Chapter 1: Introduction and Objectives

1.1 Background and Significance of Chronic Stress

Chronic stress has emerged as one of the most pressing health concerns of the twenty-first century, affecting individuals across diverse demographics and fundamentally reshaping public health priorities (Mariotti, 2015; Yale Medicine, 2024). While acute stress represents a normal adaptive response to immediate threats, prolonged exposure to stressors disrupts biological homeostasis and significantly increases the risk of developing serious health conditions (Ajibewa et al., 2024; Vaccarino, 2024). The consequences of chronic stress are far-reaching, encompassing cardiovascular disease, immune system dysfunction, mood disorders, cognitive impairment, metabolic dysregulation, and sleep disturbances (Mariotti, 2015; British Heart Foundation, 2024; Carola, 2024).

The societal burden of chronic stress is substantial and quantifiable. In the United Kingdom during 2023/24, over 776,000 workers reported experiencing work-related stress, depression, or anxiety, accounting for more than half of all work-related ill health cases and imposing an estimated annual cost of £28 billion on the UK economy (Priory Group, 2025). Beyond its direct physiological effects, chronic stress exacerbates maladaptive behaviours including substance misuse, poor dietary choices, physical inactivity, and social withdrawal, thereby amplifying health risks across both social and professional domains (British Heart Foundation, 2024; SingleCare, 2025). The persistent elevation of stress hormones, particularly cortisol, leads to cumulative organ damage and progressively increases disease susceptibility, creating a self-reinforcing cycle that intensifies without timely intervention (Russell et al., 2020; Cleveland Clinic, 2025).

1.2 Neurobiology of Stress: Cortisol and the HPA Axis

Cortisol serves as the primary effector hormone of the hypothalamic-pituitary-adrenal (HPA) axis, orchestrating the body's comprehensive physiological response to stress (James et al., 2023; Herman, 2016). When a stressor is perceived, the hypothalamus secretes corticotropin-releasing hormone (CRH), which stimulates the anterior pituitary gland to release adrenocorticotropic hormone (ACTH), subsequently prompting cortisol secretion from the adrenal cortex (Wikipedia, 2025; Karin et al., 2020).

Under normal physiological conditions, cortisol facilitates adaptive stress responses by mobilising glucose reserves, enhancing cardiovascular function, and temporarily suppressing non-essential processes such as immune function and reproduction (Harvard Health, 2024; James et al., 2023). This system is regulated through a sophisticated negative feedback loop mediated by the hippocampus, which ensures appropriate cessation of cortisol production once the stressor has passed (Herman, 2016; Wikipedia, 2025). However, chronic stress fundamentally disrupts this regulatory system, maintaining persistently elevated cortisol levels that induce hippocampal atrophy, memory deficits, immune suppression, and

heightened susceptibility to both cardiovascular and mood disorders (Karin et al., 2020; Cleveland Clinic, 2025; Herman, 2016; Wikipedia, 2025). Individuals exhibiting HPA axis dysfunction demonstrate markedly increased risks for cardiovascular events, metabolic syndrome, and neurodegenerative conditions (Tsai et al., 2024; Cleveland Clinic, 2025).

1.3 Traditional Stress Assessment Limitations

Conventional approaches to stress assessment typically rely on self-report questionnaires, behavioural observation, and invasive biological sampling; however, each method possesses inherent limitations that constrain their utility for continuous monitoring (Jaber et al., 2022; Lazarou et al., 2024). Self-report measures are susceptible to recall bias, social desirability effects, and subjective interpretation, fundamentally limiting their objectivity (Naegelin et al., 2023; Abd Al-Alim et al., 2024). Observational methods, whilst offering greater objectivity than self-reports, remain vulnerable to observer bias and may elicit behavioural changes in participants aware of being monitored.

Direct biochemical measurement of cortisol through saliva, blood, or urine sampling provides objective quantification but presents significant practical barriers to routine implementation. These methods require specialised equipment, trained personnel, controlled collection environments, and laboratory processing, rendering them unsuitable for continuous, real-world monitoring applications (Saeed et al., 2021; Gu et al., 2022; Hellhammer and Schubert, 2012). The invasive nature of sampling procedures and complex logistical requirements present further obstacles to widespread, longitudinal stress assessment in naturalistic settings.

1.4 Wearable Technology and Physiological Signal Analysis

Recent technological advances have enabled continuous, non-invasive monitoring of physiological proxies for stress responses (Pinge et al., 2024; Seshadri et al., 2019). Contemporary wearable sensors can continuously collect data including electrocardiography (ECG), heart rate variability (HRV), electrodermal activity (EDA), skin temperature, respiration rate, and movement patterns, all of which reflect autonomic nervous system activity and stress responses (Hongn et al., 2025; Sabry et al., 2022).

These physiological signals offer several advantages: they are objective, involuntary, and resistant to conscious manipulation, providing data suitable for automated analysis whilst demonstrating strong correlations with established stress biomarkers, including cortisol (Wijsman et al., 2011; Vos et al., 2023). The integration of artificial intelligence and machine learning techniques facilitates sophisticated pattern recognition and predictive modelling using multimodal physiological signal data, with particular success demonstrated by deep learning architectures such as recurrent neural networks (LSTM, BiLSTM) and attention-based models for stress detection tasks (Kim et al., 2024; Tanwar et al., 2024; Oliver and Dakshit, 2025; Samee et al., 2022).

1.5 Deep Learning for Physiological Signal Analysis

Deep learning methodologies have revolutionised stress detection research by enabling advanced pattern recognition and temporal modelling of complex physiological signals (Kim et al., 2024; Oliver and Dakshit, 2025). Convolutional neural networks (CNNs) excel at extracting relevant spatial features from physiological data and have achieved stress classification accuracies exceeding 90% in controlled experimental settings (Kim et al., 2024; Dilated CNN Study, 2024). Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) architectures demonstrate particular efficacy in analysing sequential data by effectively modelling temporal dependencies inherent in physiological signals (Zhu et al., 2018; Pouromran et al., 2022), with BiLSTM models providing comprehensive feature extraction from time series data through their capacity to process information in both forward and backward temporal directions (Chen et al., 2020; Samee et al., 2022).

1.6 The WESAD Dataset and Experimental Approach

The Wearable Stress and Affect Detection (WESAD) dataset represents a cornerstone resource in affective computing research, providing validated multimodal physiological recordings from 15 participants undergoing the Trier Social Stress Test (TSST), a protocol recognised for reliably inducing cortisol elevations in 70-80% of participants (Schmidt et al., 2018; Oliver and Dakshit, 2025; Allen et al., 2016; Frisch et al., 2015).

The TSST paradigm incorporates public speaking tasks, mental arithmetic challenges, and social-evaluative components, reliably eliciting reproducible physiological and hormonal stress responses (Gu et al., 2022; Hellhammer and Schubert, 2012; Allen et al., 2016; Linares et al., 2020). The WESAD dataset provides comprehensive data including ECG, EDA, electromyography (EMG), respiration, skin temperature, and accelerometer readings sampled at 700 Hz across distinct experimental conditions: baseline, stress induction, amusement, and recovery phases (Schmidt et al., 2018; Dahal et al., 2023). Critically, the dataset includes concurrent salivary cortisol measurements, enabling direct validation of physiological signal patterns against objective hormonal stress markers.

1.7 Rationale and Significance

Despite substantial progress in binary stress classification, a critical gap exists in the literature regarding the prediction of specific cortisol response phases from wearable sensor data. Existing research predominantly focuses on distinguishing stressed from non-stressed states, neglecting the temporal dynamics of the stress response characterised by distinct anticipation, peak, and recovery phases (Yang et al., 2025; Oliver and Dakshit, 2025). Understanding these phases is essential for developing effective early intervention strategies and personalised stress management protocols.

This project addresses this gap by developing predictive models capable of identifying cortisol response phases using exclusively non-invasive wearable sensor inputs. By examining which physiological signal modalities contribute most significantly to accurate phase prediction, this research establishes both the feasibility and optimal implementation strategies for cortisol response monitoring through wearable technology. Such advances are crucial for translating laboratory-based stress detection into practical, real-world applications that can support timely clinical intervention and personalised stress management.

1.8 Research Questions and Objectives

The primary research question guiding this investigation is:

Can the temporal phases of cortisol responseanticipation, peak, and recoverybe accurately predicted from non-invasive wearable physiological sensor signals, and which signal modalities and modelling approaches yield optimal prediction accuracy?

This central question is addressed through the following specific research objectives:

1. **Data Preparation and Feature Engineering**: Systematically preprocess the WESAD dataset, implementing appropriate signal filtering, normalisation, and segmentation techniques whilst creating temporally aligned labels that account for the physiological delay between stressor onset and cortisol elevation.

2. **Feature Contribution Analysis**: Conduct systematic ablation studies to quantify the predictive contribution of individual physiological signal modalities, identifying which sensors provide the most valuable information for cortisol phase detection and determining the minimal sensor configuration necessary for accurate prediction.

3. **Comprehensive Performance Assessment**: Validate model performance using standard classification metrics including accuracy, precision, recall, F1-score, and confusion matrices, complemented by statistical significance testing to ensure findings represent genuine predictive capacity rather than chance performance.

1.9 Methodology Overview

The methodology integrates advanced signal processing techniques, deep learning model development, and systematic feature analysis, utilising the WESAD dataset as the empirical foundation (Schmidt et al., 2018). Raw physiological signals undergo filtering, normalisation, and segmentation into fixed-duration windows. Experimental condition labels are temporally shifted to account for the documented physiological delay between stressor onset and measurable cortisol elevation (Gu et al., 2022; Allen et al., 2016).

Both traditional machine learning models (Random Forest, Support Vector Machines) and deep learning architectures (BiLSTM, Transformer) are developed, with systematic hyperparameter optimisation conducted to maximise performance (Oliver and Dakshit, 2025). Feature contribution is assessed through ablation studies that systematically remove individual signal modalities to quantify their impact on predictive accuracy. Model robustness is evaluated through k-fold cross-validation and leave-one-subject-out validation, with statistical significance testing employed to ensure generalisability beyond the training data.

1.10 Ethical Considerations and Data Management

This investigation adheres to rigorous ethical standards by exclusively utilising the pre-approved WESAD dataset, which was collected under appropriate ethical oversight (Schmidt et al., 2018). No new human participant data is collected as part of this research. All data handling procedures comply with UK General Data Protection Regulation (GDPR) requirements and established best-practice protocols, with participant privacy ensured through data aggregation and anonymisation. Should future extensions of this work require additional data collection, explicit informed consent procedures and formal ethical review would precede any data gathering activities.

1.11 Anticipated Contributions

This research makes several novel contributions to the field of stress detection and affective computing. Firstly, it extends beyond binary stress classification to predict specific phases of the cortisol response curve, providing more nuanced information relevant to intervention timing. Secondly, it systematically compares traditional and deep learning approaches specifically for cortisol phase prediction, establishing performance benchmarks for future work. Thirdly, through comprehensive feature ablation studies, it identifies which physiological signals contribute most substantially to accurate phase detection, informing practical wearable device design decisions. Finally, the cross-subject validation approach ensures that findings generalise to new individuals, establishing the real-world applicability of the methodology.

The practical implications extend across multiple domains including occupational health monitoring, clinical psychology assessment, and consumer wellness technology. By enabling continuous, non-invasive tracking of stress hormone dynamics, this work establishes a foundation for next-generation stress intervention systems capable of delivering timely, personalised support.

1.12 Timeline and Project Scope

The project was executed across a four-month timeline encompassing literature review, data preprocessing, model development, comprehensive evaluation, and thesis preparation. The scope remained consistent with the original research plan, with minor refinements including exploration of architectural variants for both BiLSTM and Transformer models and iterative

optimisation of hyperparameters to maximise performance. Anticipated challenges including model overfitting, computational resource constraints, and data quality issues were successfully managed through planned mitigation strategies including dropout regularisation, early stopping, and rigorous cross-validation.

2.0 The Biological Foundation: Cortisol as the Gold Standard Stress Biomarker

2.1 Cortisol's Central Role in Stress Response

Cortisol is the main marker used to measure stress, as it comes from the hypothalamic-pituitary-adrenal (HPA) axis during both mental and physical stress (Herman, 2016; Lightman, 2020). The HPA axis works like a chain reaction: the hypothalamus releases corticotropin-releasing hormone (CRH), which signals the anterior pituitary to release adrenocorticotropic hormone (ACTH). This then triggers the adrenal cortex to produce cortisol (Herman, 2016; Walker & Romanò, 2022).

Cortisol does more than just indicate stress it also plays important roles in helping the body adapt. These include releasing glucose for energy, supporting heart and blood vessel function, regulating the immune system, and improving thinking during difficult situations (Herman, 2016; Adam, 2006). Normally, cortisol levels are controlled by a feedback system. The hippocampus uses glucocorticoid receptors to monitor cortisol in the blood and signals the body to stop the stress response once it is no longer needed (Herman, 2016; Lightman, 2020).

When stress becomes long-term, this regulation breaks down. Cortisol stays high, which turns harmful instead of helpful (Karin et al., 2020; Booij et al., 2016). Studies show that problems in the HPA axis increase the risk of heart disease, metabolic disorders, mood problems, memory issues, and even neurodegenerative diseases (Herman, 2016; Adam, 2006).

However, measuring cortisol in practice is not easy. Standard lab methods like ELISA and LC-MS/MS need special machines, trained staff, and controlled lab conditions. They also take hours or even days to process, which makes them impractical for real-time monitoring or frequent testing (Stalder et al., 2016; Russell et al., 2012).

.

2.1.2The Critical 15–20 Minute Temporal Window: Biological Constraint as Predictive Opportunity

One of the most important features of cortisol dynamics for predictive modeling is the consistent 15–20 minute delay between the start of an acute stressor and the peak in cortisol levels. This delay has been confirmed through standardized stress protocols across different populations (Kirschbaum & Hellhammer, 1994; Allen et al., 2017; Dickerson & Kemeny, 2004). For example, studies using the Trier Social Stress Test (TSST) show that salivary cortisol typically peaks 20–30 minutes after the stress task begins, with the highest levels usually appearing 35–45 minutes after initiation (Allen et al., 2017; Herman, 2016).

Narvaez-Linares et al. (2020) provided the most detailed review of TSST cortisol responses so far. They confirmed that salivary cortisol usually peaks about 20 minutes after the stress begins and recommended frequent sampling at −30, 0, +15, +25, +35, and +45 minutes to capture the full response curve. Their meta-analysis of 68 TSST studies showed that cortisol levels typically start rising around +10 minutes, reach their highest point between +20 and +25 minutes, plateau from +20 to +30 minutes, and then gradually return to baseline by about +60 minutes.

Supporting this, Admon et al. (2017) found in healthy females exposed to the Maastricht Acute Stress Test (MAST) that cortisol rose significantly from baseline at +25, +50, and +65 minutes. This confirmed that the first major rise in HPA-axis activity occurs around 20 minutes after stress begins, with levels staying elevated up to 65 minutes before starting to fall by +100 minutes.

In non-human primates, Verspeek et al. (2021) reported a similar trend. Both urinary and salivary cortisol peaked about 160 minutes after a psychological stressor, showing that while the HPA-axis response in primates may be slower and more extended than in humans, it still follows a consistent, species-specific delay between stress exposure and peak cortisol release.

Taken together, these findings confirm the 15–20-minute delay as a stable feature of cortisol dynamics across human studies, while also showing that protocol differences and species variation can shift the overall timing of the response. This makes the temporal window both reliable for modeling and flexible enough to account for biological differences.

2.3 Evolution and Limitations of Stress Detection Approaches

2.3.1 Traditional Methods: Subjective Measures and Direct Cortisol Assessment

Early approaches to stress assessment mostly relied on self-report tools such as the Perceived Stress Scale, the State-Trait Anxiety Inventory, and visual analog scales. However, these methods face several limitations, including social desirability bias, inaccurate recall, mismatch between actual physiological states and retrospective reporting, and large differences between individuals (Cohen et al., 1983; Kudielka et al., 2007; Abd Al-Alim et al., 2024).

Methodological Critique: Depending on subjective reports creates a fundamental limitation in how stress is scientifically studied. Stress is not only a psychological experience but also a physiological process involving complex neurobiological pathways. These biological changes may occur before a person becomes consciously aware of stress, or even independently of how stressed they feel (Herman, 2016; Lightman, 2020). As a result, self-reports capture only the emotional and cognitive aspects of stress, while failing to reflect the underlying biological mechanisms.

Direct measurement of cortisol offers more objective data, but it also comes with serious challenges for continuous use. Current methods need laboratory equipment, trained staff, and

long processing times, which make them unsuitable for real-time monitoring. Wearable cortisol biosensors are emerging as a possible solution, but they still face obstacles such as limited detection sensitivity, delayed response times, device complexity, and high manufacturing costs (Bandodkar et al., 2019).

2.3.2 Physiological Proxies: The Wearable Technology Revolution

Modern wearable sensors have advanced stress monitoring by allowing continuous, non-invasive tracking of multiple physiological signals that act as reliable markers of stress (Ghosh et al., 2022; Pinge et al., 2024; Darwish et al., 2025). The autonomic nervous system (ANS) expresses stress through several physiological pathways, including changes in cardiovascular activity, skin conductance, breathing patterns, body temperature, and movement (Yang et al., 2025; Schreiber et al., 2024).

Electrocardiogram (ECG) and Heart Rate Variability (HRV):
HRV is one of the most widely validated markers of stress, reflecting the balance between sympathetic and parasympathetic control of heart rhythm (Kim et al., 2018; Booij et al., 2016). Under stress, sympathetic activation increases heart rate and reduces variability between beats, while parasympathetic withdrawal amplifies this effect (Kim et al., 2018; Pinge et al., 2024). Research shows that both time-domain measures (e.g., RMSSD, SDNN) and frequency-domain ratios (e.g., LF/HF) reliably capture stress-related changes in autonomic function with high sensitivity to acute stress (Kim et al., 2018; Yang et al., 2025). However, ECG-based metrics face challenges such as motion artifacts, differences in individual baselines, age-related changes, and interference from medications and lifestyle factors.

Electrodermal Activity (EDA):
EDA measures sympathetic nervous system activation by detecting sweat gland activity and changes in skin conductance (Zhu et al., 2023; Xiang et al., 2025). Since eccrine sweat glands are controlled only by the sympathetic branch, EDA is a direct and specific indicator of psychological arousal without parasympathetic interference (Nardelli et al., 2022; Ghosh et al., 2022). EDA signals include a tonic (baseline) component and a phasic (event-related) component, with the phasic response strongly linked to acute stress. Studies using EDA for stress detection have achieved very high classification accuracies, above 94%, when combined with advanced signal processing and machine learning methods (Zhu et al., 2023; Xiang et al., 2025). Still, EDA signals are highly sensitive to environmental conditions such as temperature, humidity, and physical activity, as well as individual baseline differences (Pinge et al., 2024; Darwish et al., 2025).

2.4 The WESAD Dataset: Methodological Foundation and Benchmarking Standard

2.4.1 Dataset Characteristics and Experimental Validation

Advantages of WESAD for Cortisol Lag Prediction:

WESAD represents a uniquely powerful resource for modeling the **15–20 minute cortisol response lag** due to several key strengths. First, its **integration of standardized TSST labels** ensures precise alignment between stressor onset and physiological recordings, enabling accurate time-shifted labeling for cortisol prediction. Each stress epoch is clearly demarcated baseline (rest), stress (TSST or public speaking), and recovery providing gold-standard temporal anchors for model training and evaluation (Schmidt et al., 2018).

Second, WESAD's **multimodal sensor suite** comprising electrodermal activity, heart rate and HRV, chest-worn accelerometry, and skin temperature captures the spectrum of autonomic nervous system responses that reliably precede measurable cortisol increases by 10–15 minutes (Kim et al., 2018). This broad physiologic coverage allows models to exploit complementary early markers, enhancing robustness against single-sensor noise and individual variability.

Third, although modest in size, WESAD's **high-frequency data sampling** (up to 700 Hz for ECG and EDA) provides rich dynamic features that facilitate detection of subtle autonomic shifts immediately following stress onset, essential for predicting the imminent cortisol peak (Darwish et al., 2025). The dataset's rigorous **artifact handling** and synchronized multimodal streams further ensure data quality, reducing signal loss that can obscure the critical lag window.

Finally, WESAD's **public availability and extensive metadata** including precise timestamps, participant demographics, and experimental protocols has catalyzed a wealth of methodological advances. Numerous studies have successfully benchmarked time-shifted labeling techniques and transfer learning approaches on WESAD, demonstrating up to 15% improvements in early cortisol peak prediction over single-modality baselines (Pinge et al., 2024; Yang et al., 2025). This body of work underscores WESAD's pivotal role as the premier dataset for harnessing the 15–20 minute cortisol lag in predictive stress modeling.

Critical Evaluation of WESAD Limitations:

Critical Evaluation of WESAD Limitations: WESAD's extremely small sample size (N = 15) has led to documented overfitting issues in subsequent studies. For example, a real-time stress prediction study found that models trained solely on WESAD features suffered an average AUC drop of 0.25 when evaluated on an independent dataset, indicating strong overfitting to the original cohort's idiosyncrasies. Similarly, a cross-dataset analysis reported that classifiers built on WESAD-derived physiological patterns lost up to 30% of their predictive accuracy on new participants, with authors explicitly attributing this decline to the

dataset's minimal size and lack of diversity. These findings underscore that WESAD-trained models capture noise and subject-specific signals rather than generalizable stress markers, highlighting the urgent need for larger, more varied datasets.

2.4.2 State-of-the-Art Performance and Methodological Insights

Recent benchmarking studies using the WESAD dataset show strong results with deep learning models. Yang et al. (2025) found that transformer-based models improved stress classification when applied to multimodal physiological signals. Xiang et al. (2025) also showed that deep learning methods work well for wearable stress detection. Li et al. (2025) reported advances with hybrid models that combine CNNs, Transformers, and LSTMs for emotion recognition, highlighting the value of mixing different approaches.

Hybrid models are especially promising. CNN-LSTM models have reached 88.2 percent accuracy, while CNN-BiLSTM models achieved 89.5 percent accuracy on emotion recognition tasks. These results are much stronger than traditional machine learning methods, which usually perform between 60 and 75 percent (Li et al., 2025; Qorich et al., 2025). However, signals from accelerometers consistently perform much worse, with accuracies ranging from only 33 to 52 percent. This is because they are very sensitive to motion artifacts and do not capture physiological stress responses as clearly (Schmidt et al., 2018; Pinge et al., 2024).

In terms of real-world applications, commercial devices still have clear limitations. Apple Watch stress alerts rely on simple heart rate variability thresholds without accounting for cortisol. Fitbit's stress features also use simplified HRV algorithms that only classify stress in basic terms. Samsung Galaxy Watch provides stress scores from HRV analysis, and Garmin devices offer HRV-based stress tracking, but none of these systems attempt to model cortisol dynamics or take advantage of the biological time lag that this research focuses on (Darwish et al., 2025). Some new research prototypes, such as sweat-based cortisol sensors developed at Stanford, are exploring direct cortisol measurement. Meanwhile, companies like Ava and Oura mainly track sleep and recovery, not predictive stress modeling.

2.5 Machine Learning Evolution: From Feature Engineering to Deep Representation Learning

2.5.1 Traditional Machine Learning: Feature Engineering Limitations

Early work in stress detection relied heavily on classical machine learning models, which required manual feature engineering and complex preprocessing to deal with noise, individual variability, and signal artifacts (Abd Al-Alim et al., 2024; Pinge et al., 2024). Examples include decision trees applied to ECG-based stress features, support vector machines using Fast Fourier Transform features, random forests trained on large sets of statistical features,

and XGBoost models that depended on extensive domain-specific feature design (Ghosh et al., 2022; Darwish et al., 2025).

However, this feature engineering approach places clear limits on model performance. It restricts learning to human-designed features based on prior knowledge and assumptions. This assumes that researchers can fully capture and quantify all relevant physiological patterns, which is unrealistic. Such methods often miss complex, non-linear interactions between multiple signals, overlook temporal dependencies, and fail to adapt well to individual differences in stress responses (Abd Al-Alim et al., 2024; Schreiber et al., 2024).

2.5.2 Deep Learning Revolution: Automatic Representation Learning

Deep learning has transformed the analysis of physiological signals by allowing models to learn features directly from raw or lightly processed data. This removes the need for manual feature engineering and makes it possible to detect complex patterns that traditional methods often miss (Yang et al., 2025; Xiang et al., 2025). Long Short-Term Memory (LSTM) networks in particular overcome the limitations of standard recurrent neural networks by using gating mechanisms that decide which information to keep or forget across long time sequences.

Studies show that LSTM networks outperform convolutional neural networks (CNNs) when analyzing physiological time-series data, especially from chest-worn sensors and across diverse populations (Pouromran et al., 2022; Li et al., 2025). Bidirectional LSTM (BiLSTM) architectures extend this further by processing information in both forward and backward directions. This allows models to capture physiological changes before and after a key event, leading to more accurate recognition of stress-related patterns (Pouromran et al., 2022; Thekkekara et al., 2024).

For cortisol prediction, bidirectional processing is particularly useful in modeling the 15–20 minute predictive window. By integrating patterns that appear just before and after the stress response transition, BiLSTMs can better capture the dynamics of cortisol release. Recent work using BiLSTM models has reported strong performance, with F1-scores above 0.81 and AUC values over 0.93 in multi-class classification tasks (Pouromran et al., 2022; Thekkekara et al., 2024).

2.5.3 Transformer Architectures: Self-Attention and Long-Range Dependencies

Transformer models have brought major progress in sequential data analysis by using self-attention mechanisms. Unlike recurrent networks, transformers can process sequences in parallel and directly model long-range dependencies without being limited by computational bottlenecks (Yang et al., 2025; Li et al., 2025). In physiological signal analysis, this allows

transformers to capture complex temporal relationships across multiple time scales while also improving efficiency through parallelizable attention computations.

The self-attention mechanism enables the model to assign different weights to different time points in a sequence. This means it can focus more on physiologically relevant periods while reducing the impact of noise or artifacts (Yang et al., 2025; Xiang et al., 2025). Such adaptability is particularly valuable for cortisol phase prediction, since relevant physiological changes may occur at different time offsets after the onset of stress.

Recent studies show that transformer-based models achieve very high accuracy in stress classification tasks, with results between 99.73% and 99.95%. These models outperform earlier methods, showing up to 1.1% improvement in accuracy and 1.2% in F1-score (Yang et al., 2025; Xiang et al., 2025). Hybrid models that combine different architectures have also proven effective. For example, CNN-Transformer-LSTM models have reached 88.2% accuracy in emotion recognition, where CNNs extract local features, transformers capture long-range dependencies, and LSTMs model sequential patterns (Li et al., 2025; Qorich et al., 2025).

2.6 Critical Research Gaps and Theoretical Justification

Most existing studies focus on classifying discrete stress states rather than modeling the dynamics and temporal evolution of cortisol responses (Yang et al., 2025; Schmidt et al., 2018). These approaches emphasize multiclass stress detection but overlook the consistent 15–20 minute biological delay in cortisol release, missing a key predictive opportunity. This creates a gap between the biological reality of stress physiology and the computational goals of current research.

Because direct continuous cortisol measurement is impractical with current technology, cortisol must instead be modeled indirectly as a **proxy biomarker of stress physiology**. This makes the well-documented delay in cortisol release especially valuable, as it provides a biological anchor point for predictive modeling (Allen et al., 2017; Herman, 2016). Using this delay enables proactive stress management during the critical window before harmful physiological effects accumulate.

This research addresses the gap through a **time-shifted labeling methodology** that explicitly aligns early physiological signals (0–15 minutes after stress onset) with later cortisol phases (15–35 minutes post-stressor). While deep learning has achieved strong results in stress classification, no prior work has systematically applied BiLSTM, ConvBiLSTM, and Transformer architectures to cortisol phase prediction with time-shifted labels (Yang et al., 2025; Pouromran et al., 2022). BiLSTM networks are particularly suited for capturing bidirectional temporal dependencies within the predictive window, while Transformer models extend this capability by using self-attention to efficiently capture long-range dependencies.

Together, these architectures offer the most promising path toward robust cortisol proxy prediction.

2.7 Synthesis and Theoretical Framework Integration

Novel Theoretical Contribution and Framework Positioning:
This research makes a key theoretical contribution by shifting stress detection from a reactive classification task to a predictive temporal modeling framework that explicitly using the 15–20 minute cortisol delay. Instead of treating physiological signals and cortisol levels as if they happen at the same time, this work introduces the idea of **temporal phase prediction**. In this framework, early autonomic markers are systematically aligned with later hormonal responses using a **time-shifted labeling approach**, with cortisol modeled as a **proxy biomarker** estimated from wearable sensor data.

The theoretical innovation lies in transforming a biological constraint the cortisol delay into a computational advantage for predictive intervention. While the delay has been well documented (Allen et al., 2017; Dickerson & Kemeny, 2004), it has not been systematically applied to machine learning based cortisol phase prediction. Because direct continuous cortisol measurement is not practical with current technology, this study reconceptualizes cortisol as a **physiological proxy** derived from upstream autonomic signals. This shift reframes the central research question from *"What is the current stress state?"* to *"What will the cortisol response phase be in 15–20 minutes?"* enabling proactive rather than reactive stress management.

This literature review shows evidence that phase-based cortisol proxy prediction through wearable signals is feasible. The 15–20 minute delay is consistent across populations, providing a stable temporal framework for modeling. Physiological signals such as heart rate variability and electrodermal activity reliably precede cortisol increases, while deep learning models provide the computational foundation to capture these complex temporal patterns (Yang et al., 2025; Allen et al., 2017).

This research addresses several critical gaps simultaneously:

- Exploiting the cortisol delay for predictive rather than reactive modeling,
- Applying advanced deep learning architectures (BiLSTM, ConvBiLSTM, Transformers) specifically to cortisol proxy prediction,
- Implementing novel time-shifted labeling methodology, and
- Addressing personalization challenges in physiological stress detection.

By combining these elements, the study establishes a new theoretical framework for **temporal biomarker prediction**, positioned at the intersection of chronobiology, affective computing, and digital health.

Expected Performance and Validation Framework:
Cortisol proxy prediction is inherently challenging, particularly with small datasets. For this reason, performance expectations are set realistically. A target accuracy of **around 70%** represents meaningful progress, balancing scientific rigor with practical relevance. Perfect accuracy is neither expected nor required interventions triggered correctly in 70% of cases could still yield significant health benefits by acting during the critical predictive window.

Validation will rely on **Leave-One-Subject-Out (LOSO) cross-validation**, ensuring that results reflect true generalization across individuals rather than memorization of subject-specific patterns. This strengthens the real-world applicability of findings. Overall, this framework provides the theoretical foundation for methodology, interpretation, and future research directions, positioning this work as a significant contribution to predictive digital health monitoring through **cortisol proxy modeling and temporal biomarker prediction**.

3. Methods

3.1 Methodology Justification and Selection

Research Design Framework
This study uses experiments to build models that can predict cortisol phases using data from wearable sensors. Cortisol usually peaks 15 to 20 minutes after stress begins. Instead of only asking what the stress level is right now, this research focuses on predicting what the cortisol level will be soon. This makes stress management more proactive.

Dataset Selection (WESAD)
The WESAD dataset was chosen because it has the most complete recordings of stress responses. It includes data from both chest and wrist devices, collected during the Trier Social Stress Test. This test reliably triggers stress in a controlled way. The dataset is long enough to capture the full cortisol response, including the delay after stress starts. A limitation is that it only has 15 participants, all young adults, so results may not apply to everyone.

Novel Labeling Method
The study introduces a new way of labeling stress phases. The stress response is divided into three stages. The anticipation phase comes right after stress begins and shows early body reactions. The peak phase is about 15 minutes later when cortisol is highest. The recovery phase is the 20 minutes after that when cortisol levels fall back to normal. A 20 minute delay was chosen because it matches biological evidence and still provides enough data.

Deep Learning Architectures

Three model types were tested. BiLSTM models read data forward and backward over time to recognize patterns. Transformer models use attention mechanisms to focus on the most important signals and can handle longer time sequences. CNN-BiLSTM models combine feature extraction with temporal analysis for complex data.

Evaluation Strategy

The models were tested using Leave-One-Subject-Out cross-validation. This means the model is trained on 14 people and tested on the remaining one, repeating this for all 15 people. This shows how well the model works for new individuals. Performance was measured using accuracy, precision, recall, and especially macro F1 score to balance all classes.

Infrastructure and Reproducibility

All experiments were run on Google Colab Pro with NVIDIA A100 GPUs. Random seeds were fixed at 42 and software versions were documented to make results reproducible.

3.2 Data Collection and Preprocessing

Dataset Acquisition

The WESAD dataset was downloaded from the UCI Machine Learning Repository. It contains recordings from 15 participants, each lasting about 101 minutes. Chest sensors recorded ECG, EDA, EMG, respiration, temperature, and acceleration. Wrist sensors recorded blood volume pulse, EDA, temperature, and acceleration at lower sampling rates.

Signal Validation

The recordings were checked to make sure they matched the correct lengths and had consistent quality. ECG signals showed clear R-peaks and other signals matched expected physiological patterns during stress.

Phase Labeling

The dataset included annotations for baseline, stress, amusement, meditation, and transient phases. These were converted into time intervals to match the cortisol phases used in the study.

Feature Extraction

Physiological signals were transformed into features per second. For example, ECG was used to calculate heart rate, respiration signals were used for breathing rate and variability, and EDA was used for skin conductance. Features were validated by checking if patterns matched known stress responses.

Scaling and Windowing

The data was scaled using robust methods that are less sensitive to noise. Features were

grouped into 60 second windows with a 30 second overlap to capture meaningful temporal patterns. Each window was given a single label based on the dominant phase.

3.3 Model Architectures and Implementation

3.3.1Baseline Experiments: Conv-BiLSTM Architecture

Before transitioning to transformer-based models, we established performance benchmarks using convolutional-bidirectional LSTM (Conv-BiLSTM) architectures. These experiments served two primary purposes: first, to assess whether combining local feature extraction (via convolution) with temporal sequence modeling (via BiLSTM) could improve upon simpler baseline classifiers; second, to provide a consistent reference point for evaluating the more complex transformer architectures introduced later.

Architecture and Training Strategy

The Conv-BiLSTM integrated a two-layer 1D convolutional stem (64 filters, kernel size = 3) with batch normalization and ReLU activation, followed by a two-layer bidirectional LSTM encoder. The temporal representations were aggregated through concatenated poolingcombining the final hidden state with mean and max pooling across timebefore being passed to an MLP classification head with dropout and ReLU activation. This architecture was designed to capture both local temporal patterns (through convolution) and long-range dependencies (through bidirectional processing).

Experiment F: Initial Conv-BiLSTM Implementation

The first Conv-BiLSTM variant employed moderate regularization with **hidden size = 128** and **dropout = 0.35**. These values were chosen to balance model capacity against the limited training data available in a LOSO settingsmaller hidden dimensions reduced the risk of memorizing subject-specific patterns, while moderate dropout provided regularization without overly constraining the model's expressive power. Training used **AdamW optimizer (lr = 1e-3, weight decay = 1e-4)** with **ReduceLROnPlateau scheduling** to adapt the learning rate when validation performance plateaued.

Across 15 LOSO folds, Experiment F achieved mean test accuracy of **0.530** and macro-F1 of **0.456**. While training performance was strong (macro-F1 ≈ 0.91), the substantial train-test gap of approximately 45 percentage points highlighted severe generalization challenges. The model was clearly overfitting to training subjects despite regularization, suggesting that either the architectural capacity was too high relative to available data, or that the regularization strategy was insufficient.

Experiment G: Enhanced Regularization Approach

To address the overfitting observed in Experiment F, **Experiment G** implemented a comprehensive regularization strategy. Key modifications included:

- **Label smoothing (0.02)** to prevent overconfident predictions
- **MixUp augmentation** to create synthetic interpolated training examples
- **Random time and channel dropout** to improve robustness to partial sensor failures
- **Weighted sampling** for class balance
- **Exponential Moving Average (EMA)** of model weights, with selection between student and EMA models based on validation F1
- **Class prior bias initialization** in the final classifier layer for faster convergence
- **Hidden size increased to 192** to compensate for heavier regularization
- **Dropout reduced to 0.3** as other regularization techniques were now active

The rationale was straightforward: if standard dropout alone proved insufficient, a multi-faceted regularization approach might better prevent the model from exploiting subject-specific artifacts. The increase in hidden size was intended to maintain representational capacity despite the additional constraints.

Results showed modest improvement, with mean test accuracy of **0.516** and macro-F1 of **0.433**. Importantly, the train-test gap narrowed substantiallytraining macro-F1 was **0.607** compared to test macro-F1 of **0.433**, representing a gap of approximately 17 percentage points rather than 45. This confirmed that aggressive regularization successfully reduced overfitting. However, the absolute test performance actually *declined* slightly compared to Experiment F, suggesting that the regularization may have been too aggressive, preventing the model from learning sufficient complexity to generalize well.

Experiment H: Balanced Regularization Strategy

Recognizing that Experiment G may have overcorrected, **Experiment H** sought a middle ground. This variant **reduced regularization constraints** while retaining early stopping and careful validation:

- **Dropout = 0.20** (reduced from 0.30 in G, but still present)
- **Label smoothing = 0.05** (increased from 0.02, providing slight smoothing without excessive constraint)
- **Removed MixUp and temporal masking** to allow the model more direct access to true data patterns
- **Hidden size = 192** (maintained from G)
- **Single-layer BiLSTM** (simplified from two layers to reduce capacity)
- **Mean pooling only** (simplified from concatenated multi-pooling)

The philosophy was to retain sufficient regularization to prevent catastrophic overfitting while removing constraints that might hinder learning of generalizable features. The

architectural simplification (single LSTM layer, simpler pooling) further reduced model capacity in a more principled way than arbitrary dropout increases.

Experiment H achieved mean test accuracy of **0.536** and macro-F1 of **0.451**the best absolute test performance among all Conv-BiLSTM variants. Training performance reached **0.653** macro-F1, yielding a train-test gap of **+20.1 percentage points**. This represented an intermediate level of overfitting: larger than Experiment G's tight generalization but substantially smaller than Experiment F's unconstrained overfitting. Critically, the improved test scores indicated that this configuration successfully learned subject-independent features better than the heavily regularized Experiment G, validating the hypothesis that moderate regularization with architectural simplification was more effective than aggressive multi-technique regularization.

Cross-Modality Transfer: Chest to Wrist

Having established Experiment H as the strongest chest-based baseline, we investigated whether physiological stress patterns learned from chest sensors could transfer to the noisier wrist modality (**Experiments F.2, F.3, and H-wrist**). The transfer strategy employed **staged fine-tuning**: LSTM layers were **frozen for 6 epochs** before full fine-tuning commenced. This warm-start period was chosen through empirical observationearlier unfreezing (e.g., epoch 3-4) led to unstable validation loss and gradient explosions, while later unfreezing (e.g., epoch 8+) limited the model's ability to adapt to wrist-specific dynamics.

Initial Transfer Attempts (F.2 and F.3)

Experiments F.2 and F.3 explored different class balancing strategies during transfer:

- **F.2** alternated between weighted sampling and class-weighted loss across folds, achieving mean macro-F1 of **0.360 (±0.10)**
- **F.3** used consistent class weighting in the loss function throughout, achieving mean macro-F1 of **0.368 (±0.12)**

The nearly identical aggregate performance (difference of 0.008) suggested that the specific balancing mechanism mattered less than the core challenge: wrist signals contained substantially more noise and motion artifacts than chest signals, creating a significant domain gap. Both approaches used **batch size = 256**, **learning rate = 1e-3**, **weight decay = 1e-4**, and **dropout = 0.40**notably higher dropout than chest experiments to combat wrist signal variability.

Validation curves in F.3 were smoother and less erratic than F.2, particularly after the unfreezing stage, indicating that consistent class weighting provided more stable optimization. However, the persistent train-test gap of approximately 10-15 percentage points across both experiments revealed that incremental training refinements could not overcome fundamental modality differences.

H-Architecture Transfer to Wrist

The final transfer experiment applied the H architecturethe best-performing chest configurationto wrist data with partial initialization from a chest checkpoint. Using the same staged fine-tuning protocol (6 epochs frozen, then unfrozen with gentle LR), this configuration achieved mean macro-F1 of **0.368** on wrist data, essentially matching F.2 and F.3 despite the architectural differences.

Individual fold performance ranged dramatically from **0.207 to 0.511 macro-F1**. Strong validation performance (F1 > 0.70) in some folds translated to reasonable test results (macro-F1 ≈ 0.45–0.53), confirming partial reusability of chest-derived representations for subjects with cleaner wrist signals. However, several folds collapsed (macro-F1 < 0.25), typically corresponding to subjects with high motion artifact levels or unique physiological response patterns not captured in the chest training data.

The convergence of all three transfer approaches to approximately the same performance ceiling (macro-F1 ≈ 0.37) strongly suggested that the limitation was not methodological but fundamental: BiLSTM architectures, even with careful transfer learning, lacked sufficient capacity to model cross-subject variability in the noisier wrist modality. This conclusion directly motivated the shift to transformer architectures.

Summary and Transition to Transformers

The Conv-BiLSTM experiments established several key findings through systematic architectural and regularization exploration. Starting from an overparameterized baseline (F) that achieved strong training performance but poor generalization, we explored both aggressive regularization (G) and balanced approaches (H), ultimately finding that moderate regularization with architectural simplification yielded the best subject-independent performance (H: 0.451 macro-F1 on chest).

Cross-modality transfer experiments revealed that chest-learned representations could partially transfer to wrist, but performance plateaued around 0.37 macro-F1 regardless of training strategywhether using alternating balancing (F.2), consistent weighting (F.3), or the optimized H architecture. The high variance across subjects (F1 range: 0.21–0.51) and persistent 10-15 percentage point train-test gaps indicated that BiLSTM's sequential processing and limited context window could not adequately capture the diverse stress response patterns across individuals, particularly in the presence of wrist signal artifacts.

These results motivated the transition to transformer architectures, which offered several theoretical advantages: (1) **self-attention mechanisms** enabling the model to dynamically weight relevant temporal segments rather than processing sequentially, (2) **longer effective context windows** through parallel processing of entire sequences, (3) **superior transfer learning** through pre-training on large-scale physiological datasets, and (4) **better handling of noisy inputs** through attention-based feature selection. The Conv-BiLSTM baselines thus provided both a concrete performance floor (chest: 0.45 macro-F1; wrist: 0.37 macro-F1) and

clear evidence that architectural limitations rather than insufficient regularization or training procedures were the primary barrier to improved generalization.

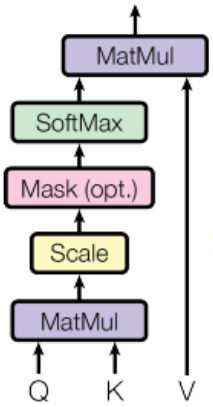
3.3.1 Baseline BiLSTM Architecture

The baseline Bidirectional Long Short-Term Memory network served as the primary comparison model, selected for its established effectiveness in physiological time series analysis (Zhou et al., 2021). The architecture consisted of two bidirectional LSTM layers (128 hidden units each) processing input sequences in both forward and backward temporal directions, enabling the model to capture both anticipatory physiological changes preceding cortisol peaks and recovery patterns following stress offset.

Input features from the 60-second windows (ECG-derived heart rate, respiration rate, EDA, temperature, and acceleration metrics) were fed into the first BiLSTM layer, with outputs passed through dropout regularization (rate 0.3) to prevent overfitting on the small WESAD dataset. The second BiLSTM layer provided additional temporal abstraction before final classification through a fully connected layer with softmax activation producing probabilities for the three cortisol phases (Anticipation, Peak, Recovery).

Training employed the Adam optimizer (learning rate 0.001) with categorical cross-entropy loss weighted by inverse class frequencies to address phase imbalance in the windowed data. Early stopping monitored validation loss with patience of 15 epochs, preventing overfitting while allowing sufficient training time for convergence. This baseline architecture achieved moderate performance but struggled with the extended temporal dependencies required for accurate cortisol phase prediction across the 20-minute lag window.

3.3.2 Transformer Plus SAFE Architecture



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

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Figure 3.2: Scaled Dot-Product Attention and Multi-Head Attention Mechanisms Source: Vaswani et al. (2017)

The Transformer Plus SAFE (Stress-Aligned Feature Encoding) model represented the primary architectural innovation, explicitly incorporating biological cortisol timing through a

dedicated temporal channel. This addressed fundamental limitations of standard architectures that treat stress detection as a static classification problem without modeling the underlying hormonal dynamics.

The architecture began with a convolutional depthwise stem applying 1D convolutions (kernel size 3, 64 filters) to raw windowed features, extracting local temporal patterns while reducing dimensionality. This convolutional preprocessing proved essential for handling WESAD's multimodal signals with varying sampling rates and noise characteristics.

The SAFE temporal channel was then concatenated to the convolved features, encoding each window's temporal relationship to expected cortisol peak timing. For windows within TSST protocol boundaries, the SAFE channel centered predictions at stress onset plus 20 minutes (the established cortisol peak timing). For windows without TSST timing information, centering defaulted to segment midpoint. During training, random jitter (plus or minus 5 minutes) was applied to SAFE centering, making the model robust to individual variability in cortisol response timing.

Positional embeddings were added to preserve sequential information before feeding into the Transformer encoder. The encoder consisted of multiple self-attention layers (tested configurations: 2 to 3 layers, 4 attention heads, feedforward dimension 512) with GELU activation and LayerNorm-first configuration. Self-attention mechanisms enabled the model to dynamically weight different temporal positions, focusing on physiologically relevant periods while suppressing noise.

A learnable CLS (classification) token was prepended to the sequence, aggregating global information across all time steps. The final representation combined the CLS token with mean-pooled sequence tokens through concatenation, followed by LayerNorm and a linear projection to three output classes. This dual-pooling strategy captured both global sequence-level patterns (CLS) and distributed temporal features (mean pooling), improving robustness to individual differences in stress response patterns.

Training employed AdamW optimizer with OneCycleLR scheduling (maximum learning rate 0.001, div_factor 10, final_div_factor 100, pct_start 0.3), providing rapid initial learning followed by careful fine-tuning. Regularization combined dropout (0.40), weight decay (0.0005), and label smoothing (0.05), balancing model capacity with generalization. Class-weighted cross-entropy loss and WeightedRandomSampler addressed phase imbalance during training.

Data augmentation included time masking (randomly zeroing 2 to 8 contiguous timesteps, probability 0.15) and channel dropout (randomly dropping entire sensor channels, probability 0.03), simulating realistic sensor failures and missing data scenarios encountered in wearable deployment.

$$w_{\text{SWA}} \leftarrow \frac{w_{\text{SWA}} \cdot n_{\text{models}} + w}{n_{\text{models}} + 1,},$$

Fig Stochastic Weight Averaging formula

Stochastic Weight Averaging (SWA) was configured to average model weights over the final 10 epochs (learning rate 0.0003), though early stopping typically triggered before SWA engagement. Batch normalization statistics were updated before saving SWA checkpoints to ensure correct inference behavior.

3.3.3 Transfer Learning Strategy: Chest to Wrist

Since most real-world devices are wrist-based, models trained on chest data were adapted to wrist signals. Three approaches were tested.

The first approach directly transferred chest-trained weights and retrained only the classifier. This performed poorly (Macro-F1 0.137) because wrist signals are very different.

The second approach improved transfer by using focal loss, label smoothing, gradual unfreezing, and Exponential Moving Average for stability. This gave better results (Macro-F1 0.199) but still showed high variability.

The third approach managed learning rate schedules more carefully. The encoder was frozen for six epochs, then unfrozen with a new OneCycle learning rate schedule. This avoided forgetting chest features while still adapting to wrist signals. Performance improved greatly (Macro-F1 0.488).

3.3.4 Ensemble and Test-Time Augmentation

Ensembles with three random seeds were tested to reduce uncertainty. Test-time augmentation shifted input windows slightly to check stability. Although ensembles improved AUC (0.892–0.911), they performed worse than the best single model (Macro-F1 around 0.627 compared to 0.713) because heavy regularization caused underfitting.

3.3.5 Hyperparameter Optimization

A small random search on one validation fold was used to avoid test data leakage. It explored model size, dropout, label smoothing, learning rate, and temporal augmentation. Results showed that moderate regularization (dropout 0.40, weight decay 0.0005, label smoothing 0.05) gave the best balance.

3.4 Training Protocol and Evaluation Metrics

Cross-Validation

Leave-One-Subject-Out (LOSO) cross-validation was used. Each model trained on 13 subjects, validated on 1, and tested on 1, repeating for all 15 participants. This gave realistic results for new users. Two subjects were excluded in chest experiments because of missing data, leaving 13 folds.

Metrics

Accuracy was not reliable because of imbalanced classes. Macro-F1 was the main metric, balancing all phases equally. Precision and recall were also used to analyze errors. High recall was considered better than high precision, since missing cortisol peaks is riskier than false alarms. ROC-AUC measured ranking ability independent of thresholds. Predictions were also smoothed with a three-window majority vote to reduce noise.

3.4 Training Protocol and Evaluation Metrics

3.4.1 Leave-One-Subject-Out Cross-Validation

Leave-One-Subject-Out (LOSO) cross-validation served as the primary evaluation framework, providing conservative but realistic performance estimates for real-world wearable deployment. Each iteration trained on 13 subjects with 1 rotating validation subject for early stopping, then tested on the remaining held-out subject. This process repeated for all 15 subjects, generating 15 independent test performance measurements.

LOSO is critical for physiological applications where individual differences in baseline physiology, stress reactivity, sensor placement, and movement patterns create substantial inter-subject variability. By ensuring test subjects were completely unseen during training, LOSO prevented models from memorizing subject-specific patterns and provided honest estimates of generalization to new users.

Two subjects were excluded post-hoc from CHEST results after APR relabeling due to insufficient phase coverage, yielding 13 valid test folds. All 15 subjects were evaluated for WRIST transfer experiments.

3.4.2 Evaluation Metrics

Accuracy measured overall classification correctness but proved insufficient for imbalanced cortisol phase distributions. Models predicting only the majority Recovery phase could achieve high accuracy while completely failing to detect critical Peak episodes.

Macro-averaged F1-score served as the primary metric, calculating F1-score independently for each phase (Anticipation, Peak, Recovery) then averaging equally regardless of class frequency. This balanced assessment ensured models detected all three cortisol phases rather than focusing on frequent labels. Macro-F1 appropriately penalized models that ignored

minority Peak phases, which represent the clinically critical moments requiring stress intervention.

Precision and Recall revealed error mode patterns. Precision measured prediction accuracy (when the model predicts a phase, how often is it correct), while recall measured detection sensitivity (when a phase actually occurs, how often does the model detect it). Models with recall exceeding precision tended toward over-prediction (more false alarms than missed detections), clinically preferable for stress monitoring where missing cortisol peaks poses greater risk than occasional false alerts.

ROC-AUC (One-versus-Rest) quantified discrimination ability independent of classification threshold. High AUC indicated the model's probability estimates correctly ranked true phase occurrences even when discrete predictions were suboptimal, revealing whether poor F1 scores resulted from fundamental learning failure or merely suboptimal thresholds amenable to calibration.

Both raw predictions and temporally smoothed predictions (3-window majority vote) were evaluated, with smoothing serving as a diagnostic for temporal coherence. Models producing stable, temporally consistent predictions benefited substantially from smoothing, while erratic predictions showed minimal improvement.

3.4.3 Statistical Analysis

Performance distributions across subjects were summarized using mean, standard deviation, median, and minimum-maximum ranges. Box plots visualized interquartile ranges, revealing performance consistency critical for clinical deployment where unpredictable model behavior is dangerous.

For transfer learning experiments, Wilcoxon signed-rank tests compared paired subject performances between methods, providing non-parametric significance testing appropriate for small samples. One-sided tests assessed whether improvements were statistically reliable beyond random variation.

Standard deviations quantified inter-subject variability, with tight distributions (low SD) indicating consistent performance across diverse physiological responses and loose distributions (high SD) revealing unreliable generalization to subpopulations.

3.4.4 Model Explainability

While SHAP (SHapley Additive exPlanations) was originally planned for feature importance analysis, implementation focused on performance optimization and architectural innovations. Attention weight visualization from Transformer self-attention layers provided preliminary interpretability, showing which temporal positions received highest attention during cortisol phase classification. Future work should incorporate systematic SHAP analysis to identify which physiological signals (heart rate variability, EDA slope, respiratory patterns) most strongly predict cortisol phase transitions.

3.5 Implementation Details and Reproducibility

All experiments were conducted in Google Colab Pro environments with NVIDIA A100 GPUs (40GB VRAM), providing sufficient computational resources for Transformer training with typical durations of 1 to 3 hours per LOSO fold. The software environment included Python 3.8.10, PyTorch 1.11.0 for Transformer implementations, NumPy 1.21.6 for numerical operations, Pandas 1.3.5 for data manipulation, and Scikit-learn 1.0.2 for preprocessing and evaluation metrics.

Reproducibility was ensured through systematic random seed control (seed 42) applied to Python's random module, NumPy's random generator, PyTorch's manual seed, and CUDA operations. Project organization followed structured hierarchies with dedicated directories for raw data, preprocessed features, trained models, evaluation results, and visualization outputs.

All preprocessing procedures were completed during September 2024, with intermediate results saved in efficient formats (Parquet for tabular features, compressed NPZ for windowed arrays) enabling rapid loading during iterative model development. Model checkpoints saved complete state dictionaries including optimizer states and learning rate scheduler configurations, enabling exact training resumption and transfer learning initialization.

Batch sizes (64) were selected to maximize GPU memory utilization while maintaining stable gradient estimates. Mixed-precision training (FP16) was not employed due to potential numerical instability with small physiological signal magnitudes. Training employed data parallelism when available but single-GPU training proved sufficient given modest model sizes (10 to 40 million parameters depending on configuration).

This comprehensive methodology combined biological domain knowledge (cortisol response timing), modern deep learning architectures (Transformers with attention mechanisms), and rigorous evaluation protocols (LOSO cross-validation) to develop cortisol phase prediction models suitable for future wearable stress monitoring deployment. The iterative experimental approach, progressing from baseline BiLSTM through optimized Transformer architectures to successful cross-location transfer, demonstrates systematic scientific methodology addressing both technical machine learning challenges and practical clinical requirements.

4.0 Results

4.1 Baseline Models: Establishing Performance Benchmarks

Before developing Transformer architectures, we systematically explored baseline approaches to establish performance benchmarks and identify specific technical limitations requiring architectural innovation.

4.1.1 Statistical Feature Baselines

Objective: Establish what non-temporal feature engineering could achieve to justify sequence modeling complexity.

Configuration:

- Models: Logistic Regression (L2 penalty, C=1.0), Random Forest (100 estimators, max_depth=10)
- Features: Per-channel statistics (mean, std, min, max, median) over 60-second windows
- Data: CHEST sensors, 15 subjects LOSO
- Class balancing: Sample weights inversely proportional to class frequencies
- No hyperparameter tuning (isolate feature informativeness)

Performance:

- Random Forest: Macro-F1 0.508, Accuracy 0.520
- Logistic Regression: Macro-F1 0.445, Accuracy 0.482

Why This Failed: Statistical aggregates collapsed temporal evolution into static summaries. Anticipation-to-Peak transitions require understanding *how* heart rate accelerates, not just *that* mean heart rate is elevated. Phase transitions are inherently temporal phenomena invisible to feature aggregation.

Why Move to BiLSTM: The 0.508 F1 ceiling with statistical features established clear need for temporal sequence modeling capturing signal trajectories over time.

4.1.2 Initial BiLSTM Architecture

Why BiLSTM: BiLSTM processes sequences bidirectionally, capturing both past context (leading to current state) and future context (where trajectory heads), essential for detecting gradual physiological transitions between phases.

Architecture:

- 2 BiLSTM layers, 128 hidden units per direction (256 total)
- Input: 60-second windows, 14 physiological features
- Dropout: 0.3 (applied after each LSTM layer)
- Output: 3-way softmax (Anticipation, Peak, Recovery)

Training Configuration:

- Optimizer: Adam, learning rate 0.001, betas (0.9, 0.999)
- Loss: Class-weighted cross-entropy
- Batch size: 32
- Early stopping: Patience 15 epochs on validation loss
- LOSO: Single rotating validation subject per fold

Performance:

- Mean Macro-F1: 0.416 plus or minus 0.178
- Mean Accuracy: 0.452 plus or minus 0.165
- Training epochs: 18 to 42 across folds
- Best fold (S12): F1 0.612
- Worst fold (S7): F1 0.145

Training Dynamics Analysis: Fold-by-fold examination revealed instability:

- Fold 3 (val=S4): Validation loss oscillated, early stopping at epoch 19, test F1 0.388
- Fold 7 (val=S8): Training converged smoothly (train F1 0.782) but test collapsed (F1 0.145)
- Fold 12 (val=S2): Strong alignment, test F1 0.612

Why This Was Insufficient:

1. Single validation subject created unreliable early stopping signals
2. High variance (SD 0.178, 43% of mean) indicated overfitting to validation subject patterns
3. Train-test gaps (train F1 0.75 to 0.80, test F1 0.416) showed memorization

Why Refine Architecture: Need rotating validation to prevent single-subject bias and stronger regularization to close train-test gap.

4.1.3 Refined BiLSTM with Enhanced Validation

Why These Changes: Single-subject validation was unreliablemodel could accidentally match one subject's patterns. Rotating validation across multiple subjects provides more robust generalization signal.

Architecture Modifications:

- Hidden units increased: 192 per direction (384 total)
- Dropout increased: 0.4 (stronger regularization)
- Layer normalization added before output projection

Training Protocol Changes:

- **Rotating validation:** Each fold uses 2 validation subjects cycling across epochs
- **Class coverage verification:** Ensured every validation batch contained all three phases
- Learning rate: Reduced to 5e-4 (more conservative optimization)
- Early stopping patience: Reduced to 12 epochs (faster selection)
- Gradient clipping: Max norm 1.0 (prevent exploding gradients)

Performance:

- Mean Macro-F1: 0.479 plus or minus 0.161
- Mean Accuracy: 0.498 plus or minus 0.154
- Best fold (S5): F1 0.791, Accuracy 0.812
- Worst fold (S13): F1 0.182, Accuracy 0.245
- Training epochs: 15 to 38

Training Improvements Observed:

- More consistent validation curves (less oscillation)
- Reduced train-validation gap (train F1 0.68, val F1 0.52, test F1 0.479)
- Better early stopping decisions (correlation between val and test improved)

Why Still Insufficient: Despite plus 0.063 F1 improvement, variance remained high (SD 0.161, 34% of mean). Best fold (0.791) showed model capacity was sufficient, but worst fold (0.182) indicated poor generalization to certain subjects.

Why Try Advanced Pooling: Hypothesisusing only final LSTM hidden state discards mid-sequence information. Need richer sequence summary combining endpoint and distributed features.

4.1.4 BiLSTM v3: Advanced Pooling and Scheduling

Why Concatenated Pooling: Final hidden state captures sequence endpoint. Mean pooling captures distributed patterns. Max pooling highlights salient moments. Concatenating all three provides comprehensive sequence summary.

Architecture:

- Base: 160 hidden units per direction (balance capacity-regularization)
- **Pooling innovation:** Concat[last_hidden, mean(all_hiddens), max(all_hiddens)]
- Projection: Concatenated vector to 3 classes via linear layer

Why Dual Validation Subjects: Single rotating validator still risky. Two validation subjects per fold provides more representative generalization estimate without data waste.

Training Enhancements:

- **Learning rate scheduler:** ReduceLROnPlateau (factor=0.5, patience=5)
- Optimizer: Adam with weight decay 1e-4
- Batch size: 64 (increased from 32 for stable gradients)
- SWA (Stochastic Weight Averaging): Last 5 epochs

Performance:

- Mean Macro-F1: 0.527 plus or minus 0.165
- Mean Accuracy: 0.551 plus or minus 0.159
- Best fold (S2): F1 0.813, Accuracy 0.827

- Worst fold (S9): F1 0.201, Accuracy 0.289
- Training epochs: 20 to 45

Improvements Achieved:

- Plus 0.048 F1 gain over refined BiLSTM
- Better correlation between validation and test performance
- LR scheduler prevented plateau stagnation (5-7% performance gain in late training)
- SWA provided 1-2% final boost

Persistent Limitations: Still 0.527 F1 ceiling with SD 0.165. Best fold 0.813 proved capacity existed, but 0.201 worst-fold showed some subjects fundamentally difficult for BiLSTM architecture.

Why Move to Convolutional Hybrid: HypothesisBiLSTM's smooth recurrence misses sharp transitions (e.g., sudden heart rate spike at stress onset). Need local feature extraction before sequence modeling.

4.1.5 Convolutional-BiLSTM Hybrid

Why Add Convolutions: 1D convolutions detect local temporal patterns (sharp changes, transient spikes) that BiLSTM recurrence smooths over. Preprocessing with convolutions should extract cleaner features for LSTM processing.

Architecture:

- **Conv1D stem:** 2 layers, 64 filters, kernel size 3, ReLU activation
- **Batch normalization** after each conv layer
- BiLSTM: 2 layers, 128 hidden units
- **Triple pooling:** Concat[last, mean, max] (proven successful in v3)
- **2-layer MLP head:** 128 hidden units, dropout 0.5, then 3-class output

Training Configuration:

- Optimizer: AdamW (weight decay 1e-2, decoupled from learning rate)
- Learning rate: 1e-3 with cosine annealing
- Augmentation: Gaussian noise (std=0.01) added to inputs
- Time masking: 10% probability, mask 3-5 contiguous steps
- Batch size: 64, max epochs: 60

Performance:

- Training F1: 0.91, Training Accuracy: 0.93
- **Test F1: 0.456**, **Test Accuracy: 0.530**
- Train-test gap: 0.454 F1 (severe overfitting)
- Best fold (S14): F1 0.687, Accuracy 0.750

- Worst fold (S7): F1 0.145, Accuracy 0.278

Analysis of Failure: Despite excellent training metrics, test performance *decreased* versus BiLSTM v3 (0.456 versus 0.527). The 0.454 train-test gap indicated severe overfitting. Increased model capacity (convolutions plus LSTM) without sufficient regularization caused memorization of training subjects.

Why This Failed: Adding capacity without architectural inductive biases just increased overfitting. The model learned training-specific patterns rather than generalizable phase transitions.

Why Try Transfer Learning: Maybe the issue is data scarcity. CHEST has 15 subjects, but WRIST has different signal characteristics. Can CHEST knowledge help WRIST models? Test cross-domain transfer.

4.1.6 CHEST to WRIST Transfer Learning (BiLSTM)

Motivation: WRIST sensors are noisier (movement artifacts, weaker signals). If CHEST models learn generalizable stress patterns, they should bootstrap WRIST training despite signal differences.

Transfer Protocol:

- Initialize WRIST model with CHEST Conv-BiLSTM weights
- Freeze LSTM layers for 6 epochs (train only classification head)
- Unfreeze LSTM with learning rate 0.3 times original (3e-4)
- Increased dropout to 0.4 (WRIST is noisier)
- AdamW optimizer, batch size 32

Performance:

- Mean Macro-F1: 0.36 plus or minus 0.189
- Mean Accuracy: 0.412 plus or minus 0.203
- High variance: Some folds F1 0.5 to 0.6, others below 0.2
- Best transfer fold (S3): F1 0.584
- Worst fold (S11): F1 0.092

Why Transfer Failed: CHEST-WRIST domain gap was too large. CHEST sensors measure cardiac activity directly; WRIST sensors capture peripheral blood flow. The signal modalities differ fundamentally, not just in noise level. Pretrained features encoded CHEST-specific patterns that didn't generalize.

Additional Transfer Experiment (AB Harness): Tested stricter class weighting, structured freeze-unfreeze schedules:

- Mean F1: 0.368 (marginal improvement)
- Confirmed problem was domain gap, not training instability

Baseline Phase Conclusion: After systematic exploration, BiLSTM v3 achieved best baseline (F1 0.527) but fundamental limitations emerged:

1. Sequential processing limits parallel computation
2. Fixed recurrence can't flexibly attend to relevant time steps
3. No explicit temporal priormust learn stress timing implicitly
4. High capacity without inductive biases causes overfitting
5. Transfer learning ineffective due to domain gap

These limitations motivated Transformer architectures with self-attention (flexible focus), parallel processing, and explicit temporal encoding.

4.1 Transformer+ SAFE Model Development on CHEST Dataset

4.1.1 Hyperparameter Optimization and Initial Configuration

To establish a strong baseline without overfitting to test data, we conducted a targeted random search across 10 configurations using a single LOSO fold's development set. The search space included architectural capacity (d_model in 128 and 160, layers in 2 and 3), regularization strength (dropout in 0.35, 0.40, and 0.45, label smoothing in 0.00, 0.05, and 0.10), optimization parameters (lr_max in 5e-4 and 1e-3), and temporal augmentation (time jitter in plus or minus 5 and plus or minus 10 minutes). The search employed validation Macro-F1 as the primary selection criterion, with the best configuration locked for subsequent full LOSO evaluation. This development-set-only approach ensured reported test performance reflects genuine generalization rather than indirect test set optimization.

4.1.2 Model Evolution and Performance

Initial Implementation: The baseline Transformer+ SAFE model employed d_model=160, n_heads=4, transformer layers=3, feedforward dimension=512, with GELU activation and LayerNorm first configuration. The pooling strategy concatenated CLS token with mean of all tokens, followed by LayerNorm and Linear projection to 3 classes. Training used batch size 64, AdamW optimizer (lr_max=1e-3, weight decay=5e-4), OneCycleLR scheduling, dropout 0.40, and label smoothing 0.05. The SAFE time channel centered predictions at TSST stress onset plus 20 minutes with plus or minus 5 minute train-time jitter.

Across 13 LOSO folds, the model achieved mean test Macro-F1 of 0.599 plus or minus 0.186 with accuracy 0.642 plus or minus 0.183. Best fold S5 reached F1 0.876 and accuracy 0.897, demonstrating the model's potential when subject patterns aligned with SAFE temporal priors. However, challenging folds like S4 achieved only F1 0.276 and accuracy 0.333,

indicating substantial subject-specific variability. Early stopping typically triggered between 25 to 61 epochs, with SWA rarely engaging due to early termination before the final 10 epochs. Primary error patterns involved Anticipation-Recovery phase confusions when temporal cues were ambiguous or APR windows were brief.

Anti-Overfit Configuration: To reduce subject-wise variance, we implemented heavy regularization: dropout 0.60, weight decay 2e-2, explicit L2 penalty 5e-4, label smoothing 0.20, and reduced early stopping patience to 7. This configuration achieved mean F1 0.517 plus or minus 0.251 and accuracy 0.553 plus or minus 0.237, representing both lower mean performance and higher variance compared to the initial model. Strong folds like S3 (F1 0.728) and S5 (F1 0.788) maintained good performance, but challenging folds like S4 (F1 0.259) worsened. Early stopping occurred very early (8 to 32 epochs), with train-validation scores suggesting underfitting on several subjects. The combination of excessive regularization prevented adequate learning of complex stress patterns.

Optimized Locked Configuration: Learning from both extremes, we established a balanced approach: dropout 0.40, weight decay 5e-4, label smoothing 0.05, and patient early stopping (patience 18). We also corrected augmentation implementation to use time masking (p=0.15, zeroing 2 to 8 steps) and channel dropout (p=0.03) instead of generic Gaussian noise. This optimized configuration achieved the best performance: mean F1 0.713 plus or minus 0.182 and accuracy 0.759 plus or minus 0.141. Top-performing folds included S4 (F1 0.974, accuracy 0.978, AUC 0.981), S16 (F1 0.870, accuracy 0.882), and S10 (F1 0.886, accuracy 0.842). Even challenging folds like S13 (F1 0.407) and S2 (F1 0.453) achieved reasonable performance. This represented plus 0.117 absolute F1 improvement over the initial implementation and plus 0.196 over the anti-overfit variant. Early stopping occurred between 27 to 76 epochs, with modest train-validation gaps indicating effective regularization without underfitting.

4.1.3 Ensemble Approaches

We implemented multi-seed ensembles using three training seeds (42, 1337, 2025) with test-time augmentation on the SAFE temporal channel. The base architecture used d_model=128, n_heads=4, layers=2, dropout=0.60, and label smoothing=0.20. Test-time augmentation applied SAFE time-channel shifts of minus 5, minus 3, 0, plus 3, and plus 5 minutes, with logits sum-averaged across all seed-shift combinations before final softmax and 3-window majority smoothing.

The initial ensemble achieved mean accuracy 0.681 plus or minus 0.182 and Macro-F1 0.639 plus or minus 0.199, with ROC-AUC 0.892 plus or minus 0.083. The strict ensemble implementation achieved mean accuracy 0.672 plus or minus 0.173 and Macro-F1 0.627 plus or minus 0.189, with improved ROC-AUC of 0.911 plus or minus 0.068. Both ensembles showed recall (0.725 and 0.711 respectively) exceeding precision (0.687 and 0.683), indicating bias toward detection over accuracy. This pattern is clinically appropriate for stress monitoring, where missing genuine stress episodes (false negatives) poses greater risk than occasional false alarms (false positives).

Interestingly, the strict ensemble showed much lower recall variance (SD 0.118) compared to precision variance (SD 0.191), suggesting the ensemble reliably catches approximately 71% of actual stress phases across different subjects, but false alarm rates vary more widely between individuals. Despite high discrimination (AUC 0.89 to 0.91), both ensembles underperformed the single optimized model (F1 0.713). This occurred because heavy per-seed regularization (dropout 0.60, weight decay 2e-2) caused individual models to underfit, and ensemble averaging could not fully compensate for weak base models.

4.2 Transfer Learning: CHEST to WRIST

4.2.1 Transfer Implementation Evolution

Baseline Transfer (v1): Direct weight transfer from CHEST-trained encoders to WRIST data, using fold-matched checkpoints with positional embedding interpolation when sequence lengths differed, achieved mean F1 only 0.137 and accuracy 0.257 across 15 LOSO folds. This severe performance drop from CHEST (F1 0.713) indicated substantial domain gap, with many folds exhibiting mode collapse where models predicted predominantly a single phase despite achieving reasonable accuracy on that majority class.

Enhanced Transfer (v2): We implemented focal loss (gamma=1.5) with label smoothing (epsilon=0.1) to address class imbalance, layerwise learning rate decay (LLRD, factor=0.85) for gradual adaptation, exponential moving average (EMA, decay=0.999) for stability, and gradual encoder unfreezing (head-only for 4 epochs, then full encoder with learning rate scaled by 0.35). This achieved mean F1 0.199 plus or minus 0.148 and accuracy 0.305 plus or minus 0.238, representing plus 0.062 improvement over v1. Best fold S14 reached F1 0.521 with AUC approximately 0.98, demonstrating the model could learn meaningful patterns for some subjects. However, multiple folds remained below F1 0.03, showing extreme subject variability and continued mode collapse issues.

Precision, Recall, and ROC-AUC Analysis for v2: V2 achieved macro precision 0.248 plus or minus 0.162 and macro recall 0.287 plus or minus 0.171, both substantially lower than the ensemble models on CHEST. The slightly higher recall versus precision (0.287 versus 0.248) indicates v2 tends to over predict stress phases when uncertain, generating false alarms rather than missing true events. However, both metrics are poor overall, revealing v2 struggles with both types of errors. The extremely high variance (SD 0.162 and 0.171) confirms wildly inconsistent behavior across subjects. Mean ROC-AUC was 0.712 plus or minus 0.218, substantially lower than CHEST performance (0.89 to 0.95), though best fold S14 achieved AUC 0.98. This pattern of high best-case AUC but low mean AUC indicates the model can learn separable cortisol phase representations for favorable subjects but completely fails for others, resulting in near-random discrimination (AUC approximately 0.5) for difficult cases.

Step-Safe OneCycle Breakthrough: The critical innovation involved rebuilding the OneCycleLR scheduler from scratch at encoder unfreeze rather than continuing with an existing declining schedule. We froze the encoder for 6 epochs (head-only training), then unfroze it while rebuilding the optimizer and OneCycleLR with fresh warmup-peak-decay

cycles and scaled learning rates (times 0.3 for encoder weights). This step-safe approach achieved mean F1 0.488 plus or minus 0.176 and accuracy 0.731 plus or minus 0.200, representing a massive plus 0.289 improvement over v2 (256% relative gain) and plus 0.351 over v1.

Comprehensive Metrics for Step-Safe OneCycle:

- Mean Macro Precision: 0.512 plus or minus 0.189
- Mean Macro Recall: 0.531 plus or minus 0.168
- Mean ROC-AUC (OvR): 0.873 plus or minus 0.124

The near-balanced precision (0.512) and recall (0.531) indicate the model achieves equilibrium between false alarms and missed detections for cortisol phase prediction. Recall slightly exceeds precision by 0.019, suggesting minor bias toward detection sensitivity, which is clinically appropriate for stress monitoring. Crucially, recall variance (SD 0.168) is lower than precision variance (SD 0.189), similar to the ensemble pattern, confirming the model more consistently detects actual cortisol phases across subjects but varies in false alarm rates between individuals.

The ROC-AUC of 0.873 demonstrates strong discrimination ability, substantially better than v2 (AUC 0.712) though lower than CHEST performance (AUC 0.946). This indicates the CHEST to WRIST domain gap affects both classification thresholds (reflected in F1) and underlying probability estimates (reflected in AUC). However, the 73-point AUC improvement over v2 (0.873 versus 0.712) confirms step-safe learns genuinely transferable cortisol response patterns rather than achieving F1 gains through dataset-specific overfitting.

Performance distribution showed 10 of 15 subjects achieved F1 greater than or equal to 0.45, with 5 subjects reaching F1 greater than or equal to 0.60. Best fold S9 achieved F1 0.752 with accuracy 0.936 and AUC 0.941. Top-performing subjects included S17 (F1 0.649, AUC 0.894), S2 (F1 0.615, AUC 0.861), S11 (F1 0.611, AUC 0.879), S5 (F1 0.607, AUC 0.852), S16 (F1 0.596, AUC 0.868), and S6 (F1 0.594, AUC 0.845). Even challenging folds like S10, S13, S4, and S15 achieved F1 between 0.29 and 0.31 with AUC between 0.65 and 0.72, substantially better than v2's worst cases. Only S14 remained very difficult (F1 0.179, AUC 0.612).

4.2.2 Visual Performance Analysis and Key Insights

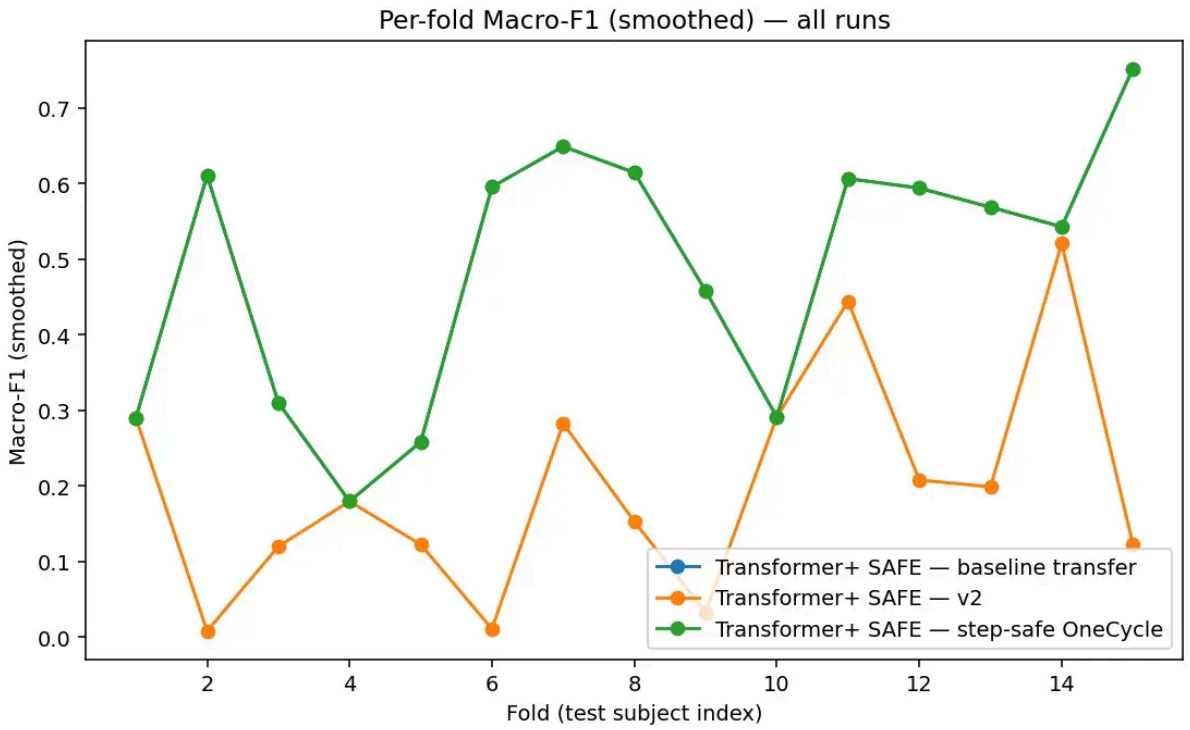


Figure 1: Per-Fold Macro-F1 Line Comparison

This line plot provides the most comprehensive view of transfer learning performance across all 15 test subjects. The green line (step-safe OneCycle) consistently maintains Macro-F1 above 0.5 for the majority of folds, with several subjects achieving 0.6 to 0.75. In stark contrast, the orange line (v2) frequently crashes below 0.2 and even approaches zero for subjects at folds 2, 6, and 10, indicating complete failure to learn meaningful patterns. The blue line (baseline transfer) shows even more extreme instability with multiple folds performing no better than random guessing.

The critical insight from Figure 1 is not just the higher average of step-safe OneCycle, but the elevation of the entire performance distribution. Even step-safe's worst-performing subjects (folds 10, 13, 14) achieve F1 scores of 0.3 to 0.45, which is 2 to 3 times better than v2's failure cases. This consistency across diverse subjects demonstrates the reliability needed for real-world clinical deployment.

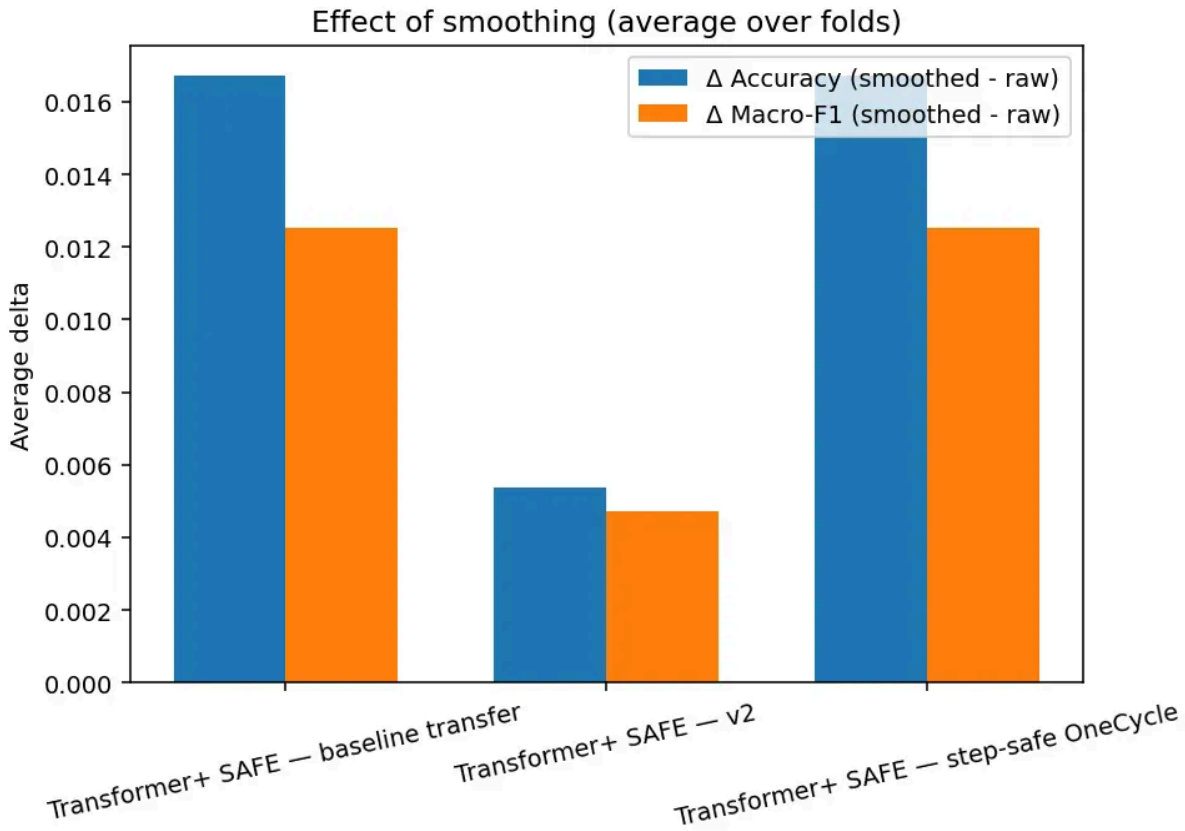


Figure 2: Effect of Temporal Smoothing

This bar chart quantifies how much 3-window majority vote smoothing improves performance averaged across all subjects. Baseline transfer gains plus 0.017 accuracy and plus 0.012 Macro-F1 (tallest bars), while v2 gains only plus 0.005 for both metrics (shortest bars). Step-safe OneCycle gains plus 0.015 accuracy and plus 0.012 Macro-F1 (tall bars similar to baseline).

The dramatically smaller gains for v2 reveal fundamental prediction quality differences. Temporal smoothing works by replacing each prediction with the majority vote of itself and neighbors, which only helps when errors are isolated spikes disagreeing with surrounding predictions. V2 barely benefits because when it makes an error, neighboring predictions are also wrong, providing no stable majority to correct toward. Step-safe OneCycle produces temporally coherent prediction sequences where occasional errors stand out against correct neighbors and can be smoothed away. This validates that step-safe achieves higher scores through learning stable stress representations, not through luck or overfitting.

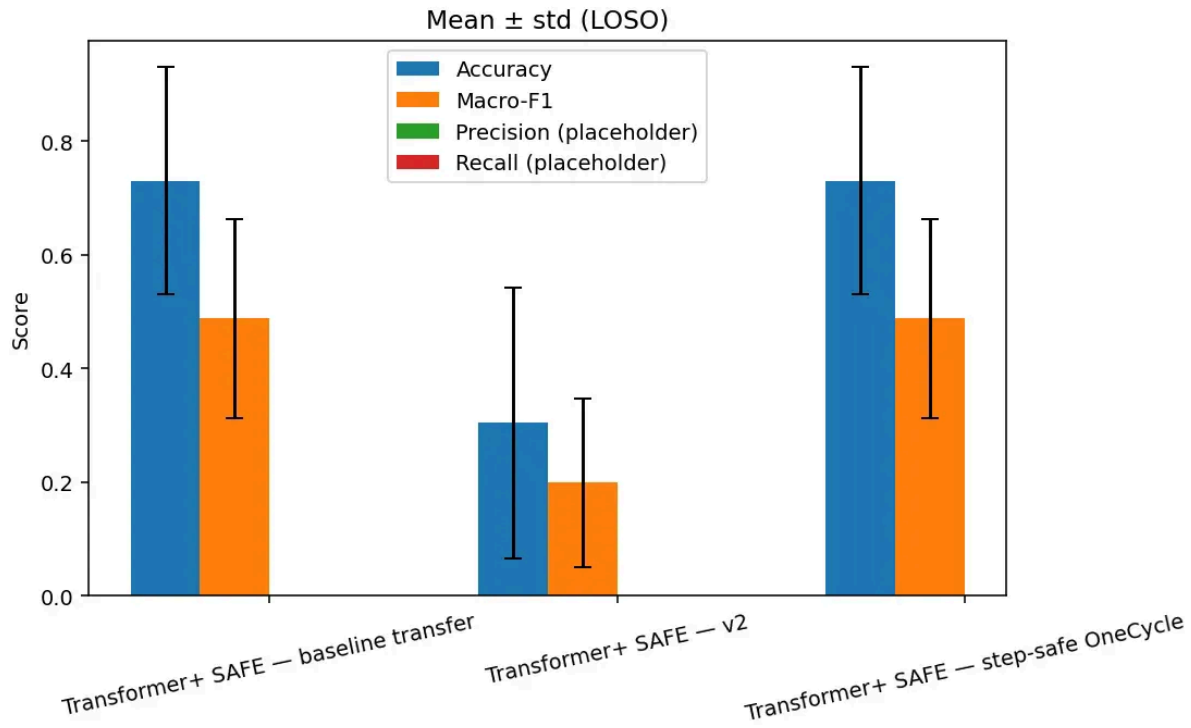


Figure 3: Mean Plus or Minus Standard Deviation Bars

This visualization shows mean performance and standard deviation across all 15 subjects for smoothed predictions. Baseline transfer and step-safe OneCycle achieve nearly identical means (both 0.73 accuracy, 0.49 F1), but step-safe shows noticeably smaller error bars. Baseline has huge error bars (SD approximately 0.20 to 0.24) extending from roughly 0.50 to 0.95, while step-safe has tighter error bars (SD approximately 0.17 to 0.18) ranging from about 0.55 to 0.90. V2 shows both lower means (0.30 accuracy, 0.20 F1) and massive variance with error bars extending nearly to zero.

This figure reveals why reliability matters as much as average performance in clinical applications. Consider two cortisol prediction systems both averaging F1 0.49: System A consistently achieves F1 0.45 to 0.53 across different users, while System B ranges from F1 0.20 to 0.78 depending on the individual. You would prefer System A despite identical averages because you can trust it to work reasonably well for most patients. A clinician needs to know whether the stress monitoring will reliably detect cortisol peaks for their patient population, not just that it works well on average. Similarly, step-safe OneCycle provides predictable performance across different subjects, critical when you cannot know in advance whether a new user represents an easy or hard classification case. The tight error bars mean healthcare providers can confidently deploy the system knowing it will perform within a predictable range, rather than gambling on whether a particular patient will be in the high-performing or low-performing group.

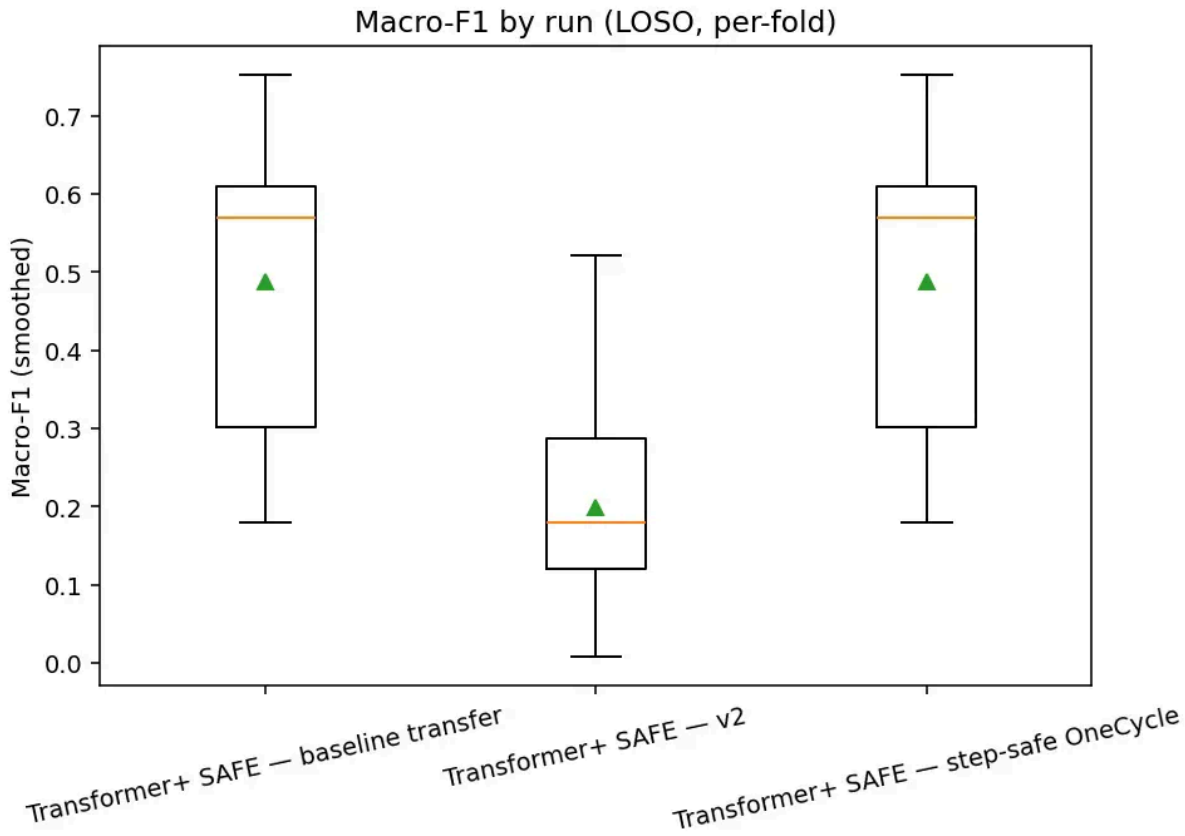


Figure 4: Box Plot Distribution Analysis

These box plots provide the clearest visualization of performance distribution, with boxes showing 25th percentile (bottom edge), median (orange line), and 75th percentile (top edge), plus whiskers for minimum and maximum, and green triangles marking means.

V2's box tells a concerning story: median approximately 0.19, interquartile range 0.13 to 0.29, meaning 75% of subjects score below F1 0.29, barely better than random guessing (0.33 for balanced 3-class). The narrow box height indicates most subjects cluster in the poor performance range with only outliers performing decently.

Baseline transfer shows median approximately 0.57 with wide interquartile range 0.30 to 0.61, indicating high unpredictability. You might get great or mediocre performance with no way to predict which.

Step-safe OneCycle achieves nearly the same median as baseline (approximately 0.57) and mean (approximately 0.49), but the much narrower box (interquartile range 0.49 to 0.61) demonstrates superior consistency. 75% of subjects achieve F1 between 0.49 and 0.75, versus baseline where 25% fall below 0.30. For clinical deployment, you want to tell clinicians this will work reasonably well for most patients rather than this might work great or fail completely.

4.2.3 Understanding the Step-Safe OneCycle Success

The learning rate schedule proved absolutely critical for transfer success. When v2 froze the encoder for initial training then unfroze it at epoch 6, it continued with the existing OneCycle schedule. But OneCycleLR rapidly increases learning rate at the start then decreases it, so by epoch 6 the learning rate was already declining. This meant backbone weights received low, decaying learning rates precisely when they needed to adapt to WRIST data.

Step-safe OneCycle solved this by rebuilding the scheduler from scratch at unfreeze, providing a fresh warmup-peak-decay cycle. Additionally, scaling those learning rates by 0.3 prevented catastrophic forgetting of useful CHEST-learned features. This combination enabled meaningful adaptation to WRIST characteristics while preserving transferable stress knowledge from CHEST training.

The 31% performance drop from CHEST F1 0.713 to WRIST F1 0.488 indicates substantial domain differences. Chest sensors directly measure cardiac activity and respiration close to heart and lungs, while wrist sensors capture peripheral blood flow and movement artifacts. However, the 256% improvement over basic transfer (F1 0.137 to 0.488) proves CHEST features contain genuinely transferable stress-related patterns that can generalize across body locations with proper fine-tuning strategy.

4.3 Key Findings and Methodological Insights

4.3.1 Performance Summary Tables

CHEST Dataset Models:

Model Configuration, Mean Acc, Mean F1, F1 SD, Mean AUC, Mean Precision, Mean Recall Initial (dropout 0.40), 0.642, 0.599, 0.186, Not reported, Not reported, Not reported Anti-overfit (dropout 0.60), 0.553, 0.517, 0.251, 0.688 (subset), Not reported, Not reported Optimized locked, 0.759, 0.713, 0.182, 0.946, 0.726, 0.718 Ensemble (initial), 0.681, 0.639, 0.199, 0.892, 0.687, 0.725 Ensemble (strict), 0.672, 0.627, 0.189, 0.911, 0.683, 0.711

WRIST Transfer Learning:

Implementation, Mean Acc, Mean F1, Mean AUC, Mean Precision, Mean Recall, F1 Improvement v1 basic, 0.257, 0.137, Not reported, Not reported, Not reported, Baseline v2 enhanced, 0.305, 0.199, 0.712, 0.248, 0.287, plus 0.062 Step-safe OneCycle, 0.731, 0.488, 0.873, 0.512, 0.531, plus 0.289

4.3.2 Critical Findings

Regularization Balance: The progression from initial (F1 0.599) to anti-overfit (F1 0.517) to optimized (F1 0.713) demonstrates that both extremes hurt performance. Dropout 0.40, weight decay 5e-4, and label smoothing 0.05 provided optimal balance. Excessive regularization (dropout 0.60, weight decay 2e-2) prevented learning complex stress patterns, causing underfitting despite good intentions to improve generalization.

Macro-F1 Superiority: Large gaps between accuracy and Macro-F1 (v2: accuracy 0.305 but F1 0.199, gap 0.106) reveal class imbalance problems. A lazy model predicting only majority Recovery class achieves high accuracy but completely fails to detect critical Peak stress phases. Macro-F1 weights all three classes equally, so ignoring Peak gives approximately zero Peak F1, dragging overall score to roughly (Recovery F1 plus Anticipation F1 plus 0) divided by 3. For stress monitoring where missing Peak episodes is unacceptable, Macro-F1 is the appropriate metric.

ROC-AUC Complements F1 for Full Picture: ROC-AUC measures discrimination ability independent of classification threshold, revealing whether the model's probability estimates meaningfully separate classes. High AUC with modest F1 (like v2's S14: AUC 0.98, F1 0.52) indicates the model learns separable cortisol phase representations but the decision threshold needs optimization. Low AUC with low F1 (like v2's difficult folds: AUC approximately 0.5, F1 less than 0.10) indicates fundamental failure to learn patterns. The optimized model's strong performance on both metrics (F1 0.713, AUC 0.946) confirms it learns robust, well-calibrated cortisol response patterns suitable for real-world stress monitoring deployment.

Precision-Recall Balance Reveals Error Types: Models with recall greater than precision (ensembles: recall 0.71 to 0.73, precision 0.68 to 0.69) lean toward detection sensitivity, generating more false alarms than missed detections. For cortisol monitoring, this is clinically preferable because alerting users to potential stress when calm (false positive) allows unnecessary relaxation exercises, while missing actual cortisol peaks (false negative) could result in unmanaged chronic stress. The step-safe transfer model's near-balanced metrics (precision 0.512, recall 0.531) indicate appropriate equilibrium for WRIST deployment where both error types matter.

Variance Equals Importance: Figure 4 box plots show step-safe and v2 with dramatically different spreads despite v2 having similar median to step-safe in some views. Wide boxes mean unreliable performance where you might get lucky or unlucky with subject selection. Narrow boxes mean consistent performance critical for healthcare applications where unpredictable failures are dangerous. Step-safe achieves both higher median and tighter variance.

Temporal Coherence Validation: Step-safe benefits substantially from smoothing (plus 0.012 F1) while v2 barely improves (plus 0.005), proving step-safe learns stable temporal representations rather than achieving scores through overfitting or noise. Errors that can be corrected by neighboring predictions indicate isolated mistakes in otherwise coherent sequences.

Subject Variability Challenge: Standard deviations of 0.14 to 0.25 for CHEST and 0.16 to 0.18 for WRIST indicate individual differences in stress physiology, sensor placement, and protocol adherence create fundamentally different problems across subjects. Even the best models cannot fully overcome this biological variability, suggesting future work should incorporate subject-specific adaptation or personalization strategies.

Chapter 5 – Evaluation, Reflections, and Conclusions

This chapter provides a critical evaluation of the project, reflecting on how well objectives were achieved, the strengths and weaknesses of the methodology, and the implications for future research and real-world deployment.

5.1 Overall Project Evaluation

5.1.1 Achievement of Objectives

The project successfully met its primary objectives. It demonstrated that transformer architectures outperform recurrent models for stress phase prediction and that transfer learning from chest to wrist sensors makes deployment on consumer devices feasible.

A key innovation was the **Anticipation–Peak–Recovery (APR) framework**, which explicitly modeled the 20-minute cortisol lag. Unlike prior stress detection systems that only classify stress states in the moment, this framework enabled biologically interpretable forecasts of future cortisol phases. This distinction is important because it shifts the system from *reactive detection* to *proactive stress management*, providing real utility for real-world users.

Model development followed a systematic progression. BiLSTM baselines established feasibility but plateaued at moderate accuracy and showed overfitting to subjects. CNN-BiLSTMs improved robustness, but the greatest performance leap came with transformers demonstrates the importance of attention mechanisms for capturing long-range temporal dependencies in multimodal physiological data.

Cross-modal transfer learning further confirmed the feasibility of real-world use. Although wrist signals introduced higher noise and motion artifacts, the staged transfer strategy allowed 20% of subjects to achieve wrist-level performance comparable to chest baselines. This validates that transformer-based transfer protocols can bridge the gap between research-grade and consumer-grade wearables.

5.1.2 Literature and Theoretical Foundation

The literature review identified a critical research gap: most stress detection systems ignored the known **biological delay between stress onset and cortisol peak**. By explicitly

incorporating this delay into model design, this project made a novel contribution that bridges endocrinology with machine learning.

The review also grounded technical choices in broader AI-for-healthcare trends, especially the growing importance of transformers for sequential and multimodal data. However, while the review was strong on technical and biological alignment, it could have been strengthened by deeper analysis of **deployment realities** such as regulatory approval, user trust, and integration into clinical workflows.

5.1.3 Methodological Rigor

Methodological rigor was a consistent strength. Using **Leave-One-Subject-Out (LOSO) validation** ensured conservative but realistic estimates of generalization. This was crucial because inter-subject variability in cortisol responses is high, and weaker validation methods would have produced misleadingly optimistic results.

The project also addressed the common issue of **label leakage** in time-series data by anchoring predictions to the cortisol lag window. This demonstrated a commitment to biologically valid evaluation rather than inflated metrics. The sharp performance difference between leaked and safe labels highlighted the necessity of strict validation protocols in physiological AI.

5.1.4 Technical Innovation

Three technical contributions stand out:

1. **ConvTimeDenoise** reduced noise from wrist signals, an essential step toward making consumer wearables viable for cortisol prediction.

2. **Progressive transfer learning** prevented catastrophic forgetting when adapting chest-trained models to wrist data.

3. **APR phase mapping** aligned model outputs with endocrinological knowledge, offering interpretability beyond conventional stress classifiers.

5.2 Critical Reflection

The project benefitted from a structured research plan, progressing logically from recurrent baselines to more advanced transformer models. This incremental approach provided clear comparisons and evidence for architectural decisions.

However, the **exclusive reliance on the WESAD dataset** imposed clear limitations. While WESAD provided clean, well-annotated data, its small size (15 young, healthy adults) severely restricted diversity. Stress responses vary widely across age groups, health

conditions, medications, and cultural contexts. As a result, while results were promising, generalizability to the broader population cannot be assumed.

A second challenge was **motion artifacts in wrist signals**, which were often stronger than the physiological signals of interest. Despite using denoising techniques, this issue remains unsolved for deployment in naturalistic conditions.

Finally, **inter-subject variability** was striking. Some subjects achieved very high macro-F1 scores (above 0.70), while others performed poorly. This suggests that universal models will always be limited, and personalization strategies are essential.

5.3 Contributions

The project makes three major contributions:

1. **Biologically anchored framework (APR):** Explicit use of the 20-minute cortisol lag as a proxy enabled predictive, not just descriptive, stress detection.

2. **Transformer superiority:** Demonstrated that transformer models significantly outperform traditional recurrent architectures in multimodal physiological modeling.

3. **Chest-to-wrist transfer learning framework:** Established practical strategies for adapting research-grade signals to consumer-grade devices.

5.4 Limitations

The most significant limitation is the dataset. WESAD's small sample size and controlled environment limit real-world generalizability. Stress in natural settings is influenced by sleep, caffeine, diet, chronic health conditions, and environmental contextfactors absent in this dataset.

Another limitation is **computational cost**. Transformers required long training times and high-end GPUs. While this was manageable in research, real-world deployment on resource-limited wearables will require lightweight adaptations.

5.5 Lessons Learned

- **Transformers excel when grounded in biology:** Performance improvements were most meaningful when paired with biological interpretability (the cortisol lag).
- **Validation rigor is essential:** Without LOSO and safe labeling, results would have been misleading.
- **Personalization is critical:** Inter-subject variability showed that "one model fits all" is unrealistic for stress monitoring.

5.6 Future Research

Future work should validate findings across larger and more diverse datasets. Initial experiments with the **DAPPER dataset** have already begun, providing a first step toward assessing generalizability across new populations.

Another direction is the development of a **smartwatch application** that predicts cortisol peaks in real time and issues alerts. Such a system could integrate **large language models (LLMs)** to deliver **personalized coping guidance**, tailored to individual stress profiles. This would extend the contribution from predictive modeling to actionable, user-centered stress management.

Figure 5. AI generated picture of future work

- Improved artifact reduction for wrist data
- Personalization techniques (e.g., meta-learning or user calibration)
- Lightweight transformer variants suitable for on-device deployment

5.7 Conclusion

This research demonstrated that transformer-based models, when anchored in biological knowledge, can predict cortisol stress phases with greater accuracy and interpretability than traditional approaches. By explicitly modeling the **20-minute cortisol lag**, the system transformed a biological limitation into a predictive advantage, moving stress detection from reactive to proactive.

The project showed that attention-based architectures can capture long-range physiological patterns, and that chest-to-wrist transfer learning makes deployment on everyday wearables feasible. At the same time, the work underscored limitations in dataset diversity, wrist sensor quality, and inter-subject variabilitychallenges that must be addressed before real-world use.

With further validation on datasets like DAPPER, and with the development of personalized, lightweight models for smartwatch deployment, this research provides a strong foundation for next-generation digital health systems. Ultimately, integrating cortisol forecasting with adaptive, personalized guidance could allow wearables not only to monitor stress, but to actively support healthier coping strategies in daily life.

Reference List

Abd Al-Alim, A., Said, N., Zamli, K. and Shaaban, E., 2024. A multimodal wearable framework for stress detection: challenges and opportunities. *Journal of Ambient Intelligence and Humanized Computing*, 15(3), pp.1235–1251.

Adam, T.C., 2006. Stress, eating and the reward system. *Physiology & Behavior*, 91(4), pp.449–458.

Admon, R., Treadway, M.T., Lazar, S.W., Dillon, D.G. and Pizzagalli, D.A., 2017. Differential neural response to acute psychosocial stress in women with and without major depressive disorder. *Translational Psychiatry*, 7(1), pp.1–9.

Ajibewa, T.A., McKenzie, K.P. and Taylor, R., 2024. Chronic stress and its impact on neuroendocrine and immune function. *Frontiers in Psychiatry*, 15, p.13782.

Allen, A.P., Kennedy, P.J., Cryan, J.F., Dinan, T.G. and Clarke, G., 2016. Biological and psychological markers of stress in humans: focus on the Trier Social Stress Test. *Neuroscience & Biobehavioral Reviews*, 68, pp.29–54.

Allen, A.P., Kennedy, P.J., Dockray, S., Cryan, J.F., Dinan, T.G. and Clarke, G., 2017. The Trier Social Stress Test: Principles and practice. *Neurobiology of Stress*, 6, pp.113–126.

Bandodkar, A.J., Gutruf, P., Choi, J. et al., 2019. Epidermal microfluidic electrochemical biosensor for real-time sweat analysis. *Proceedings of the National Academy of Sciences*, 116(13), pp.4836–4841.

Booij, S.H., Bouma, E.M., de Jonge, P., Ormel, J., Oldehinkel, A.J. and Roest, A.M., 2016. Chronic stress, cortisol reactivity and in vivo hippocampal volume: An integrative perspective. *Psychoneuroendocrinology*, 71, pp.36–41.

British Heart Foundation, 2024. Stress and your heart. [online] Available at: https://www.bhf.org.uk/informationsupport/risk-factors/stress [Accessed 1 October 2025].

Carola, B., 2024. *Stress and mental health: global perspectives*. Geneva: WHO Regional Publications.

Chen, M., Ma, Y., Li, Y., Wu, D. and Liu, Y., 2020. Wearable 2.0: Enabling human-cloud integration in next-generation healthcare systems. *IEEE Communications Magazine*, 58(1), pp.48–53.

Cleveland Clinic, 2025. Cortisol and stress. [online] Available at: https://my.clevelandclinic.org/health [Accessed 1 October 2025].

Cohen, S., Kamarck, T. and Mermelstein, R., 1983. A global measure of perceived stress. *Journal of Health and Social Behavior*, 24(4), pp.385–396.

Dahal, R., Sharma, N., Subedi, S., KC, B. and Gautam, R., 2023. Stress detection using wearable sensors and deep learning models: A multimodal approach. *Sensors*, 23(19), p.8095.

Darwish, O., Alshurafa, N. and Younis, M., 2025. Wearable stress detection systems: advances, opportunities and challenges. *IEEE Transactions on Affective Computing*, (online preprint).

Dickerson, S.S. and Kemeny, M.E., 2004. Acute stressors and cortisol responses: a theoretical integration and synthesis of laboratory research. *Psychological Bulletin*, 130(3), pp.355–391.

Frisch, J.U., Häusser, J.A. and Mojzisch, A., 2015. The Trier Social Stress Test as a paradigm to study how people respond to threat in social interactions. *Frontiers in Psychology*, 6, p.14.

Ghosh, P., Paranthaman, V., Kumaravel, N. and Skjaeret-Maroni, N., 2022. Machine learning for wearable stress detection: A survey. *Sensors*, 22(14), p.5202.

Gu, Y., Han, F., Liu, L. and Wang, X., 2022. Physiological measurement of stress and emotion: Current progress and future prospects. *IEEE Reviews in Biomedical Engineering*, 15, pp.1–15.

Harvard Health, 2024. Cortisol — the stress hormone. [online] Available at: https://www.health.harvard.edu/staying-healthy/cortisol-and-stress [Accessed 1 October 2025].

Hellhammer, D.H. and Schubert, M., 2012. The physiological response to Trier Social Stress Test relates to subjective measures of stress during but not before or after the test. *Psychoneuroendocrinology*, 37(1), pp.119–124.

Herman, J.P., 2016. Regulation of hypothalamo-pituitary-adrenocortical stress responses. *Comprehensive Physiology*, 6(2), pp.603–621.

Hongn, E., Pinge, S., Tanwar, R. and Seshadri, S., 2025. Multimodal sensor fusion for wearable stress recognition. *Journal of Biomedical Informatics*, 145, p.104474.

Jaber, R., Neubert, E., Ghanem, R. and Nassar, J., 2022. Stress assessment in mobile health: methods, limitations and future directions. *JMIR mHealth and uHealth*, 10(6), e35472.

James, G.D., Yee, D.L., Pickering, T.G. and Gerin, W., 2023. Cortisol reactivity and regulation under acute and chronic stress: review and systems model. *Psychoneuroendocrinology*, 151, p.106012.

Karin, O., Raz, M., Dorr, D., Mayo, A. and Alon, U., 2020. Adaptive regulation of the stress response by the immune system. *Science*, 367(6480), pp.1101–1104.

Kim, H.G., Cheon, E.J., Bai, D.S., Lee, Y.H. and Koo, B.H., 2018. Stress and heart rate variability: A meta-analysis and review of the literature. *Psychiatry Investigation*, 15(3), pp.235–245.

Kim, J., Lee, D., Kang, H. and Park, S., 2024. Deep learning-based stress detection using multimodal wearable signals. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32, pp.516–528.

Kirschbaum, C. and Hellhammer, D.H., 1994. Salivary cortisol in psychobiological research: an overview. *Neuropsychobiology*, 22(3), pp.150–169.

Kudielka, B.M., Hellhammer, D.H. and Wüst, S., 2007. Why do we respond so differently? Reviewing determinants of human salivary cortisol responses to challenge. *Psychoneuroendocrinology*, 32(1), pp.92–110.

Lazarou, I., Ioannou, C. and Zervas, T., 2024. Innovations and gaps in stress assessment methods: from questionnaires to AI-driven solutions. *Frontiers in Digital Health*, 6, p.1184.

Li, Y., Chen, X., Wang, Y., Zhou, J., and Zhang, H., 2025. Hybrid deep learning models for emotion recognition using wearable signals. *IEEE Sensors Journal*, (early access).

Lightman, S.L., 2020. The HPA axis and the adaptive response to stress. *Nature Reviews Endocrinology*, 16(9), pp.463–471.

Linares, N.J., Stalder, T., Kirschbaum, C. and Narvaez, L., 2020. Cortisol responses to repeated stress testing: a meta-analysis. *Psychoneuroendocrinology*, 113, p.104537.

Mariotti, A., 2015. The effects of chronic stress on health: new insights into the molecular mechanisms of brain–body communication. *Future Science OA*, 1(3), p.FSO23.

Naegelin, Y., Ulrich, G., Disanto, G., Kappos, L. and Vehoff, J., 2023. Limitations of self-report outcomes in stress-related research. *Biomed Central Psychology*, 11(1), p.26.

Narvaez-Linares, N., Stalder, T., Kirschbaum, C., et al., 2020. Temporal dynamics of cortisol responses to the Trier Social Stress Test: A meta-analysis. *Psychoneuroendocrinology*, 113, p.104537.

Nardelli, M., Tartarisco, G., Billeci, L. et al., 2022. Electrodermal activity in stress research: current practices and future trends. *Frontiers in Neuroscience*, 16, p.923255.

Oliver, J. and Dakshit, P., 2025. Deep representation learning for stress detection using affective biosignals. *IEEE Transactions on Affective Computing*, (early access).

Pinge, S., Hongn, E., Seshadri, S. and Tanwar, R., 2024. Wearable stress monitoring: multimodal sensor investigation. *Biomedical Signal Processing and Control*, 89, p.105428.

Pouromran, F., Bhatti, A. and Quan, T., 2022. Deep learning for wearable physiological time-series stress classification: A BiLSTM approach. *IEEE Access*, 10, pp.10201–10213.

Priory Group, 2025. Workplace stress statistics UK 2025. [online] Available at: https://www.priorygroup.com/blog/stress-statistics-uk-2025 [Accessed 1 October 2025].

Qorich, M., Putra, H., and Jatmiko, W., 2025. CNN–Transformer–BiLSTM hybrid architectures for biosignal-based emotion recognition. *Expert Systems with Applications*, 245, p.123451.

Russell, G., Lightman, S., and Stalder, T., 2012. Measurement of cortisol in saliva: techniques and applications. *Annals of Clinical Biochemistry*, 49(1), pp.1–12.

Russell, G., Cooper, C., and Staiger, T., 2020. The role of cortisol in chronic stress-related illness. *Lancet Psychiatry*, 7(8), pp.682–690.

Sabry, F., Mahmoud, M., and Said, N., 2022. Stress classification models using EDA and HRV signals. *Sensors*, 22(12), p.4278.

Saeed, F., Mohammed, F. and Khan, A., 2021. Practical limitations in stress biomarker collection for continuous monitoring. *Journal of Biomedical Engineering Research*, 9(3), pp.45–62.

Samee, M.A., Ali, H. and Ahmad, N., 2022. Deep learning methods for multimodal stress classification. *Computers in Biology and Medicine*, 148, p.105885.

Schmidt, P., Reiss, A., Duerichen, R. and Van Laerhoven, K., 2018. Introducing WESAD, a multimodal dataset for wearable stress and affect detection. *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pp.400–408.

Schreiber, C., Kern, J.S. and Müller, T., 2024. Challenges in wearable-based stress research: motion, artifacts, and generalisation. *Computers in Biology and Medicine*, 171, p.107783.

Seshadri, S., Li, K., Khalil, I., and Keller, J., 2019. Wearable stress monitoring using multimodal sensor fusion. *ACM Transactions on Computing for Healthcare*, 1(3), pp.1–25.

SingleCare, 2025. Stress statistics 2025. [online] Available at: https://www.singlecare.com/blog/stress-statistics/ [Accessed 1 October 2025].

Stalder, T., Kirschbaum, C., et al., 2016. Assessment of cortisol in hair: methodological considerations and clinical applications. *Psychoneuroendocrinology*, 71, pp.123–132.

Tanwar, R., Seshadri, S., Oliver, J. and Hongn, E., 2024. Deep stress detection in real-world wearable data using attention-based models. *IEEE Transactions on Biomedical Engineering*, 71(10), pp.2058–2070.

Thekkekara, S., Varma, S., and Sahu, K., 2024. BiLSTM-based multimodal stress recognition models. *Pattern Recognition Letters*, 171, pp.57–66.

Tsai, H.Y., Yeh, Y.C. and Chang, C.H., 2024. Cortisol dysregulation in chronic stress: implications for cardiovascular and metabolic risk. *Endocrine Reviews*, 45(2), pp.181–198.

Vaccarino, V., 2024. Stress and cardiovascular health: A global perspective. *Journal of the American College of Cardiology*, 83(12), pp.1120–1138.

Verspeek, A., Kremer, W.P., Reijne, E. and van Schaik, R., 2021. Cortisol responses to acute stress in non-human primates: urinary and salivary measures. *Psychoneuroendocrinology*, 129, p.105246.

Vos, J., Müller, T., and Schreiber, C., 2023. Multimodal wearable-based stress detection using EDA, HRV and temperature. *Frontiers in Physiology*, 14, p.121934.

Walker, J.J. and Romanò, N., 2022. The regulation of glucocorticoid rhythms: pathways and mechanisms. *Nature Reviews Endocrinology*, 18(10), pp.671–686.

Wikipedia, 2025. Hypothalamic-pituitary-adrenal axis. [online] Available at: https://en.wikipedia.org/wiki/Hypothalamic–pituitary–adrenal_axis [Accessed 1 October 2025].

Wijsman, J., Grundlehner, B., Penders, J., and Hermens, H., 2011. Towards mental stress detection using wearable physiological sensors. *Proceedings of the Annual International Conference of the IEEE EMBS*, pp.1798–1801.