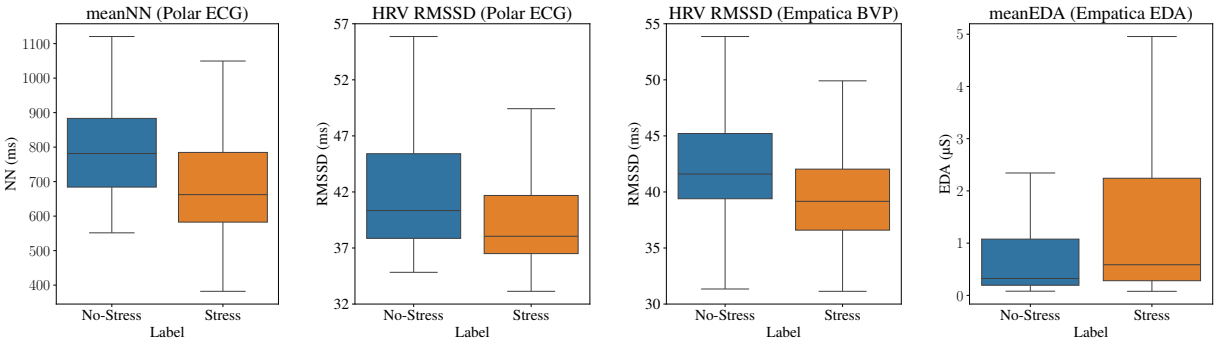


Figure B.6: Box plots illustrating key features extracted from the Polar and Empatica devices.

Binary Classification			Four-Level Classification		
Device	**Feature**	**Gini Importance**	**Device**	**Feature**	**Gini Importance**
Polar	HRV_MeanNN	0.0474	Empatica	max_EDA	0.0279
Polar	HRV_Prc80NN	0.0440	Empatica	mean_EDA	0.0272
Polar	HRV_MedianNN	0.0434	Empatica	min_EDA	0.0253
Polar	HRV_MedianNN	0.0351	Empatica	sum_amp_SCR	0.0248
Polar	HRV_Prc20NN	0.0339	Polar	HRV_MeanNN	0.0242
Polar	mean_IBI	0.0296	Empatica	mean_SCL	0.0219
Polar	max_IBI	0.0253	Polar	HRV_MedianNN	0.0218
Polar	HRV_Prc80NN	0.0250	Polar	HRV_MedianNN	0.0214
Polar	HRV_MinNN	0.0207	Polar	HRV_Prc80NN	0.0190
Myndsens	O2Hb_max	0.0183	Polar	HRV_Prc20NN	0.0183
Empatica	sum_amp_SCR	0.0167	Polar	mean_IBI	0.0177
Polar	HRV_MeanNN	0.0160	Empatica	area_SCR	0.0174
Polar	min_IBI	0.0149	Empatica	mean_SCR	0.0168
Empatica	mean_EDA	0.0144	Myndsens	O2Hb_max	0.0151
Polar	HRV_MaxNN	0.0143	Polar	HRV_Prc80NN	0.0144
Polar	HRV_MaxNN	0.0137	Polar	max_IBI	0.0119
Empatica	mean_SCL	0.0136	Empatica	std_SCL	0.0107
Empatica	max_EDA	0.0136	Myndsens	O2Hb_mean	0.0106
Empatica	area_SCR	0.0135	Empatica	std_SCR	0.0105
Myndsens	O2Hb_mean	0.0124	Empatica	range_EDA	0.0103

Table B.4: Ranked Gini importance scores of the top features for binary classification and four-class tasks with updated data.

Binary Classification			Four-Level Classification		
Device	**Sensor**	**Gini Importance**	**Device**	**Sensor**	**Gini Importance**
Empatica	EDA	0.2470	Empatica	EDA	0.2792
Polar	ECG	0.2402	Myndsens	fNIRS	0.1530
Myndsens	fNIRS	0.1368	Polar	ECG	0.1376
Polar	IBI	0.1180	Empatica	BVP	0.1337
Empatica	BVP	0.1060	Polar	IBI	0.1180
Quattrocento	Upper Trap EMG	0.0431	Empatica	Wrist ACC	0.0469
Quattrocento	Mastoid EMG	0.0344	Quattrocento	Mastoid EMG	0.0427
Polar	Chest ACC	0.0342	Quattrocento	Upper Trap EMG	0.0358
Empatica	Wrist ACC	0.0326	Polar	Chest ACC	0.0338
Empatica	TEMP	0.0077	Empatica	TEMP	0.0193

Table B.5: Ranked Gini importance scores of the top sensors for binary classification and four-class tasks on the MUSED dataset.

Subject	S1	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18
Accuracy	0.901	0.864	0.933	0.882	0.944	0.932	0.954	0.874	0.902	0.781	0.935	0.761	0.823	0.784	0.763	0.743

Table B.6: fNIRS modality performance on a per-subject basis using the fixed model and LOSO-CV.

B.4 Computational Performance

		CPU Inference Time (ms)			GPU Inference Time (ms)		
Modalities	**Parameters**	No Cache	KV Cache	AS Cache	No Cache	KV Cache	AS Cache
1	2,210	4.389 ± 0.441	1.464 ± 0.429	1.944 ± 0.422	3.696 ± 0.427	1.319 ± 0.373	1.76 ± 0.409
2	8,228	7.837 ± 0.981	2.499 ± 0.474	3.263 ± 0.461	6.667 ± 0.686	2.272 ± 0.405	3.017 ± 0.438
3	18,054	12.502 ± 1.42	3.982 ± 0.611	5.012 ± 0.577	10.677 ± 0.952	3.662 ± 0.504	4.654 ± 0.544
4	31,688	20.748 ± 2.196	6.456 ± 0.839	7.926 ± 0.77	17.741 ± 1.42	5.839 ± 0.658	7.285 ± 0.713
5	49,130	32.294 ± 3.282	9.917 ± 1.158	12.004 ± 1.041	27.489 ± 2.066	9.021 ± 0.884	11.114 ± 0.959
6	70,380	46.48 ± 4.616	14.369 ± 1.568	17.249 ± 1.389	39.273 ± 2.847	13.152 ± 1.176	16.014 ± 1.275
7	95,438	64.117 ± 6.275	19.808 ± 2.069	23.66 ± 1.814	54.673 ± 3.868	18.121 ± 1.529	21.902 ± 1.654
8	124,304	81.707 ± 7.93	26.238 ± 2.661	31.236 ± 2.317	69.613 ± 4.858	24.073 ± 1.951	29.008 ± 2.111
9	156,978	105.095 ± 10.13	33.656 ± 3.344	39.977 ± 2.897	89.47 ± 6.174	30.823 ± 2.429	37.073 ± 2.63
10	193,300	122.733 ± 11.789	42.063 ± 4.119	49.884 ± 3.555	104.538 ± 7.173	38.511 ± 2.974	46.283 ± 3.223

Table B.7: Computational performance comparison across different caching methods for the generalized model. Tested on Intel Core i9-10900 CPU @ 2.80GHz × 20 with NVIDIA GeForce RTX 3090, 64GiB DDR4 Synchronous 2133MHz RAM.

		CPU Inference Time (ms)			GPU Inference Time (ms)		
Modalities	**Parameters**	No Cache	KV Cache	AS Cache	No Cache	KV Cache	AS Cache
1	2,210	7.025 ± 0.443	2.782 ± 0.431	2.643 ± 0.424	4.447 ± 0.429	1.990 ± 0.375	2.274 ± 0.411
2	8,228	15.097 ± 0.986	4.715 ± 0.477	4.435 ± 0.463	9.049 ± 0.689	3.299 ± 0.407	3.767 ± 0.44
3	18,054	24.084 ± 1.427	7.514 ± 0.614	6.811 ± 0.579	14.492 ± 0.956	5.317 ± 0.506	6.135 ± 0.546
4	31,688	39.969 ± 2.206	12.179 ± 0.843	10.772 ± 0.774	24.080 ± 1.427	8.477 ± 0.661	9.602 ± 0.716
5	49,130	62.211 ± 3.297	18.709 ± 1.163	16.317 ± 1.046	37.312 ± 2.076	13.097 ± 0.888	14.650 ± 0.964
6	70,380	89.539 ± 4.638	27.106 ± 1.575	23.445 ± 1.396	53.308 ± 2.861	19.094 ± 1.182	21.108 ± 1.281
7	95,438	123.515 ± 6.305	37.369 ± 2.079	32.158 ± 1.823	74.212 ± 3.886	26.308 ± 1.536	28.868 ± 1.662
8	124,304	157.401 ± 7.967	49.497 ± 2.674	42.456 ± 2.328	94.491 ± 4.881	34.948 ± 1.96	38.235 ± 2.121
9	156,978	202.456 ± 10.178	63.492 ± 3.36	54.337 ± 2.911	121.443 ± 6.203	44.749 ± 2.441	48.864 ± 2.643
10	193,300	236.433 ± 11.845	79.352 ± 4.138	67.803 ± 3.572	141.895 ± 7.207	55.91 ± 2.988	61.004 ± 3.238

Table B.8: Computational performance comparison across different caching methods for the generalized model. Tested on Intel Core i7-10510U CPU @ 1.80GHz × 8 with NVIDIA GeForce MX250, 12GiB DDR4 2400 Synchronous 1200Hz RAM.

B.5 Personalization

WESAD			UBFC			MUSED		
Subject	Generalized	Personalized	Subject	Generalized	Personalized	Subject	Generalized	Personalized
S2	0.979	0.986	S2	0.893	0.931	S1	0.768	0.883
S3	0.969	0.976	S7	0.853	0.873	S4	0.773	0.891
S4	0.949	0.968	S15	0.902	0.919	S5	0.817	0.911
S5	0.989	0.989	S16	0.956	0.955	S6	0.974	0.974
S6	0.957	0.977	S18	0.847	0.867	S7	0.933	0.941
S7	0.972	0.979	S20	0.847	0.851	S8	0.974	0.968
S8	0.945	0.955	S23	0.960	0.965	S9	0.974	0.975
S9	0.960	0.970	S24	0.910	0.922	S10	0.828	0.871
S10	0.987	0.989	S29	0.833	0.893	S11	0.949	0.936
S11	0.872	0.912	S34	0.896	0.916	S12	0.831	0.831
S13	0.967	0.981	S36	0.833	0.858	S13	0.942	0.944
S14	0.852	0.902	S43	0.833	0.837	S14	0.956	0.967
S15	0.925	0.932	S44	0.877	0.910	S15	0.974	0.977
S16	0.963	0.969	S46	0.743	0.761	S16	0.761	0.771
S17	0.832	0.872	S51	0.755	0.778	S17	0.757	0.797
						S18	0.778	0.855

Table B.9: Generalized model vs. personalized model performance on a per-subject basis for each of the three datasets tested using LOSO-CV.

B

Bibliography

[1] B. D. Winslow, R. Kwasinski, J. Hullfish, *et al.*, "Automated stress detection using mobile application and wearable sensors improves symptoms of mental health disorders in military personnel," *Frontiers in Digital Health*, vol. 4, Aug. 2022. DOI: 10.3389/fdgth.2022.919626. [Online]. Available: https://doi.org/10.3389/fdgth.2022.919626.

[2] C. Maslach and S. E. Jackson, "The measurement of experienced burnout," *Journal of Organizational Behavior*, vol. 2, no. 2, pp. 99–113, Apr. 1981. DOI: 10.1002/job.4030020205. [Online]. Available: https://doi.org/10.1002/job.4030020205.

[3] C. Maslach, S. E. Jackson, and M. P. Leiter, "Maslach burnout inventory: Third edition.," in (Evaluating stress: A book of resources.), Evaluating stress: A book of resources. Lanham, MD, US: Scarecrow Education, 1997, pp. 191–218, ISBN: 0-8108-3231-3 (Hardcover).

[4] *Burnout*, https://mentalhealth-uk.org/burnout/, Accessed: 2023-04-02, 2020.

[5] M. R. Salleh, "Life event, stress and illness," *The Malaysian journal of medical sciences*, vol. 15, no. 4, 2009, ISSN: 1394-195X.

[6] "Covax delivers its 1 billionth covid-19 vaccine dose. covid-19 pandemic triggers 25 percent increase in prevalence of anxiety and depression worldwide," eng, *Saudi medical journal*, vol. 43, no. 4, pp. 438–439, 2022, ISSN: 0379-5284.

[7] F. Pasquier, P. Denise, A. Gauthier, N. Bessot, and G. Quarck, "Impact of galvanic vestibular stimulation on anxiety level in young adults," *Frontiers in Systems Neuroscience*, vol. 13, Apr. 2019. DOI: 10.3389/fnsys.2019.00014. [Online]. Available: https://doi.org/10.3389/fnsys.2019.00014.

[8] Brainpatch.ai, *Safe burnout prevention and stress relief for businesses*, https://www.brainpatch.ai/, Accessed: 2024-09-03.

[9] G. Crétot-Richert, M. De Vos, S. Debener, M. G. Bleichner, and J. Voix, "Assessing focus through ear-EEG: A comparative study between conventional cap EEG and mobile in- and around-the-ear EEG systems," en, *Front. Neurosci.*, vol. 17, p. 895 094, Sep. 2023.

[10] J. H. Lee, H. Gamper, I. Tashev, *et al.*, "Stress monitoring using multimodal bio-sensing headset," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '20, <conf-loc>, <city>Honolulu</city>, <state>HI</state>, <country>USA</country>, </conf-loc>: Association for Computing Machinery, 2020, pp. 1–7, ISBN: 9781450368193. DOI: 10.1145/3334480.3382891. [Online]. Available: https://doi.org/10.1145/3334480.3382891.

[11] R. Alcaide, N. Agarwal, J. Candassamy, *et al.*, "EEG-based focus estimation using neurable's enten headphones and analytics platform," Jun. 2021.

[12] U. Lundberg, R. Kadefors, B. Melin, *et al.*, "Psychophysiological stress and EMG activity of the trapezius muscle," en, *Int. J. Behav. Med.*, vol. 1, no. 4, pp. 354–370, 1994.

[13] J. Wijsman, B. Grundlehner, J. Penders, and H. Hermens, "Trapezius muscle EMG as predictor of mental stress," en, *ACM Trans. Embed. Comput. Syst.*, vol. 12, no. 4, pp. 1–20, Jun. 2013.

[14] N. Hakimi and S. K. Setarehdan, "Stress assessment by means of heart rate derived from functional near-infrared spectroscopy," en, *J. Biomed. Opt.*, vol. 23, no. 11, pp. 1–12, Nov. 2018.

[15] F. Al-Shargie, M. Kiguchi, N. Badruddin, S. C. Dass, A. F. M. Hani, and T. B. Tang, "Mental stress assessment using simultaneous measurement of eeg and fnirs," *Biomed. Opt. Express*, vol. 7, no. 10, pp. 3882–3898, Oct. 2016. DOI: 10.1364/BOE.7.003882. [Online]. Available: https://opg.optica.org/boe/abstract.cfm?URI=boe-7-10-3882.

[16] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ser. ICMI '18, ACM, Oct. 2018. DOI: 10.1145/3242969.3242985. [Online]. Available: http://dx.doi.org/10.1145/3242969.3242985.

[17] R. M. Sabour, Y. Benezeth, P. De Oliveira, J. Chappé, and F. Yang, "Ubfc-phys: A multimodal database for psychophysiological studies of social stress," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 622–636, 2023. DOI: 10.1109/TAFFC.2021.3056960.

[18] M. Naegelin, R. P. Weibel, J. I. Kerr, *et al.*, "An interpretable machine learning approach to multimodal stress detection in a simulated office environment," *Journal of Biomedical Informatics*, vol. 139, p. 104 299, 2023, ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2023.104299. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046423000205.

[19] A. Iranfar, A. Arza, and D. Atienza, "Relearn: A robust machine learning framework in presence of missing data for multimodal stress detection from physiological signals," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021, pp. 535–541. DOI: 10.1109/EMBC46164.2021.9630040.

[20] A. Tlija, K. Węgrzyn-Wolska, and D. Istrate, "Missing-data imputation using wearable sensors in heart rate variability," *Bulletin of the Polish Academy of Sciences Technical Sciences*, pp. 255–261, Apr. 2020, ISSN: 2300-1917. DOI: 10.24425/bpasts.2020.133118. [Online]. Available: http://dx.doi.org/10.24425/bpasts.2020.133118.

[21] J.-Y. Jiang, Z. Chao, A. L. Bertozzi, W. Wang, S. D. Young, and D. Needell, "Learning to predict human stress level with incomplete sensor data from wearable devices," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, ser. CIKM '19, Beijing, China: Association for Computing Machinery, 2019, pp. 2773–2781, ISBN: 9781450369763. DOI: 10.1145/3357384.3357831. [Online]. Available: https://doi.org/10.1145/3357384.3357831.

[22] N. Rashid, T. Mortlock, and M. A. A. Faruque, *Stress detection using context-aware sensor fusion from wearable devices*, 2023. DOI: 10.48550/ARXIV.2303.08215. [Online]. Available: https://arxiv.org/abs/2303.08215.

[23] H. Yu, T. Vaessen, I. Myin-Germeys, and A. Sano, "Modality fusion network and personalized attention in momentary stress detection in the wild," in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, Sep. 2021. DOI: 10.1109/acii52823.2021.9597459. [Online]. Available: http://dx.doi.org/10.1109/ACII52823.2021.9597459.

[24] A. Alberdi, A. Aztiria, and A. Basarab, "Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review," *Journal of Biomedical Informatics*, vol. 59, pp. 49–75, Feb. 2016, ISSN: 1532-0464. DOI: 10.1016/j.jbi.2015.11.007. [Online]. Available: http://dx.doi.org/10.1016/j.jbi.2015.11.007.

[25] L. Rochette and C. Vergely, "Hans selye and the stress response: 80 years after his "letter" to the editor of nature," *Annales de Cardiologie et d'Angéiologie*, vol. 66, no. 4, pp. 181–183, Sep. 2017. DOI: 10.1016/j.ancard.2017.04.017. [Online]. Available: https://doi.org/10.1016/j.ancard.2017.04.017.

[26] H. SELYE, "A syndrome produced by diverse nocuous agents," *Nature*, vol. 138, no. 3479, pp. 32–32, Jul. 1936. DOI: 10.1038/138032a0. [Online]. Available: https://doi.org/10.1038/138032a0.

[27] S. Levine, "A definition of stress?" In *Animal Stress*. Springer New York, 1985, pp. 51–69, ISBN: 9781461475446. DOI: 10.1007/978-1-4614-7544-6_4. [Online]. Available: http://dx.doi.org/10.1007/978-1-4614-7544-6_4.

[28] R. Murison, *Neuroscience of Pain, Stress, and Emotion*. Elsevier, 2016. DOI: 10.1016/c2013-0-16065-5. [Online]. Available: https://doi.org/10.1016/c2013-0-16065-5.

[29] Harvard Health Publishing. "Understanding the stress response," Harvard Medical School. (2024), [Online]. Available: https://www.health.harvard.edu/staying-healthy/understanding-the-stress-response (visited on 07/09/2024).

[30] G. Aguilera, *The Hypothalamic-Pituitary-Adrenal Axis and Neuroendocrine Responses to Stress*. 2012, ISBN: 9780123750976.

[31] A. Sanchez, "Changes in norepinephrine and epinephrine concentrations in adrenal gland of the rats submitted to acute immobilization stress," *Pharmacological Research*, vol. 48, no. 6, pp. 607–613, Dec. 2003. DOI: 10.1016/s1043-6618(03)00241-x. [Online]. Available: https://doi.org/10.1016/s1043-6618(03)00241-x.

[32] M. D. Breed and J. Moore, "Neurobiology and endocrinology for animal behaviorists," in *Animal Behavior*, Elsevier, 2012, pp. 25–65. DOI: 10.1016/b978-0-12-372581-3.00002-7. [Online]. Available: https://doi.org/10.1016/b978-0-12-372581-3.00002-7.

[33] *Parasympathetic nervous system - an overview*, https://www.sciencedirect.com/topics/computer-science/parasympathetic-nervous-system, Accessed: 2023-04-05, 2023.

[34] S. M. Mugo, W. Lu, and S. Robertson, "A wearable, textile-based polyacrylate imprinted electrochemical sensor for cortisol detection in sweat," *Biosensors*, vol. 12, no. 10, 2022, ISSN: 2079-6374. DOI: 10.3390/bios12100854. [Online]. Available: https://www.mdpi.com/2079-6374/12/10/854.

[35] Psychology Lexicon, *Parasympathetic rebound*, Accessed on July 09, 2024, Psychology Lexicon, 2024. [Online]. Available: https://www.psychology-lexicon.com/cms/glossary/49-glossary-p/13941-parasympathetic-rebound.html.

[36] R. Castaldo, P. Melillo, U. Bracale, M. Caserta, M. Triassi, and L. Pecchia, "Acute mental stress assessment via short term HRV analysis in healthy adults: A systematic review with meta-analysis," *Biomedical Signal Processing and Control*, vol. 18, pp. 370–377, Apr. 2015. DOI: 10.1016/j.bspc.2015.02.012. [Online]. Available: https://doi.org/10.1016/j.bspc.2015.02.012.

[37] "Abstract 15680: Validation of remote measurement of the qtc intervals using an apple watch," eng, *Circulation (New York, N.Y.)*, vol. 142, no. Suppl$_3$*Suppl*3, A15680–A15680, 2020, ISSN: 0009-7322.

[38] "Polar h10 heart rate sensor system," Jan. 2019. [Online]. Available: `https://www.polar.com/sites/default/files/static/science/white-papers/polar-h10-heart-rate-sensor-white-paper.pdf`.

[39] T. Park, M. Lee, T. Jeong, Y.-I. Shin, and S.-M. Park, "Quantitative analysis of eeg power spectrum and emg median power frequency changes after continuous passive motion mirror therapy system," *Sensors*, vol. 20, no. 8, p. 2354, Apr. 2020, ISSN: 1424-8220. DOI: `10.3390/s20082354`. [Online]. Available: `http://dx.doi.org/10.3390/s20082354`.

[40] S.-H. Seo, J.-T. Lee, and M. Crisan, "Stress and eeg," *Convergence and hybrid information technologies*, vol. 27, 2010.

[41] E. Perez-Valero, M. A. Vaquero-Blasco, M. A. Lopez-Gordo, and C. Morillas, "Quantitative assessment of stress through EEG during a virtual reality stress-relax session," *Frontiers in Computational Neuroscience*, vol. 15, Jul. 2021. DOI: `10.3389/fncom.2021.684423`. [Online]. Available: `https://doi.org/10.3389/fncom.2021.684423`.

[42] P. Pinti, I. Tachtsidis, A. Hamilton, *et al.*, "The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience," en, *Ann. N. Y. Acad. Sci.*, vol. 1464, no. 1, pp. 5–29, Mar. 2020.

[43] T. Wilcox and M. Biondi, "fNIRS in the developmental sciences," en, *Wiley Interdiscip. Rev. Cogn. Sci.*, vol. 6, no. 3, pp. 263–283, May 2015.

[44] K. Dedovic, C. D'Aguiar, and J. C. Pruessner, "What stress does to your brain: A review of neuroimaging studies," *The Canadian Journal of Psychiatry*, vol. 54, no. 1, pp. 6–15, Jan. 2009, ISSN: 1497-0015. DOI: `10.1177/070674370905400104`. [Online]. Available: `http://dx.doi.org/10.1177/070674370905400104`.

[45] X. Cui, S. Bray, and A. L. Reiss, "Speeded near infrared spectroscopy (NIRS) response detection," en, *PLoS One*, vol. 5, no. 11, e15474, Nov. 2010.

[46] W. B. Baker, A. B. Parthasarathy, D. R. Busch, R. C. Mesquita, J. H. Greenberg, and A. G. Yodh, "Modified beer-lambert law for blood flow," *Biomedical Optics Express*, vol. 5, no. 11, p. 4053, Oct. 2014, ISSN: 2156-7085. DOI: `10.1364/boe.5.004053`. [Online]. Available: `http://dx.doi.org/10.1364/BOE.5.004053`.

[47] P. Jack C. Waymire, *Acetylcholine neurotransmission (section 1, chapter 11)*, Neuroscience Online: An Electronic Textbook for the Neurosciences, Department of Neurobiology and Anatomy, The University of Texas Medical School at Houston, 2024. [Online]. Available: https://nba.uth.tmc.edu/neuroscience/m/s1/chapter11.html.

[48] J. Bertilsson, D. C. Niehorster, P. J. Fredriksson, *et al.*, "Stress levels escalate when repeatedly performing tasks involving threats," *Frontiers in Psychology*, vol. 10, Jul. 2019, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2019.01562. [Online]. Available: http://dx.doi.org/10.3389/fpsyg.2019.01562.

[49] D. Bozovic, M. Racic, and N. Ivkovic, "Salivary cortisol levels as a biological marker of stress reaction," *Med Arch*, vol. 67, no. 5, pp. 374–377, 2013.

[50] D. Leiner, A. Fahr, and H. Früh, "EDA positive change: A simple algorithm for electrodermal activity to measure general audience arousal during media exposure," en, *Commun. Methods Meas.*, vol. 6, no. 4, pp. 237–250, Oct. 2012.

[51] T. Fekete, D. Rubin, J. M. Carlson, and L. R. Mujica-Parodi, "The nirs analysis package: Noise reduction and statistical inference," *PLoS ONE*, vol. 6, no. 9, X.-N. Zuo, Ed., e24322, Sep. 2011, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0024322. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0024322.

[52] J. J. Braithwaite, D. G. Watson, R. Jones, and M. Rowe, "A guide for analysing electrodermal activity (eda) and skin conductance responses (scrs) for psychological experiments," University of Birmingham, Birmingham, UK, Technical Report, 2015.

[53] N. Thammasan, I. V. Stuldreher, E. Schreuders, M. Giletta, and A.-M. Brouwer, "A usability study of physiological measurement in school using wearable sensors," *Sensors*, vol. 20, no. 18, p. 5380, Sep. 2020, ISSN: 1424-8220. DOI: 10.3390/s20185380. [Online]. Available: http://dx.doi.org/10.3390/s20185380.

[54] J. Wijsman, B. Grundlehner, J. Penders, and H. Hermens, "Trapezius muscle emg as predictor of mental stress," *ACM Transactions on Embedded Computing Systems*, vol. 12, no. 4, pp. 1–20, Jun. 2013, ISSN: 1558-3465. DOI: 10.1145/2485984.2485987. [Online]. Available: http://dx.doi.org/10.1145/2485984.2485987.

[55] S. Wüst, I. S. Federenko, E. F. van Rossum, J. W. Koper, and D. H. Hellhammer, "Habituation of cortisol responses to repeated psychosocial stress—further characterization and impact of genetic factors," *Psychoneuroendocrinology*, vol. 30, no. 2, pp. 199–211, Feb. 2005, ISSN: 0306-4530. DOI: 10.1016/j.psyneuen.2004.07.002. [Online]. Available: http://dx.doi.org/10.1016/j.psyneuen.2004.07.002.

[56] D. Roger and J. Jamieson, "Individual differences in delayed heart-rate recovery following stress: The role of extraversion, neuroticism and emotional control," *Personality and Individual Differences*, vol. 9, no. 4, pp. 721–726, Jan. 1988, ISSN: 0191-8869. DOI: 10.1016/0191-8869(88)90061-x. [Online]. Available: http://dx.doi.org/10.1016/0191-8869(88)90061-X.

[57] J. Koenig and J. F. Thayer, "Sex differences in healthy human heart rate variability: A meta-analysis," *Neuroscience & Biobehavioral Reviews*, vol. 64, pp. 288–310, May 2016, ISSN: 0149-7634. DOI: 10.1016/j.neubiorev.2016.03.007. [Online]. Available: http://dx.doi.org/10.1016/j.neubiorev.2016.03.007.

[58] T. Iqbal, A. J. Simpkin, D. Roshan, *et al.*, "Stress monitoring using wearable sensors: A pilot study and stress-predict dataset," *Sensors*, vol. 22, no. 21, p. 8135, Oct. 2022, ISSN: 1424-8220. DOI: 10.3390/s22218135. [Online]. Available: http://dx.doi.org/10.3390/s22218135.

[59] O. George and G. F. Koob, "Individual differences in prefrontal cortex function and the transition from drug use to drug dependence," *Neuroscience amp; Biobehavioral Reviews*, vol. 35, no. 2, pp. 232–247, Nov. 2010, ISSN: 0149-7634. DOI: 10.1016/j.neubiorev.2010.05.002. [Online]. Available: http://dx.doi.org/10.1016/j.neubiorev.2010.05.002.

[60] B. M. Kudielka, D. Hellhammer, and S. Wüst, "Why do we respond so differently? reviewing determinants of human salivary cortisol responses to challenge," *Psychoneuroendocrinology*, vol. 34, no. 1, pp. 2–18, Jan. 2009, ISSN: 0306-4530. DOI: 10.1016/j.psyneuen.2008.10.004. [Online]. Available: http://dx.doi.org/10.1016/j.psyneuen.2008.10.004.

[61] G. Lamotte, C. J. Boes, P. A. Low, and E. A. Coon, "The expanding role of the cold pressor test: A brief history," *Clinical Autonomic Research*, vol. 31, no. 2, pp. 153–155, Mar. 2021, ISSN: 1619-1560. DOI: 10.1007/s10286-021-00796-4. [Online]. Available: http://dx.doi.org/10.1007/s10286-021-00796-4.

[62] B. M. Kudielka and C. Kirschbaum, "Sex differences in hpa axis responses to stress: A review," *Biological Psychology*, vol. 69, no. 1, pp. 113–132, 2005, Current Trends in Women's Health Research, ISSN: 0301-0511. DOI: https://doi.org/10.1016/j.biopsycho.2004.11.009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S030105110400170X.

[63] A. P. Allen, P. J. Kennedy, S. Dockray, J. F. Cryan, T. G. Dinan, and G. Clarke, "The trier social stress test: Principles and practice," en, *Neurobiol. Stress*, vol. 6, pp. 113–126, Feb. 2017.

[64] L. Stappen, A. Baird, L. Christ, E.-M. Meßner, and B. Schuller, *Ulm-TSST dataset - raw data (MuSe2021)*, 2021.

[65] O. Kelly, K. Matheson, A. Martinez, Z. Merali, and H. Anisman, "Psychosocial stress evoked by a virtual audience: Relation to neuroendocrine activity," en, *Cyberpsychol. Behav.*, vol. 10, no. 5, pp. 655–662, Oct. 2007.

[66] P. Zimmer, B. Buttlar, G. Halbeisen, E. Walther, and G. Domes, "Virtually stressed? a refined virtual reality adaptation of the trier social stress test (TSST) induces robust endocrine responses," en, *Psychoneuroendocrinology*, vol. 101, pp. 186–192, Mar. 2019.

[67] Y. Shiban, J. Diemer, S. Brandl, R. Zack, A. Mühlberger, and S. Wüst, "Trier social stress test in vivo and in virtual reality: Dissociation of response domains," en, *Int. J. Psychophysiol.*, vol. 110, pp. 47–55, Dec. 2016.

[68] E. Montero-López, A. Santos-Ruiz, M. C. García-Ríos, R. Rodríguez-Blázquez, M. Pérez-García, and M. I. Peralta-Ramírez, "A virtual reality approach to the trier social stress test: Contrasting two distinct protocols," en, *Behav. Res. Methods*, vol. 48, no. 1, pp. 223–232, Mar. 2016.

[69] O. D. Kothgassner, A. Felnhofer, H. Hlavacs, *et al.*, "Salivary cortisol and cardiovascular reactivity to a public speaking task in a virtual and real-life environment," *Comput. Human Behav.*, vol. 62, pp. 124–135, Sep. 2016.

[70] O. D. Kothgassner, H. Hlavacs, L. Beutl, L. M. Glenk, R. Palme, and A. Felnhofer, "Two experimental virtual paradigms for stress research: Developing avatar-based approaches for interpersonal and evaluative stressors," in *Entertainment Computing - ICEC 2016*, ser. Lecture notes in computer science, Cham: Springer International Publishing, 2016, pp. 51–62.

[71] S. Liszio, "Relaxation, distraction, and fun: Improving well-being in situations of acute emotional distress with virtual reality," en, Ph.D. dissertation, 2021. DOI: 10.17185/DUEPUBLICO/74774. [Online]. Available: https://duepublico2.uni-due.de/receive/duepublico_mods_00074774.

[72] K. Linnig, S. Seel, B. von Dawans, *et al.*, "Open tsst vr: Psychobiological reactions to an open version of the trier social stress test in virtual reality," Mar. 2024. DOI: 10.31234/osf.io/fpe2h. [Online]. Available: http://dx.doi.org/10.31234/osf.io/fpe2h.

[73] C. D. Spielberger, *State-Trait Anxiety Inventory: Bibliography*, 2nd ed. Palo Alto, CA: Consulting Psychologists Press, 1989.

[74] C. Carmin and R. L. Ownby, "Chapter 2 - assessment of anxiety in older adults," in *Handbook of Assessment in Clinical Gerontology (Second Edition)*, P. A. Lichtenberg, Ed., Second Edition, San Diego: Academic Press, 2010, pp. 45–60, ISBN: 978-0-12-374961-1. DOI: https://doi.org/10.1016/B978-0-12-374961-1.10002-8. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123749611100028.

[75] W. S. Helton and K. Näswall, "Short stress state questionnaire," en, *Eur. J. Psychol. Assess.*, vol. 31, no. 1, pp. 20–30, Jun. 2015.

[76] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: The PANAS scales," en, *J. Pers. Soc. Psychol.*, vol. 54, no. 6, pp. 1063–1070, Jun. 1988.

[77] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," en, *J. Behav. Ther. Exp. Psychiatry*, vol. 25, no. 1, pp. 49–59, Mar. 1994.

[78] A. Tazarv, S. Labbaf, A. Rahmani, N. Dutt, and M. Levorato, "Active reinforcement learning for personalized stress monitoring in everyday settings," Apr. 2023. arXiv: 2305.00111 [cs.LG].

[79] B. D. Womack and J. H. L. Hansen, "Classification of speech under stress using target driven features," *Speech Commun.*, vol. 20, pp. 131–150, 1996. [Online]. Available: https://api.semanticscholar.org/CorpusID:17615543.

[80] K. H. Kim, S. W. Bang, and S. R. Kim, "Emotion recognition system using short-term monitoring of physiological signals," *Med. Biol. Eng. Comput.*, vol. 42, pp. 419–427, 2004.

[81] M.-H. Lee, G. Yang, H.-K. Lee, and S. Bang, "Development stress monitoring system based on personal digital assistant (pda)," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2004, pp. 2364–2367.

[82] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 2, pp. 156–166, Jun. 2005.

[83] J. Zhai and A. Barreto, "Stress detection in computer users based on digital signal processing of noninvasive physiological variables," in *Proc. Conf. IEEE Eng. Med. Biol. Soc.*, 2006, pp. 1355–1358.

[84] . Ren, A. Barreto, Y. Gao, and M. Adjouadi, "Affective assessment by digital processing of the pupil diameter," *IEEE Trans. Affect. Comput.*, vol. 4, no. 1, pp. 2–14, Jan. 2013.

[85] S. Baltaci and D. Gokcay, "Stress detection in human–computer interaction: Fusion of pupil dilation and facial temperature features," *Int. J. Hum.–Comput. Interaction*, vol. 32, pp. 956–966, 2016.

[86] K. Lai, S. N. Yanushkevich, and V. P. Shmerko, "Intelligent stress monitoring assistant for first responders," *IEEE Access*, vol. 9, pp. 25 314–25 329, 2021. DOI: 10.1109/access.2021.3057578.

[87] S. Ghosh *et al.*, "Classification of mental stress from wearable physiological sensors using image-encoding-based deep neural network," *Biosensors*, vol. 12, no. 12, p. 1153, 2022. DOI: 10.3390/bios12121153.

[88] M. Jaén-Vargas, K. M. Reyes Leiva, F. Fernandes, *et al.*, "Effects of sliding window variation in the performance of acceleration-based human activity recognition using deep learning models," *PeerJ Computer Science*, vol. 8, e1052, Aug. 2022, ISSN: 2376-5992. DOI: 10.7717/peerj-cs.1052. [Online]. Available: http://dx.doi.org/10.7717/peerj-cs.1052.

[89] P. Sa-Nguannarm, E. Elbasani, and J.-D. Kim, "Human activity recognition for analyzing stress behavior based on bi-lstm," *Technology and Health Care*, vol. 31, no. 5, pp. 1997–2007, 2023.

[90] L. Malviya, S. Mal, R. Kumar, *et al.*, "Mental stress level detection using lstm for wesad dataset," in *Proceedings of Data Analytics and Management: ICDAM 2022*, Springer, 2023, pp. 243–250.

[91] L. Xia, Y. Feng, Z. Guo, *et al.*, "Mulhita: A novel multiclass classification framework with multibranch lstm and hierarchical temporal attention for early detection of mental stress," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, pp. 9657–9670, 2023. DOI: 10.1109/TNNLS.2022.3159573.

[92] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[93] X. Zhang, X. Wei, Z. Zhou, *et al.*, "Dynamic alignment and fusion of multimodal physiological patterns for stress recognition," *IEEE Transactions on Affective Computing*, vol. 15, no. 2, pp. 685–696, 2024. DOI: 10.1109/TAFFC.2023.3290177.

[94] R. Child, S. Gray, A. Radford, and I. Sutskever, *Generating long sequences with sparse transformers*, 2019. DOI: 10.48550/ARXIV.1904.10509. [Online]. Available: https://arxiv.org/abs/1904.10509.

[95] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, *Transformers are rnns: Fast autoregressive transformers with linear attention*, 2020. DOI: 10.48550/ARXIV.2006.16236. [Online]. Available: https://arxiv.org/abs/2006.16236.

[96] I. Beltagy, M. E. Peters, and A. Cohan, *Longformer: The long-document transformer*, 2020. DOI: 10.48550/ARXIV.2004.05150. [Online]. Available: https://arxiv.org/abs/2004.05150.

[97] A. Bhatti *et al.*, "Attentive cross-modal connections for deep multimodal wearable-based emotion recognition," in *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, IEEE, 2021, pp. 1–5. DOI: 10.1109/aciiw52867.2021.9666360.

[98] L. Huynh, T. Nguyen, T. Nguyen, S. Pirttikangas, and P. Siirtola, "Stressnas: Affect state and stress detection using neural architecture search," in *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*, ser. UbiComp '21, ACM, Sep. 2021. DOI: 10.1145/3460418.3479320. [Online]. Available: http://dx.doi.org/10.1145/3460418.3479320.

[99] B. Dasarathy, "Sensor fusion potential exploitation-innovative architectures and illustrative applications," *Proceedings of the IEEE*, vol. 85, no. 1, pp. 24–38, 1997, ISSN: 0018-9219. DOI: 10.1109/5.554206. [Online]. Available: http://dx.doi.org/10.1109/5.554206.

[100] M. Yan, Z. Deng, B. He, C. Zou, J. Wu, and Z. Zhaoju, "Emotion classification with multichannel physiological signals using hybrid feature and adaptive decision fusion," *Biomedical Signal Processing and Control*, vol. 71, p. 103 235, Jan. 2022. DOI: 10.1016/j.bspc.2021.103235.

[101] T. Goncalves, I. Rio-Torto, L. F. Teixeira, and J. S. Cardoso, "A survey on attention mechanisms for medical applications: Are we moving toward better algorithms?" *IEEE Access*, vol. 10, pp. 98 909–98 935, 2022, ISSN: 2169-3536. DOI: 10.1109/access.2022.3206449. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2022.3206449.

[102] P. Gao, Z. Jiang, H. You, *et al.*, "Dynamic fusion with intra- and inter-modality attention flow for visual question answering," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA: IEEE, Jun. 2019.

[103] Z. Zohourianshahzadi and J. K. Kalita, "Neural attention for image captioning: Review of outstanding methods," *Artificial Intelligence Review*, vol. 55, pp. 3833–3862, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:244729197.

[104] H.-G. Kim, E.-J. Cheon, D.-S. Bai, Y. H. Lee, and B.-H. Koo, "Stress and heart rate variability: A meta-analysis and review of the literature," *Psychiatry Investig.*, vol. 15, no. 3, pp. 235–245, Mar. 2018.

[105] H. Yang, H. Yu, K. Sridhar, T. Vaessen, I. Myin-Germeys, and A. Sano, "More to less (m2l): Enhanced health recognition in the wild with reduced modality of wearable sensors," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2022, pp. 3253–3256. DOI: 10.1109/EMBC48229.2022.9871472.

[106] P. Shi, M. Hu, X. Shi, and F. Ren, "Deep modular co-attention shifting network for multimodal sentiment analysis," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 20, no. 4, Jan. 2024, ISSN: 1551-6857. DOI: 10.1145/3634706. [Online]. Available: https://doi.org/10.1145/3634706.

[107] H. Han, J. Yang, and W. Slamu, "Cascading modular multimodal cross-attention network for rumor detection," in *2023 IEEE International Conference on Control, Electronics and Computer Technology (ICCECT)*, 2023, pp. 974–980. DOI: 10.1109/ICCECT57938.2023.10140211.

[108] H. Liu, F. Wu, W. Wang, *et al.*, "Nrpa: Neural recommendation with personalized attention," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '19, ACM, Jul. 2019. DOI: 10.1145/3331184.3331371. [Online]. Available: http://dx.doi.org/10.1145/3331184.3331371.

[109] H. Yu and A. Sano, "Passive sensor data based future mood, health, and stress prediction: User adaptation using deep learning," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020, pp. 5884–5887. DOI: 10.1109/EMBC44109.2020.9176242.

[110] S. Salehizadeh, D. Dao, J. Bolkhovsky, C. Cho, Y. Mendelson, and K. H. Chon, "A novel time-varying spectral filtering algorithm for reconstruction of motion artifact corrupted heart rate signals during intense physical activities using a wearable photoplethysmogram sensor," *Sensors*, vol. 16, no. 1, p. 10, 2016.

[111] G. Bianchi and R. Sorrentino, *Electronic Filter Simulation & Design*. McGraw-Hill Professional, 2007, pp. 17–20, ISBN: 978-0-07-149467-0.

[112] S. D. Yusuf, F. C. Maduakolam, I. Umar, A. Z. Loko, and L. W. Lumbi, "Comparative analysis of savitzky-golay and butterworth filters for electrocardiogram de-noising using daubechies wavelets," *Asian J. Res. Cardiovasc. Dis.*, vol. 2, no. 1, pp. 15–29, 2020.

[113] M. A. Rahman, M. A. Rashid, and M. Ahmad, "Selecting the optimal conditions of savitzky–golay filter for fnirs signal," *Biocybernetics and Biomedical Engineering*, vol. 39, no. 3, pp. 624–637, 2019, ISSN: 0208-5216. DOI: https://doi.org/10.1016/j.bbe.2019.06.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0208521618305667.

[114] N. Rashid, L. Chen, M. Dautta, A. Jimenez, P. Tseng, and M. A. A. Faruque, "Feature augmented hybrid cnn for stress recognition using wrist-based photoplethysmography sensor," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, 2021, pp. 2374–2377.

[115] A. Chatterjee and U. K. Roy, "Ppg based heart rate algorithm improvement with butterworth iir filter and savitzky-golay fir filter," in *2018 2nd International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech)*, 2018, pp. 1–6. DOI: 10.1109/IEMENTECH.2018.8465225.

[116] L. Cao, H. Wu, S. Chen, *et al.*, "A novel deep learning method based on an overlapping time window strategy for brain–computer interface-based stroke rehabilitation," *Brain Sciences*, vol. 12, no. 11, p. 1502, Nov. 2022, ISSN: 2076-3425. DOI: 10.3390/brainsci12111502. [Online]. Available: http://dx.doi.org/10.3390/brainsci12111502.

[117] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, 2018. DOI: 10.48550/ARXIV.1810.04805. [Online]. Available: https://arxiv.org/abs/1810.04805.

[118] PyTorch Contributors, *Multiheadattention*, Accessed: 2023-07-02, 2023. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.MultiheadAttention.html.

[119] PyTorch Contributors, *Kv cache*, Accessed: 2023-07-02, 2023. [Online]. Available: https://pytorch.org/torchtune/stable/_modules/torchtune/modules/kv_cache.html.

[120] F. Chen, G. Datta, S. Kundu, and P. Beerel, *Self-attentive pooling for efficient deep learning*, 2022. DOI: 10.48550/ARXIV.2209.07659. [Online]. Available: https://arxiv.org/abs/2209.07659.

[121] S. Oviatt, B. Schuller, P. Cohen, D. Sonntag, G. Potamianos, and A. Krüger, *The Handbook of Multimodal-Multisensor Interfaces, Volume 2: Signal Processing, Architectures, and Detection of Emotion and Cognition*. Morgan & Claypool, 2018.

[122] A. Pakrashi and B. Mac Namee, "Kalman filter-based heuristic ensemble (kfhe): A new perspective on multi-class ensemble classification using kalman filters," *Information Sciences*,

vol. 485, pp. 456–485, Jun. 2019, ISSN: 0020-0255. DOI: 10.1016/j.ins.2019.02.017. [Online]. Available: http://dx.doi.org/10.1016/j.ins.2019.02.017.

[123] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2999–3007. DOI: 10.1109/ICCV.2017.324.

[124] The HDF Group, *Introduction to hdf5*, Accessed: 2024-09-04, The HDF Group, 2024. [Online]. Available: https://docs.hdfgroup.org/hdf5/v1_14/_intro_h_d_f5.html.

[125] M. H. Imam *et al.*, "Effect of ecg-derived respiration (edr) on modeling ventricular repolarization dynamics in different physiological and psychological conditions," *Medical & Biological Engineering & Computing*, vol. 52, no. 10, pp. 851–860, Aug. 2014. DOI: 10.1007/s11517-014-1188-0. [Online]. Available: https://doi.org/10.1007/s11517-014-1188-0.

[126] L. Holper *et al.*, "Short-term pulse rate variability is better characterized by functional near-infrared spectroscopy than by photoplethysmography," *J. Biomed. Opt.*, vol. 21, no. 9, p. 091 308, 2016.

[127] G. E. Billman, "The lf/hf ratio does not accurately measure cardiac sympatho-vagal balance," *Front. Physiol.*, vol. 4, p. 26, 2013.

[128] M. Ahmed, M. Grillo, A. Taebi, M. Kaya, and P. Thibbotuwawa Gamage, "A comprehensive analysis of trapezius muscle emg activity in relation to stress and meditation," *BioMedInformatics*, vol. 4, no. 2, pp. 1047–1058, 2024, ISSN: 2673-7426. DOI: 10.3390/biomedinformatics4020058. [Online]. Available: https://www.mdpi.com/2673-7426/4/2/58.

[129] R. Jerath, M. Syam, and S. Ahmed, "The future of stress management: Integration of smartwatches and hrv technology," *Sensors*, vol. 23, no. 17, p. 7314, Aug. 2023, ISSN: 1424-8220. DOI: 10.3390/s23177314. [Online]. Available: http://dx.doi.org/10.3390/s23177314.

[130] S. Heinzel, F. B. Haeussinger, T. Hahn, A.-C. Ehlis, M. M. Plichta, and A. J. Fallgatter, "Variability of (functional) hemodynamics as measured with simultaneous fnirs and fmri during intertemporal choice," *NeuroImage*, vol. 71, pp. 125–134, 2013, ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2012.12.074. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S105381191300013X.

[131] M. D. Hssayeni and B. Ghoraani, "Multi-modal physiological data fusion for affect estimation using deep learning," *IEEE Access*, vol. 9, pp. 21 642–21 652, 2021, ISSN: 2169-3536. DOI: 10.1109/access.2021.3055933. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2021.3055933.

[132] N. Rashid, T. Mortlock, and M. A. A. Faruque, "Stress detection using context-aware sensor fusion from wearable devices," *IEEE Internet of Things Journal*, vol. 10, no. 16, pp. 14 114–14 127, Aug. 2023, ISSN: 2372-2541. DOI: 10.1109/jiot.2023.3265768. [Online]. Available: http://dx.doi.org/10.1109/JIOT.2023.3265768.

[133] V. T. Ninh, S. Smyth, M. T. Tran, and C. Gurrin, "Analysing the performance of stress detection models on consumer-grade wearable devices," *arXiv preprint arXiv:2203.09669*, 2022. DOI: 10.3233/faia210050.

[134] R. Sah and H. Ghasemzadeh, *Stress classification and personalization: Getting the most out of the least*, Jul. 2021.

[135] A. Liapis, E. Faliagka, C. Katsanos, C. Antonopoulos, and N. Voros, "Detection of subtle stress episodes during ux evaluation: Assessing the performance of the wesad bio-signals dataset," in *Human-Computer Interaction–INTERACT 2021: 18th IFIP TC 13 International Conference, Bari, Italy, August 30–September 3, 2021, Proceedings, Part III 18*, Springer International Publishing, 2021, pp. 238–247. DOI: 10.1007/978-3-030-85613-7_17.

[136] M. Albaladejo-González, J. A. Ruipérez-Valiente, and F. Gómez Mármol, "Evaluating different configurations of machine learning models and their transfer learning capabilities for stress detection using heart rate," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 8, pp. 11 011–11 021, 2023. DOI: 10.1007/s12652-022-04365-z.

[137] M. O. Dobrokhvalov and A. Y. Filatov, "Two-branch convolutional networks for stress detection from biomedical data," in *2024 XXVII International Conference on Soft Computing and Measurements (SCM)*, 2024, pp. 449–452. DOI: 10.1109/SCM62608.2024.10554273.

[138] A. Almadhor *et al.*, "Wrist-based electrodermal activity monitoring for stress detection using federated learning," *Sensors*, vol. 23, no. 8, p. 3984, 2023. DOI: 10.3390/s23083984.

[139] S. Samyoun, A. S. Mondol, and J. A. Stankovic, "Stress detection via sensor translation," in *2020 16th International Conference on Distributed Computing in Sensor Systems (DCOSS)*, IEEE, 2020, pp. 19–26.

[140] Y. Hasanpoor, K. Motaman, B. Tarvirdizadeh, K. Alipour, and M. Ghamari, "Stress detection using ppg signal and combined deep cnn-mlp network," in *2022 29th National and 7th International Iranian Conference on Biomedical Engineering (ICBME)*, IEEE, 2022, pp. 223–228. DOI: 10.1109/icbme57741.2022.10052957.

[141] G. Vos, K. Trinh, Z. Sarnyai, and M. Rahimi Azghadi, "Ensemble machine learning model trained on a new synthesized dataset generalizes well for stress prediction using wearable devices," *Journal of Biomedical Informatics*, vol. 148, p. 104 556, 2023, ISSN: 1532-0464. DOI: https://doi.org/10.1016/j.jbi.2023.104556. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046423002770.

[142] J. Birjandtalab, D. Cogan, M. B. Pouyan, and M. Nourani, "A non-eeg biosignals dataset for assessment and visualization of neurological status," in *2016 IEEE International Workshop on Signal Processing Systems (SiPS)*, 2016, pp. 110–114. DOI: 10.1109/SiPS.2016.27.

[143] S. Koldijk, M. Sappelli, S. Verberne, M. A. Neerincx, and W. Kraaij, "The swell knowledge work dataset for stress and user modeling research," in *Proceedings of the 16th International Conference on Multimodal Interaction*, ser. ICMI '14, Istanbul, Turkey: Association for Computing Machinery, 2014, pp. 291–298, ISBN: 9781450328852. DOI: 10.1145/2663204.2663257. [Online]. Available: https://doi.org/10.1145/2663204.2663257.

[144] K. Dahal, B. Bogue-Jimenez, and A. Doblas, "Global stress detection framework combining a reduced set of hrv features and random forest model," *Sensors*, vol. 23, no. 11, 2023, ISSN: 1424-8220. [Online]. Available: https://www.mdpi.com/1424-8220/23/11/5220.

[145] C.-Y. Chang, C. Hsu, Y. C. Wu, S. Wang, D. Tsui, and T.-P. Jung, *Online mental stress detection using frontal-channel eeg recordings in a classroom scenario*, 2024. DOI: 10.48550/ARXIV.2405.11394. [Online]. Available: https://arxiv.org/abs/2405.11394.

[146] T. Nozawa and Y. Miyake, "Capturing individual differences in prefrontal activity with wearable fnirs for daily use," in *2020 13th International Conference on Human System Interaction (HSI)*, 2020, pp. 249–254. DOI: 10.1109/HSI49210.2020.9142689.

[147] G. Vos, K. Trinh, Z. Sarnyai, and M. Rahimi Azghadi, "Generalizable machine learning for stress monitoring from wearable devices: A systematic literature review," *International Journal of Medical Informatics*, vol. 173, p. 105 026, 2023, ISSN: 1386-5056. DOI: https://doi.org/10.1016/j.ijmedinf.2023.105026. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1386505623000436.

[148] S. Hosseini, R. Gottumukkala, S. Katragadda, *et al.*, "A multimodal sensor dataset for continuous stress detection of nurses in a hospital," *Scientific Data*, vol. 9, no. 1, Jun. 2022, ISSN: 2052-4463. DOI: 10.1038/s41597-022-01361-y. [Online]. Available: http://dx.doi.org/10.1038/s41597-022-01361-y.

[149] NVIDIA Corporation, *Cuda c programming guide*, Accessed: 2024-09-10, 2023.

[150] W. R. Lovallo, T. L. Whitsett, M. al'Absi, B. H. Sung, A. S. Vincent, and M. F. Wilson, "Caffeine stimulation of cortisol secretion across the waking hours in relation to caffeine intake levels," en, *Psychosom. Med.*, vol. 67, no. 5, pp. 734–739, Sep. 2005.

[151] B. M. Kudielka, N. C. Schommer, D. H. Hellhammer, and C. Kirschbaum, "Acute hpa axis responses, heart rate, and mood changes to psychosocial stress (tsst) in humans at different times of day," *Psychoneuroendocrinology*, vol. 29, no. 8, pp. 983–992, Sep. 2004, ISSN: 0306-4530. DOI: 10.1016/j.psyneuen.2003.08.009. [Online]. Available: http://dx.doi.org/10.1016/j.psyneuen.2003.08.009.

[152] J. Koenig, M. N. Jarczok, W. Kuhn, *et al.*, "Impact of caffeine on heart rate variability: A systematic review," *Journal of Caffeine Research*, vol. 3, no. 1, pp. 22–37, Mar. 2013, ISSN: 2156-5368. DOI: 10.1089/jcr.2013.0009. [Online]. Available: http://dx.doi.org/10.1089/jcr.2013.0009.

[153] M. J. Tipton, A. Harper, J. F. R. Paton, and J. T. Costello, "The human ventilatory response to stress: Rate or depth?" en, *J. Physiol.*, vol. 595, no. 17, pp. 5729–5752, Sep. 2017.

[154] N. Ma, D. F. Dinges, M. Basner, and H. Rao, "How acute total sleep loss affects the attending brain: A meta-analysis of neuroimaging studies," en, *Sleep*, vol. 38, no. 2, pp. 233–240, Feb. 2015.