Stress Detection Using Wearable Device Data: A Knowledge Discovery and Recurrent Deep Learning Approach

by

Arick Sarkar
22101016
Ibraheem Ibn Anwar
22101040
Sharmin Ahmed Nova
22101028
Fahad Al Mahmood
22101033

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
January 2026.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Arick Sarkar

22101016

Ibraheem Ibn Anwar

22101040

Sharmin Ahmed Nova

22101028

Fahad Al Mahmood

22101033

Approval

The thesis/project titled ”Stress Detection Using Wearable Device Data: A Knowledge Discovery and Recurrent Deep Learning Approach” submitted by

1. Arick Sarkar (22101016)
2. Ibraheem Ibn Anwar (22101040)
3. Sharmin Ahmed Nova (22101028)
4. Fahad Al Mahmood (22101033)

of Fall 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in January 31, 2026.

Examining Committee:

Supervisor:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Associate Professor & Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

Increased attention to mental well-being in recent days has created a need for non-invasive and continuous methods of stress monitoring. Conventional assessment tools like EEG based systems are accurate although are clinical and unsuitable for day to day usage. A hybrid methodology combining knowledge discovery techniques with deep learning to analyze multimodal physiological signals collected from public datasets such as Wearable Stress and Affect Detection (WESAD) provide accelerometer (ACC), skin temperature (ST), and heart rate (HR) data that are collected using consumer-grade wearables like smart wristbands. For the preprocessing part, signal cleaning, Z-score normalization, followed by time-frequency domain feature extraction has been brought out to identify stress relevant trends. The study implements a modified version of recurrent deep learning model currently called Modified DA-SKIP RNN to handle temporal dependencies and nonlinear meaningful patterns in the data stream,using Advanced multi-modal architectures combining per-channel embedding, hierarchical-attention, bidirectional GRU, and for personalized affective computing, subject adaptation has been performed using transfer learning procedure. This architecture is further refined with a Projection Layer for dimensionality compression and a Domain-Aware Normalization step that helps representing subject specific and contextual results leading to better generalization across individuals. The Modified DA-SKIP RNN achieves 99.18% accuracy for binary classification and 96.39% four-class accuracy under non-subject-independent evaluation, while subject-independent Leave-One-Subject-Out (LOSO) evaluation demonstrates 94.34% accuracy after a short calibration phase, confirming effective generalization to unseen individuals. The implications of this research is towards the developers, educators, and mental health practitioners for a easily accessible and data driven stress management tool. With the emerging focus on personal well being in recent times, this study will contribute to the real world implementation.

Keywords: Stress Detection, Wearable Devices, Multimodal Physiological Signals, Deep Learning, DA-SKIP RNN, Hierarchical Attention, Per-Channel Embedding, Domain-Aware Normalization, Knowledge Discovery, Time-Series Analysis

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Background

Psychological stress is a widespread phenomenon caused by a range of factors such as: stressful schedule, conflict with other people, serious changes in life and many more that may result in the significant influence on the physical and mental condition of a person causing such problems as anxiety, cardiovascular diseases, and gradual declinement in working performance [2]. Stress management in traditional practices has been accomplished through various methods like mindful exercises, therapy or drugs that in most cases are not feasible in real time application, and cannot be implemented properly unless the person is in the practice. Stress management practices have limited success in continuous observation of the situation. The modes of measurements, such as the self administered questionnaires or the electroencephalography (EEGs) are not amicable in ensuing daily measurement since they are not easily accessible, include high cost in affording, or need special devices. In their turn, wearable gadgets may be one of the alternatives since they will enable sensing stress in realtime in a nonprocedural way by utilizing physiological data. The measurements include heart rate also denoted as HR, electrodermal activity (EDA), skin temperature (ST) and motion when measured by accelerometers (ACC) and all these are correlated with the stress responses and are collected by instruments like Empatica E4 wristband [3]. The proposed study by indexing these bio signals through newer techniques of machine learning and recurrent deep learning will be able to develop new intelligent systems which will be able to use these bio signals so that the user can recognize the stress in sequential data and be able to manage the stress proactively. The proposed solution could be used to make scalable solutions to detect and counteract stress more consumer friendly and hence reduce the need to use invasive, resource intensive solutions. As the field of digital health grows rapidly, this piece of work aims to improve individual well being by continuous detection of stress. In contrast to limitations of traditional approaches, this study involves incorporating multimodal physiological sensing accompanied by adjusting the deep learning structure. It has been augmented with DA-SKIP Recurrent Neural Network having attention-based fusion, adaptive temporal gating and subject aware normalization. This inclusion enables the machine to intellectually learn long term dependencies, customize answers in light of individual variation and run in real time with wearable data. The suggested framework provides a step towards the availability of affordable digital mental health options that people can use every day.

1.2 Rational of the Study or Motivation

Though wearable sensors and machine learning have already delivered some encouraging breakthroughs even with the existing saturated market, there is still a number of practical and scientific issues in actual introduction of stress detecting systems. Physiological responses to stress differ greatly among individuals such that it is hard to develop a general rule of detecting stress. Additionally, the recorded biosignals through wearable are usually noisy, incomplete, or imbalanced, and that makes it hard to build solid machine learning models. Although the EEG-based stress monitoring methods are very accurate, their operational complexity fails to make them ideal in real-life and real-time application. Henceforth, it becomes apparent that such a system that will continually be able to detect the stress through the wearable devices with a degree of interpretation and accurate detection is needed in greater capacity. The motivation behind this research lies in filling the gap between sophisticated stress detection methods by developing a personalized stress detection framework that can adapt to the highly variable physiological responses observed across individuals.

1.3 Problem Statement

This paper deals with the requirement of an effective and smart stress monitoring system capable of functioning on the basis of physiological measures acquired by consumer level sensors. Namely, the current challenge is how to develop a system that will be capable of effectively processing the noisy and incomplete multimodal biosignals and generalize across individual stress responses (regardless of the inter-subject variability in the latter) and be computationally lightweight to be used in a real-time environment. In this study, an attempt to address each of these issues is proposed by unifying standard methods of domain-driven knowledge discovery with a variant of the DA-SKIP RNN architecture in order to achieve a scalable and personalized stress detection solution that can be applicable to real life settings. The key research areas this study aims at finding are,
Questions:

- RQ1) What significance does knowledge discovery technique imparts in improving stress detection from noisy wearable sensor data?
- RQ2) Can personalization of stress detection for diverse users be possible with Recurrent deep learning models with diverse physiological responses?

1.4 Objective

Psychological stress presents a great threat to the welfare of individuals which is usually aggravated by pressurizing lifestyles and living conditions . There has been an increasing demand in effective and real time stress monitoring since most stress monitoring techniques cannot be used daily because they are either invasive such as

EEG or complex such as self reports[42]. In view of this gap, this research study seeks to fill it by relying on wearable technology which offers a discrete and scalable solution to stress detection. Using physiological data on the heart rate, electrodermal activity, skin temperature, and motion, this research study aims at creating a system that blends well into people's lives. High Tech computational methods such as machine learning and deep learning allow deriving meaningful patterns in complex biosignals [27].

Such a method will not only increase the accuracy of perception of stress but will also favour proactive management of mental health. The availability of the solution through the use of consumer grade wearable devices makes it viable since it improves accessibility. The main purposes of the current research are:

- To create a stress sensing system using data from wearable devices which focuses on physiological signals (e.g., EDA, HR, ST, and ACC) to monitor stress.
- To use knowledge discovery techniques for feature extraction, pattern recognition, and preprocessing in order to produce data that is clearer and easier to understand.
- To create and evaluate recurrent deep learning models specifically, modified DA-SKIP—RNN in order to identify intricate patterns in physiological signals and learn temporal dependencies.
- To apply proposed model on publicly available datasets such as WESAD and evaluate their performance using powerful evaluation metrics (e.g., F1-score, AUC, etc.).

To complete this study's objectives we have come up with a solid workflow that incorporates traditional approach towards stress classification but is implemented in our modified structure. Figure 1.1 is a detailed workflow that complements the study objectives.

1.5 Methodology in Brief

This research study proposes an instantaneous stress detection system that integrates a knowledge-discovery workflow with recurrent deep learning models using physiological signals from consumer grade wearable devices. The system is developed and evaluated using the publicly available Wearable Stress and Affect Detection dataset commonly known as WESAD. This is a multimodal dataset which contains synchronized recordings from 15 subjects collected with the Empatica E4 wristband across multiple affective states.

To ensure temporal consistency across heterogeneous sampling rates, all sensor streams are downsampled to a unified sampling frequency of 4 Hz and segmented into overlapping windows of 30 s (120 samples) with 75% overlap. Frame-level labels are resampled, transient segments are excluded, and each window is assigned a single class label using majority voting.

Preprocessing is performed under Leave-One-Subject-Out (LOSO) cross-validation to prevent leakage. Z-score normalization is fitted only on training subjects, and a

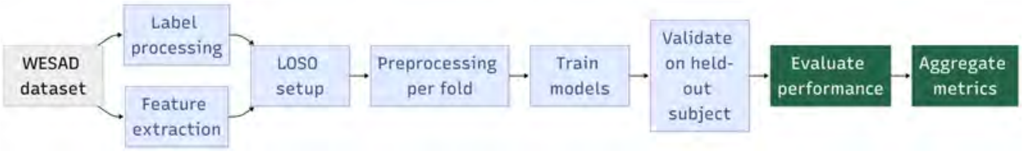

Figure 1.1: Research design framework for stress detection using wearable signals and deep learning models.

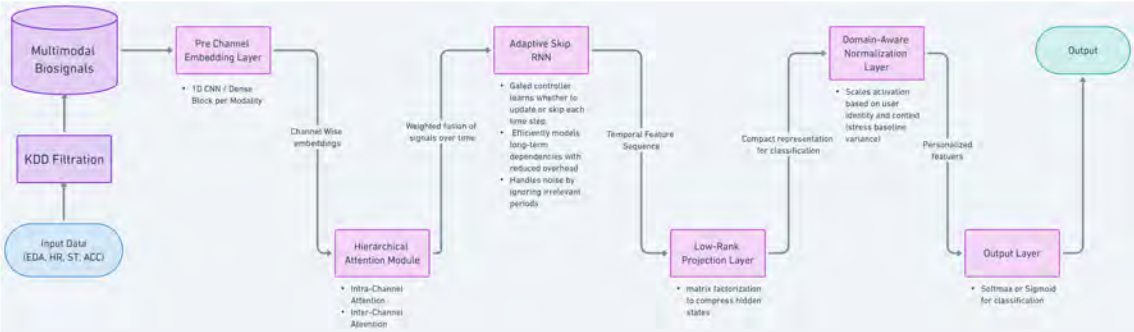

Figure 1.2: Modified DA-SKIP-RNN architecture with per-channel embedding, hierarchical attention, and adaptive skip gating.

domain-aware normalization mechanism is integrated into the network to adapt to inter-subject variability.
A hybrid feature engineering strategy is applied in parallel with raw sequence learning. For each window, six statistical descriptors per modality and cross-signal correlations are extracted, yielding an approximately 30-dimensional feature vector that is fused with deep temporal representations inside the model.
Model development focuses on DA-SKIP variants (original and modified). The modified architecture introduces per-channel embeddings and hierarchical attention to capture both intra-channel temporal salience and inter-channel modality importance. Temporal modeling is performed with recurrent units, and adaptive skip gating selectively skips hidden-state updates to improve long-range dependency modeling.
To mitigate class imbalance, conservative balancing is applied only to training folds, while validation and test data remain unchanged. Models are trained under a consistent regime (class-weighted cross-entropy, Adam optimization, learning-rate scheduling, and early stopping). Performance is reported using macro and weighted F1-scores, precision, recall, confusion matrices, and AUC-ROC aggregated across LOSO folds.

1.6 Scopes and Challenges

This research study is mainly devoted to the creation of a real time noninvasive system of stress detection based on the data of widely available wearable devices, including smartwatches, wristbands. The main idea behind it is to explore how heart rate, electrodermal activity EDA, skin temperature, and motion physiological signals may be processed with the help of machine learning and deep learning methods to properly define stress levels. It also uses repeated deep learning techniques,

i.e., DASKIP architecture, which successfully convey the complex and timelong process of physiological data. Moreover, the algorithm has incorporated methods of knowledge discovery to increase the readability and efficiency of models through reducing noise in the data and focusing on the meaningful features. The overall aim is to have an impact on mental health surveillance, workplace wellbeing, and digital health, since the goal is to provide scalable, fuly dependent on non invasive signals collected from werable devices. With the help of data collected from consumer level wearable devices, this system is intended to be completely integrated with daily life such that it does not require any obtrusive medical devices [42].
Along with its significant promising outcomes a number of important issues are faced in this study. Sensors worn on the human body often suffer noise and missing values and artifacts due to body motion, device placement and other environmental conditions.Thus, a strong data preprocessing technique is an important step towards producing quality inputs. In the real world, the datasets are usually class imbalanced which mainly signifies that a predominance is present for "no stress" occurrences compared to "stress" occurrences, which makes development of the balanced models difficult and requires using the sophisticated methods, such as SMOTE or cost sensitive learning [26]. The variability in the physiological response to stress among individual persons is a problem as formulation of models tend to shift towards extensive personalization for individuals instead of generalizing the whole process [28]. Here it is not possible as it would require customization of large amounts of data. In addition, the modeling design process is challenging to construct models that are precise and lightweight enough to be processed in realtime on resource constrained wearable technology. The subjectivity of stress makes it difficult to acquire ground truth labels which are reliable since self reports and physiological thresholds may not be consistent [10]. Lastly, the fact that deep learning models have certain aspects of a black box makes it all the more relevant to make them more interpretable and transparent especially in the context of medical systems. The current research is designed to create a basis of feasible, dependable, and accessible stress monitoring systems based on wearable devices by means of researching these issues.

1.7 Research Objectives

The primary objective of this research is to design and evaluate a scalable, non-invasive, and real-time stress detection framework using multimodal physiological data acquired from consumer-grade wearable devices, supported by knowledge discovery techniques and recurrent deep learning architectures.

In Chapter 2, the objective is to systematically review existing literature on wearable-based stress detection, knowledge discovery methodologies, and deep learning approaches, with particular emphasis on recurrent neural networks, multimodal feature fusion, personalization strategies, and lightweight model design, in order to identify research gaps, limitations of current systems, and opportunities for improving generalization and real-world deployment.

Chapter 3 aims to specify the functional, technical and operational requirements of the proposed stress detection system through hardware, software, data and ethical requirements analysis and also environmental, societal and economic implications of the proposed solution are also evaluated to make sure that the proposed solution is practical, sustainable and suitable to work with wearables in real-time.

In Chapter 4, the task is to introduce a detailed methodology framework that will combine signal preprocessing, feature engineering, and multimodal data fusion with our proposed DA-SKIP RNN architecture, to successfully characterize the temporal dependencies, subject variability and cross-modal reactions by using physiological signals.

Chapter 5 aims to determine how the proposed models perform across subject-independent and subject-dependent experimental setups, the evaluation metrics in the form of standardized F1-score, AUC, precision, recall and confusion matrices, will be used to analyze how well the model performs, the model stability, and the ability to personalize the models.

Finally, in Chapter 6, the objective is to summarize the key findings of the study, highlight the contributions of the proposed stress detection framework to wearable-based digital mental health systems, and outline future research directions, including model optimization, real-world deployment strategies, and extensions toward explainable and energy-efficient physiological monitoring systems.

Chapter 2

Literature Review

2.1 Preliminaries

Wearable gadgets like the Empatica E4 wristband record multiple signals that contain electrodermal activity (EDA), photoplethysmography (PPG) related to heart rate and HRV, skin temperature, and motion. Other devices, such as the Shimmer3 ECG unit or sensors worn on the chest, have more detailed data including the ECG, respiration, and EMG are more practical in a controlled environment [4]. On contrary to medical devices most study focuses on data from consumer friendly devices such as smartwatches and wearable scanners which are easy to use and relatively accurate in addition to being cheap.

Raw sensor data is converted to more useful information with knowledge discovery in order to understand the data better. This is normally carried out in four steps. The initial step is to clean the raw data by deleting errors and noise, as well as, missing values. The next step is data transformation where we standardize measurements between sensors and also simplify it by deleting features that are not of the most importance using methods such as PCA [1]. We will then apply the machine learning process to the data to find any hidden patterns behind stress. Lastly, we also interpret these findings in order to relate them with known stress responses.

Deep learning models learn its valuable features automatically. We apply the advanced RNNs architecture that can fit the time series data in our research. Both the GRU and the bidirectional GRU (BiGRU) effectively learn trends and allow processing (in unidirectional or bidirectional) sequences. In the DA-SKIP model, it adopts attention and skip connections to decompose the task into nonlinearity, periodicity, and linearity [16]. The combination of these models give a potent but streamlined approach to identifying stress given wearable sensor readings.

The WESAD dataset is a multimodal dataset that is recorded using the devices such as the Empatica E4 and chest mounted sensors to record the different physiological signals such as ECG, EDA, EMG, RESP, ACC during, stress, baseline, and amusement. It presents the correctness of 93 percent in classifying binary and as a three class problem, it is about 80 percent [6]. This is quite ideal in real world tests and studies . StressPredict on the other hand, was a subset of WESAD that targets physiological indicators among the nurses and reached a precision of approximately

84.13 as an indicator of occupational stress [7]. Datasets gathered by different labs with a broad choice of sensors (Shimmer3 or Empatica E4) prove to be very different in terms of accuracy, reaching about 75%-76% under LR/SVM and almost 98% under RF in binary classification [6]. Nonetheless, these datasets are diverse and different in size and context, so they are not as generalizable. Sets which are openly available contain such signals as EDA, PPG or HRV. They are helpful in validation, and deep learning models are prone to achieving accuracies of 77.91 percent, but they are restricted when it comes to standardization [11].

The WESAD dataset would be the best option in our study of deep learning based stress detection since it contains multisensor data and abundant annotations that allow comprehensive knowledge discovery and model buildup. Its great rate of accuracy in various classification problems and its reality in real life applications made it fit our deep learning model. There is no doubt that StressPredict dataset can be used to investigate such phenomenon as occupational stress, but the publicly available data can be used as a secondary dataset to verify our findings additionally, the feature points about the standardization and diversity of the WESAD dataset make it the most suitable primary dataset for our evaluation and model development purposes.

2.2 Review of Existing Research

2.2.1 Deep Learning and RNN Based Approaches

With the growing accessibility of physiological data, recorded by wearable sensors, has exploded the number of studies examining the use of machine learning and deep learning to continuously determine degrees of stress. In recent years Recurrent Neural Networks (RNNs) especially Gated Recurrent Units (GRU) and Long ShortTerm Memory (LSTM) networks have emerged as one of the main building blocks in the modelling of sequential relationships in the biosignals sequential in nature like heart rate and electrodermal activity.

Xu et al. [43] (2025) tried solving the gradient vanishing issue in Recurrent Spiking Neural Networks (RSNNs) combining the investigation of models of the neurons and recurrent units which led to Adaptive Skip Recurrent Connection Spiking Neural Network (ASRC-SNN) as a solution. They ran on various temporal sequence data sets: sequential MNIST (S-MNIST), permuted sequential MNIST (PS-MNIST), Google Speech Commands (GSC), Spiking Speech Commands (SSC) and sequential CIFAR (SCIFAR), each of them posing a test to long-term temporal modeling. The models were regular RSNNs, Leaky Integrate-and-Fire (LIF) neurons, Skip Recurrent Connection SNN (SRC-SNN) and their primary contribution ASRC-SNN. SRC was selected as a way to prevent the disappearance of vanishing gradients through skip connections across a time dimension, but ASRC has further advanced this by learning the optimal skip span per layer through an adaptively-learned temperature-scaled Softmax to improve both modeling capacity and robustness. Classification accuracy was applied as a measure of performance, and ASRC-SNN has state-of-the-art performance: 99.57% on S-MNIST, 96.62% on PS-MNIST, 81.93% on SSC, and 96.29% on GSC, and outperformed SRC-SNN and earlier methods. The most important indicators were accuracy and the resistance to fluctuation of hyperparameters. This was achieved successfully because, through the success of ASRC-SNN, new standards were established in long-term sequence learning when compared to SNNs in capturing the time in a superior way and also in their robustness. These findings indicate that the paper is a breakthrough in the area of temporal modeling with SNNs and provides a viable and flexible solution to the gradient vanishing issue and introduces a new record of long-term sequence learning in spiking neural networks.

Nevertheless, RNNs have generally been associated with huge computational costs, vanishing gradients, and a slow learning process with regard to long physiological sequence. To this end, the article by Campos et al. (2018) [5], entitled "Skip RNN: Learning to Skip State Updates in Recurrent Neural Networks" proposes the new method to overcome all the inefficiencies in the application of standard RNNs. The Skip RNN network learns dynamically when to skip or update the internal states of itself and drastically save up the amount of calculation without losing the predictive capabilities. This is attained by using trainable gating mechanism to identify and decide whether the RNN will or will not update its hidden state depending on his or her input relevance. To constrain the amount of updates, the model is outlined using a budget loss function that provides a tunable performance vs computation tradeoff. This method is very applicable to the wearable sensor data stress detection, as the length of the input sequences of this data is long and, in many cases, redundant. Integrating skip connections, Skip RNNs allow the computation resources to concentrate on informative physiological behaviour and disregard noisy or low variance parts. This has the advantage especially on real time systems implemented on wearable devices with power constraints. Moreover, Skip RNN's compatibility with GRU architectures makes it directly applicable to our proposed model using BiGRU. Incorporation of skip mechanisms has the ability to make our stress detection more scalable and the overall system efficient as well as responsive. Notably, the end to end training in Skip RNNs is also applicable where it is easier to implement standard backpropagation and the flexibility of adaptation.

As an extension to the usefulness of the RNN architectures, Sharma et al. (2022) [17] developed a Deep Recurrent Neural Network (DRNN) based Stress Detection model specific to working professionals, a group that was most vulnerable to occupational stress due to deadlines, workloads, and environmental pressures. The authors furthermore identified stress detection as identifying psychological distress by use of physiological and behavioral indicators, and were attracted to DRNNs because of their capabilities to adapt to time series dependencies, the keystone to describing temporal trends in stress reaction. The paper was keen to discuss the shortcomings of historical models of supervised learning including Support Vector Machines (SVMs), Decision Trees (DTs) and Naive Bayes classifiers, which were incapable of capturing the temporal detail needed to effectively model stress as an evolving phenomenon. Such legacy models viewed input data as mere snapshots rather than the sequential data that usually asserts itself with regard to the progression of stress. According to Sharma et al., this way of thinking does not consider the multifactorial and changing nature of workplace stress. In order to overcome these problems, the authors developed a system with a DRNN architecture and augmented by optimal feature selection processes, which led to an increase in both generalization capacity and computing strength. Experimental comparison of the model under implementation is achieved by indicating a gain of over 90% on the classification accuracy, and an increased precision, recall, and F1score than the models used as baseline comparisons. The model also revealed the ability to work with varying levels of stresses in the workplace conditions, which appears to be highly flexible and resilient. The ways this was successfully integrated limited background enduring modeling and sensible feature selection, the proposed framework eradicated the restrictions placed on the previous methods and provided with the potential of scaling and real time performance on occupational stress detection. These results confirm that DRNN provides a very appropriate structure in dealing with stress at the workplace, particularly when compatible wearable or behavioral surveillance devices are involved. Nevertheless, the research admitted its weakness by discussing them as opposed to scalability to larger datasets and the necessity of real world installations validations. In general, the study filled the gap between theory and practice in mental health analytics of workplace using deep learning methods, which is a good precedent in the future studies of the sphere.

Alshamrani (2021) [14] referred to a noninvasive, real time detection of psychological stress using solely the data obtained by a smartwatch in their study. The reasoning behind the proposal can be described as an effort to create a convenient, cheap, and broadly available approach to constant monitoring of psychological stress. To the WESAD dataset, the researchers applied metrics regarding physiological and motion data collecting by a sensor embedded in the chest and the wrist, but they only selected the wrist, being the data most peculiar to smartwatches providers. To categorize time series data, they have applied 2 deep learning modules, namely Fully Convolutional Network (FCN) and ResNet and settled on FCN, since it is easy to understand, has a shortened training duration of 8 minutes when compared to ResNet of 1.5 hours as well as better performance that was empirically observed. There was a substantial preprocessing of the data, which involved eliminating outliers, filtering of the signals, and resampling. Leaveoneout cross validation results showed that the model was able to achieve an average accuracy of 85% and weighted F1 score of 0.84 and area under the curve (AUC) of 0.89 across 15 subjects. The stress class revealed specifically an accuracy of 0.68 and recall of 0.59. Although one of the previous models reported a slightly better accuracy (87.4 percent) than the current one (87.3 percent), it took 255 training cycles against 15 of the present study, which is why the methodology is more fruitful. According to these parameters, there is a high likelihood of using the proposed system in real life providing an efficient and scalable solution to wearable based stress detection. In comparison to Sharma et al.'s work, Alshamrani emphasizes a lower complexity model that still retains robust performance, highlighting the importance of computation time considerations in real world wearable deployments.

Moser et al. (2024) [29] proposed an explainable deep learning model that can improve the accuracy and understandability of stress with wearable biosensors, including Empatica E4 (located on the surface of the skin) that records such signals as Electrodermal Activity (EDA) and Heart Rate (HR). The authors acknowledged the limitations of the existing models especially that they rely on the manually engineered features and covert decisionmaking mechanisms and joined Long ShortTerm Memory (LSTM) networks with the Conditional Generative Adversarial Networks (GANs) and Explainable AI strategies to enhance model transparency alongside data efficiency. The LSTM networks were selected due to the ability to incorporate temporal relations in physiological data, which in this case are really suitable to biosensor signals that vary according to time. Synthetic data augmenting Conditional GANs was used to balance the classes that were underrepresented, to resolve class imbalance and overfitting problems. To accomplish interpretability, the study integrated an explainable AI method, Integrated Gradients (IG) that grades features of the input data and, therefore, expresses physiological characteristics most important to the model predictions. Their propounded system reported an improvement of 3 percent in recall and a 7.18 percent enhancement in precision in comparison to the conventional base line models, which amounts to the success of their implementation of a hybrid framework. IG assessments reiterated that model was in a constant habit of zeroing in on biologically gauging qualities, EDA spikes, during stress forecasts. The desirable property of integrating the generative model is the fact that the training set is expanded beyond, whilst maintaining the characteristics of the underlying physiology signal, thus allowing generalization across different subjects. The multipronged architecture solved the issues that have previously restricted the implementation of deep learning in wearable stress monitoring systems, i.e., the black box problem and limited data availability. Therefore, Moser et al. proved that LSTM combined with GANs and explainable AI can create a highly efficient and explainable framework to measure stress in real time with the help of wearables. The study, however, pointed out the necessity to validate it in uncontrolled and real life environments and increase energy efficiency to use this kind of implementation constantly in wearable devices. Like in the study by Alshamrani, this study focuses on the practicality; this study extends it by addressing the concept of interpretability, which is an essential characteristic of a real life medical and mental health system.

Abd Al Aleem et al. [24] employed deep learning to detect stress in the multiclass case based on the WESAD dataset. The researchers designed a feed forward deep neural network (DNN) including 5 hidden layers and dropout layers to prevent the overfitting, seeking a classification of NoStress, Medium, and High level of stress. Data was processed using a five second window and subsequently two common optimizers, Adam and SGD were used with various batch sizes. The results of the test revealed that the model achieved the highest accuracy, approximately 90.26% using Adam optimizer and batch size=16. It also had good precision, recall and F1 scores. The method's dependence on lab collected data may limit its applicability in noisy, real world settings, despite its potential to effectively use wearable sensor data for stress classification. On the whole this paper demonstrates the possibilities to find stress using deep neural networks and the differences in features between the models which would be better in results. Comparing the results to those of Sharma or Alshamrani, the study shows the advantage of not being satisfied with binary classification but rather implement a more detailed classification scheme that is of importance when the application needs the stress level to be more sensitively monitored i.e., monitoring workplace productivity or clinical assessment.

Furia et al. [21] introduced a new deep reinforcement learning (DRL) algorithm that does not only enable early stress identification but also dynamically adapts its specifications regarding the level of data necessary until a decision is made, which permits to reply quicker. Their solution with the help of the Soft ActorCritic algorithm will teach an agent to make decisions based on the assessment of whether it has obtained adequate data about the stress state to classify it as physical, cognitive, emotional, or relaxation stress using multimodal time series data obtained using the wearable sensor. A dynamic observation window method can enable early classification, and the findings indicate positive developments: the model achieved an accuracy of about 48.5 per cent in a fourclass scenario with data augmentation, whereas clustering similar stress forms in a threeclass task achieved an accuracy of about 68.9 per cent. The method paves the path to the real time uses of stress measurements by significantly reducing the required time of classification (it requires less than 2 time steps in 3 class case on average). The biggest drawbacks were the limitation of training the DRL model and mediocre task results in four classes, which is likely due to the problem of the data, such as its imbalance and the lack of subject diversity. Despite these drawbacks, the introduction of a learning framework that can dynamically determine when it has "seen enough" information represents a paradigm shift in how stress can be monitored in adaptive, low latency systems which makes for a viable option for future works.

To achieve this goal, Ahmad et al. [13] (2021) tried to identify both complex spatiotemporal individual and group activities concurrently based on game spatiotemporal data to get a better insight into player behavior and the game. They built and labeled two datasets based on the games BoomTown and DotA 2 which have a detailed log of low-level actions of players with spatial and time data and also label both individual and group activities. To model them, they created a hierarchical dual attention-based recurrent neural network (RNN) architecture, which introduces both feature attention and temporal attention into hLSTM network architecture. This was preferred to get the significance of certain features at every time step and the time dependence across sequences in addition to group dynamics modeling using the aggregated representations of individual players. The models were coded in PyTorch and weighted cross-entropy loss was used to deal with label imbalance. In BoomTown, the top model (Feature and Temporal Attention LSTM) reached a 75.4 percent accuracy in singular activity recognition, which is greater than the machine learning benchmarks, which use traditional machine learning. Hierarchical Feature Attention LSTM has had a high performance of 66.8 percent accuracy on DotA 2 outsmarting non-hierarchical and classical models in group activity recognition. The measures involved accuracy, recall, and F1-score with confusion matrices emphasizing the important and dissimilarities in the recognition of labels. The models performed better with long periods of repeating activities and less with infrequent or less permanence labels. This aim of the paper is achieved with success, as it sets a new state-of-the-art concerning the activity recognition in games with the use of spatiotemporal data, and the suggested hierarchical dual attention RNN framework is now a major contribution to casual activity recognition not only per individual but also groupwise in complicated multi-agent scenarios.

2.2.2 Knowledge Discovery and Attention Mechanisms

To achieve this goal, Ahmad et al. [13] (2021) tried to identify both complex spatiotemporal individual and group activities concurrently based on game spatiotemporal data to get a better insight into player behavior and the game. They built and labeled two datasets based on the games BoomTown and DotA 2 which have a detailed log of low-level actions of players with spatial and time data and also label both individual and group activities. To model them, they created a hierarchical dual attention-based recurrent neural network (RNN) architecture, which introduces both feature attention and temporal attention into hLSTM network architecture. This was preferred to get the significance of certain features at every time step and the time dependence across sequences in addition to group dynamics modeling using the aggregated representations of individual players. The models were coded in PyTorch and weighted cross-entropy loss was used to deal with label imbalance. In BoomTown, the top model (Feature and Temporal Attention LSTM) reached a 75.4 percent accuracy in singular activity recognition, which is greater than the machine learning benchmarks, which use traditional machine learning. Hierarchical Feature Attention LSTM has had a high performance of 66.8 percent accuracy on DotA 2 outsmarting non-hierarchical and classical models in group activity recognition. The measures involved accuracy, recall, and F1-score with confusion matrices emphasizing the important and dissimilarities in the recognition of labels. The models performed better with long periods of repeating activities and less with infrequent or less permanence labels. This aim of the paper is achieved with success, as it sets a new state-of-the-art concerning the activity recognition in games with the use of spatiotemporal data, and the suggested hierarchical dual attention RNN framework is now a major contribution to casual activity recognition not only per individual but also groupwise in complicated multi-agent scenarios.

Unlike deep learning based techniques, this part will be devoted to the area of stress detection using classical machine learning techniques. The presented works place their focus on knowledge discovery by means of strict feature engineering, data preprocessing, and model selection, with the help of physiological, and behavioral data gathered by wearable devices. Well Known models, such as Support Vector Machines (SVM), Random Forests, and Decision Trees have been considered on the basis of efficiency, interpretability and comparatively low computational costs. Such techniques usually include extracting measurements of handcrafted features out of raw biosignals and using statistical learning to discriminate between states of stress. They are not as accurate as the deep learning models but retain that information for simple interpretability and small resource or real time environments.
With such a premise, a number of authors have developed on the idea of improving the classical machine learning method by incorporating user centric specification and user specific patterns improvements to make the models more accurate and generalizable when applying it to real life. The study by Yekta Said Can, Niaz Chalabianloo, Deniz Ekiz, Javier FernandezAlvarez, Giuseppe Riva, and Cem Ersoy [9] was focused on improving stress detection with smartwatches based on proposing two new techniques, which were personal stress level clustering and decision level smoothing. They aimed at achieving better results on person indifferent stress diagnosis systems in real life and this can be done without monitoring personalized data over a long period of time . The authors gathered the data of 32 participants

who participated in a summer school through the smart gadgets on their wrists (Empatica E4 and Samsung Gear smartwatches). The sessions consisted of baseline, lecture (cognitive load), exam (stress) and recovery (stress management) and among the data measured were heart rate variability (HRV), electrodermal activity (EDA), skin temperature, and accelerometer records. The subjective stress data was also collected with the help of NASA TLX questionnaires and Perceived Stress Scale (PSS14). They used the conventional machine learning models such as Support Vector Machine (SVM), Random Forest, kNearest Neighbors, Linear Discriminant Analysis (LDA), and MultiLayer Perceptron (MLP) to categorize the level of stress. The selection of these models is due to proven success in literature on the previous detection of stresses. The team employed the data preprocessing, artifact removal, modality specific feature extraction, and class balancing processes to increase accuracy. The new hybrid method used baseline measures of stress (low, medium, high) to cluster the participants, so that more effective generalization was possible than with purely person independent models. The paper reports of the study that it reached 82.70 % and 81.82 % accuracy in 2 classes classification through heart rate input and just EDA respectively. Using the decision level smoothing values went up to 92.89% and 94.44% respectively. In some cases, EDA attained a 100 percent session level accuracy using high level accuracy calculation. Classification accuracy, Area Under Curve (AUC) and correlation with self reports were chosen as the key parameters that were assessed. On the basis of these findings, the paper shows that in real life scenarios, the method of stress detection based on the combination of machine learning with the knowledge stemming out of the domain of interest like personal clustering and smoothing is possible and effective. The offered system has superior performance to the baseline models and can be used in a real life context as health monitoring in mobile and wearable devices. Nevertheless, one of the significant weaknesses of the study is that it was based on a comparatively small sample of participants and resorted to using the classic machine learning models that can be rather limited in terms of scalability and flexibility in comparison to the more compelling deep neural networks.

Conducting further research on the same topic, other researchers have considered how overall wearable device monitoring can be limited to general sensors resolution and concentrate on eliminating the necessity of special equipment by conducting real time control and observation. Pekka Siirtola [8] did a research and published an article to identify and verify the real possibility of monitoring stress continuously with the sensors that are available in the currently available smart watches without considering the use of electrodermal activity (EDA) sensor which is not available in the majority of such contrivance. The study made use of available WESAD dataset, which included biosignal data of 15 participants gathered using monster wearable sensor (Empatica E4 wristband) i.e, biosignals, accelerometer, and skin temperature, blood volume pulse (BVP), heart rate (HR), and heart rate variability (HRV). The research was conducted on a binary classification problem where they merged the amusement and relaxed scenarios into one non-stressed category. It used a standard three phase application in human activity recognition of segmentation of the signal into time windows, extraction of features, and subsequently the classifier. The linkage strategy, and the cross validation strategy were compared with different window values (15120 seconds) and classifiers, Linear Discriminant Analysis (LDA), Quadratic Discriminant analysis (QDA) and Random Forest (RF) in a bid to arrive at good configurations. The use of LDA as the main model was prompted by its universal performance, where the best performance was realized with LDA and features of skin temperature, BVP and HR without considering EDA, with the accuracy value of 87.4%. The balanced accuracy was used to measure the performance of the classifications basing on the performance on a per class basis to deal with possible overrepresentation of stressed instances relative to non stressed instances. It was revealed that the larger were the window sizes the better the detection rate and a 120 second window yielded the most accurate result and that heart directed measures of signals was more effective in detecting stress as compared to motionbased values by the accelerometer. Even though the general objective of constructing a user independent stress detection model was achieved, the way it was recognized by the participants differed greatly and ranged between 62.6 to 98.8 percent, which can serve as an indication that the focus on personalized models may prove useful in the further research. In this paper, there exists great evidence that precise and continuous stress detection is a viable phenomenon with the available smart watch sensors in the market and promotes the invention of constant monitoring of the wearable based systems of stress monitoring which are independent of specific biosensors such as EDA .

Moreover, the usefulness of classical models on biosignal dataset has been further highlighted by utilizing the said WESAD dataset in a study by Shedage et al. [31] that evaluated a multiclass classification model that classifies the state as both baseline and stress, amusement and meditation. Shedage et al. [23] utilized the publicly available WESAD dataset that is comprised of physiological signals such as Blood Volume Pulse (BVP), Electrodermal Activity (EDA), and body Temperature (TEMP) obtained on the basis of wearable device on the wrist. They wished to have control upon the ability to distinguish between four mental states, including baseline, stress, amusement, and meditation. First, the data of the BVP were downsampled so that it aligns with lower frequencies because the measurements were taken using different sampling speeds (BVP 64 Hz vs. EDA and TEMP 4 Hz). It was followed by data cleaning based on such methods as the elimination of outliers with the interquartile range (IQR) and Tukey fences. The data was standardised using Zscore normalisation first and then the features reduced using Principal Component Analysis (PCA). Separate models were then developed based on the 15 subjects as well as a global model based on all the subjects. These three classifiers were compared namely; Logistic Regression, Decision Tree and Random Forest. Subsequently, a method known as Stacking Ensemble Learning was applied that essentially took advantage of all the three entailing a Random Forest as the ultimate predictor. The accuracy obtained by this ensemble approach was very high, 99% and 91% in personalized models and general models respectively.

Latest research on ECG-based stress detection highlight the growing importance of attention mechanisms for improving both classification performance and robustness in deep learning models. In contrast to earlier WESAD-based approaches that relied on handcrafted heart rate variability features and traditional classifiers propose an attention-driven hybrid CNN–RNN architecture that explicitly models temporal importance within ECG sequences [37]. The attention mechanism enables the network to focus on physiologically salient time steps associated with stress-induced autonomic responses, rather than treating all temporal segments equally. Experimental results on the WESAD dataset clearly demonstrate the effectiveness of this design: the CNN–BiLSTM model with attention achieves an accuracy of 95.6%, a macro F1-score of 0.954, and an AUC of 0.997, outperforming the CNN–BiGRU with attention as well as previously reported non-attention-based models. The analysis of errors also reveals that the attention-enhanced BiLSTM architecture has lower equal error rates, which makes it a more reliable structure. These results affirm that attention processes are important in stress detection through the change in ability of temporal feature discrimination, classification accuracy and robustness in the real-world ECG-based wearables monitoring framework.

Recent studies on the classification of stresses have shown that combination of attention mechanisms with feature fusion can be used to enhance the classification accuracy and strength. A ResNet architecture with attention and two attention block Attention Module (CBAM) and shallow feature fusion in the classification of stress levels with the WESAD dataset is also present since 2024 [30]. As compared to traditional methods of deep learning which use deep representations alone, the method combines deep features learnt by an attention-based ResNet with handcrafted statistical, time-domain and frequency-domain features of multimodal physiological signals. The CBAM module allows the network to focus on informative channels and time areas adaptively and in effect minimize noise and enhance discriminative ability between stress states.Experimental results strongly validate the effectiveness of this approach: the proposed model achieves an accuracy of 92.74% and an F1-score of 91.19% for four-class classification (baseline, stress, amusement, meditation), and further improves to 97.58% accuracy and 97.51% F1-score in the three-class setting. Ablation studies confirm that CBAM-based attention outperforms squeeze-and-excitation mechanisms and that feature fusion substantially boosts performance compared to deep features alone. These results highlight the critical role of attention-guided feature fusion in capturing complementary physiological patterns, enabling robust and interpretable stress classification suitable for real-world wearable health monitoring systems.

2.2.3 LOSO,Personalization and Transfer learning techniques

With the growing scope of stress detection with wearable devices, the Leave-One-Subject-Out (LOSO) cross-validation approach has become a solid tool for measuring model generalisation across patients, especially when implemented on datasets such as WESAD. Li et al. used a convolutional encoder and a feed-forward neural network to classify three emotional states (neutral, stress, and amusement) based on physiological signals from 15 participants in the WESAD dataset [22]. Their experiment contrasted personalized models, which are trained and tested on the same subject, with generalized models. In particular, the subject-exclusive generalized model was trained using data from all subjects except one and tested on the held-out subject, achieving an average accuracy of 67.65% and an F1-score of 43.05. Conversely, the subject-inclusive generalized model, which incorporated data from all subjects in both training and testing splits, achieved slightly lower performance with an accuracy of 66.95% and an F1-score of 42.50. It has better performance on personalized model, which has an accuracy of 95.06% and F1-score of 91.71, indicating inter-subject variation that negatively affects generalized performance. The authors revealed that electrocardiography (ECG) and electrodermal activity (EDA) are some of the physiological indicators that exhibit interpersonal differences and that LOSO is required to give an objective measurement. This approach identifies the weaknesses of generalized models in real life scenarios where new users might not fit training distributions and a further study of hybrid approaches to minimise variance is suggested.

The recent studies have also revealed that personalization approaches have come to be adequate to overcome the difficulties of inter-subject variations in detecting stresses. Individual data fine-tuning baseline models have demonstrated good results. In Hoang et al., the WESAD dataset was referred to to developing a stress detection system that is specific to university students, and XGBoost was selected as the baseline model due to its efficiency and compromise between precision and speed of computing results [36]. The baseline model was tested and trained on aggregated data of 15 participants between 24 to 29 years old and fitted with physiological measurements of heart rate, EDA, and skin temperature during the baseline, stress, amusement, and meditation conditions. It was relatively successful, however, personalization by means of transfer learning, i.e., fine-tuning the model on personal data, yielded a significant improvement in metrics, and in one of the cases, the F1-score rose by 0.84 to 0.95. All in all, the custom models were always more precise, more recalling and more accurate and showed the possibility of adapting to the specific physiological responses. This article validates the idea of personalization as the viable solution to improve detection in the academic setting where mental health and performance are impacted by stress.

Another stream of personalization uses neural algorithms to fit models to personal physiology it has been shown that customized models work better in classifying acute stress. Stewart et al. postulated that neural processes represented the meta-learning approach that involves classifying stress based on wearable sensor measurements but puts into consideration inter-subject differences mediated through the hypothalamic-pituitary-adrenal axis and the autonomic nervous system [12]. This method was tested on two datasets with the leave-one-participant-out (LOSO) cross-validation of neural process classifiers against standard machine learning models like support vector machines and random forests. The results showed that neural processes were more accurate than conventional models when individual context points in the situations of both stress and baseline were used. The authors have stressed that the model was able to utilize the general trends, but to individualize to unique manifestations of stress, such as variations in heart rate and galvanic skin response. It also has limitations such as the necessity of preliminary contextual information and computational resources, but the findings imply that there are a lot of potentials of scalable, non-invasive mobile health interventions.

Transfer learning has also been explored about enhancing the adaptability of the stress detection models in various environments, and the evaluation of heart rate and variability measures showed both advantages and disadvantages. Albaladejo-Gonza et al. tested the alternatives of supervised and unsupervised anomaly detecting models on the WESAD and SWELL-KW datasets in terms of their transfer learning performance on binary stress detection based on heart rate data [18]. Five supervised models, such as Multi-layer Perceptron (MLP), five anomaly detection models, such as Local Outlier Factor (LOF), were tested. MLP obtained 99.03%

F1-score and 88.89% LOF on WESAD, but 82.75% and 77.17% on SWELL-KW. Performance declined much when models were transferred to the other dataset in MLP (to 57.28% percent on WESAD and 28.41% percent on SWELL-KW), but LOF showed better resilience with 85.00% on WESAD and 70.66% on SWELL-KW. Training on combined datasets also improved generalization: LOF yielded 87.92% and 85.51% on WESAD test data, and 78.03% and 82.16% on SWELL-KW, while MLP achieved 78.36% and 81.33% on WESAD, and 79.37% and 80.68% on SWELL-KW. The authors concluded that anomaly detection models like LOF offer superior transfer learning due to their focus on deviations rather than labeled patterns, recommending multi-context training for real-world applications in workplaces or education. This approach addresses gaps in prior work by highlighting unsupervised methods' potential in non-invasive wearables for stress management.

2.3 Summary of Key Findings

Ref.	Title	Tasks	Datasets	Architecture	Accuracy	F1-score
[6]	Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection	Multimodal stress and affect detection (3-class; binary)	WESAD	LDA, RF, AdaBoost, kNN	3-class: 80%; binary: 93%	3-class: 80%; binary: 93%
[15]	Feature Augmented Hybrid CNN for Stress Recognition Using Wrist-based Photoplethysmography Sensor	Stress recognition (3-class; 2-class)	WESAD	Hybrid CNN (H-CNN)	3-class: 75.21%; 2-class: ∼80%	3-class: 64.15%; 2-class: ∼65%
[36]	Personalized Stress Detection for University Students Using Wearable Devices	Personalized stress detection (binary)	WESAD	XGBoost with fine-tuning	up to 99%	up to 0.97
[37]	ECG-Based Stress Surveillance Using an Attention-Driven Hybrid CNN–RNN Model	Stress surveillance (3-class)	WESAD	CNN–BiLSTM + Attention	95.6%	0.954
[19]	Wrist-Based Electrodermal Activity Monitoring for Stress Detection Using Federated Learning	Multi-class stress detection (3-class)	WESAD	Federated Learning with DNN	99.1% (client); 86.82% (server)	0.97 (avg.); up to 0.99
[18]	Evaluating different configurations of machine learning models and their transfer learning capabilities for stress detection using heart rate	Stress detection with transfer learning (binary)	WESAD, SWELL-KW	MLP, LOF	MLP: 99.03% (WESAD); 82.75% (SWELL-KW)	MLP: 99.03% (WESAD); 82.75% (SWELL-KW)

Table 2.1: Deep learning models in stress Detection

Finally, **Deep Learning Models in Stress Detection** have emerged as effective approaches for stress recognition and related affective-state classification in wearable settings. Within the research study, the recent wearable-based stress detection developments have been facilitated by the presence of multimodal datasets like WESAD, AffectiveROAD, and SWELL, from which high-quality physiological measurements comprise of heart rate, EDA, skin temperature and motion are collected. Knowledge discovery along with time and frequency-domain features derived by using time and frequency-profiling tools such as wavelet transform and FFT have shown desirable results in identifying stress-related trends. SVM, Random Forest, and KNN models and other traditional models of machine learning have been commonly utilized; however, deep-learning techniques with RECURRENT structures (GRU, LSTM) and CNN-RNN architectures have proved to be more effective in reflecting and controlling temporal patterns and non-linear relationships in biosignals. Stresses have been classified with a higher than 90 percent accuracy in some studies and the effectiveness of these models has been proved. Data augmentation, lightweight neural networks and subject-specific fine-tuning are the proposed solutions that will enhance generalizability and feasibility. The overall direction of field is towards more individually edited, dynamic, and computationally efficient models that can aid in the creation of real-time computerized systems of digital mental health surveillance with the help of low-cost wearable devices.

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

In this study, a sophisticated stress detection system is developed using multimodal (EDA, ACC, TEMP, BVP) signals obtained from consumer-level wearable devices. The architecture achieves robust, real-time classification of affective states by integrating knowledge discovery techniques with deep learning models. It relies on high-quality multimodal data captured under precisely controlled experimental conditions, advanced software frameworks to support deep learning operations, and dedicated hardware for model training. These requirements ensure that the system is ready for deployment and remains scalable and reproducible for practical stress-monitoring wearables.

3.1.1 Hardware Requirements

Throughout this study there has been usage of NVIDIA GeForce RTX 4080 Super graphics processing unit with 16GB of GDDR6X memory running at 23 Gbps and 10,240 CUDA cores based on the Ada Lovelace architecture to implement the Modified DA-SKIP RNN and Attention-Enhanced LSTM Multimodal CNN architectures. With a boost clock of 2.55 GHz and memory bandwidth of 736 GB/s, the RTX 4080 Super offers remarkable computational throughput that enables effective training of deep recurrent networks on extensive physiological time series data.While maintaining gradient stability during backpropagation through time, the large VRAM capacity allows batch processing of 30-second windowed sequences across 15 subjects. While maintaining numerical precision, the tensor cores speed up mixed-precision training operations, cutting training time by about 40% when compared to CPU-only implementations. The system also makes use of an AMD Ryzen 9 7950X processor with 16 cores and 64GB of DDR5-5200 RAM to manage feature extraction operations, parallel subject-level processing during Leave-One-Subject-Out cross-validation, and data preprocessing pipelines. Efficient performance on resource intensive tasks such as bidirectional neural network management, hierrarchial attention during architecture training inlcuding subject adaptation in testing phase.

3.1.2 Software Requirements

Machine Learning Framework: The core deep learning system implemented is the PyTorch 2.0.1 which was selected due to the strong automatical differentiation engine along with the extensive support of custom recurrent architectures as well as the capability of having dynamic computational graphs. There are more complicated architectural constructs like per-channel embedding layers, adaptive skip connections and element-wise gating, hierarchical attention modules and multi-head self-attention, and domain-aware batch normalization and learnable subject-specific parameters that can be implemented more easily with PyTorch. The system is powered with NumPy 1.24.3 to perform effective array operations, Pandas 2.0.1 to handle any structured data and SciPy 1.10.1 to perform signal processing tasks such as resampling and filtering data to manipulate and preprocess data. Normalization and standardization utilities , resampling testing and training frameworks including various performance metrics are computed through Scikit-learn 1.2.2. To produce figures of publication quality, confusion matrices, attention weight distributions and visualizations Matplotlib 3.7.1 and Seaborn 0.12.2 have been employed. Implementation of Tensor Board is also done throughout each of the training steps to track validation metrics, gradient steps, and loss curves in real time on per training epochs .

3.1.3 Data requirements

Dataset: This study is primarily focused on the WESAD (Wearable Stress and Affect Detection) dataset, which has become the gold standard optimal benchmark for affective computing research due to its rigorous experimental protocol, multimodal sensor fusion (from chest worn and wrist worn devices), and laboratory verified comprehensive annotation schema. WESAD includes physiological recordings from 15 participants (2 females and 13 males, ages 27.5 ± 2.4 years) who completed the Trier Social Stress Test (TSST) protocol, a validated psychosocial stressor that combines mental math and public speaking exercises. The dataset records synchronized measurements from wrist-worn Empatica E4 wearables (64 Hz blood volume pulse, 4 Hz electrodermal activity, 4 Hz skin temperature, and 32 Hz 3-axis acceleration) and chest-worn RespiBAN Professional sensors (700 Hz ECG, 700 Hz EMG, 100 Hz respiration, and 32 Hz 3-axis acceleration). This design with two sensors is able to compare clinical grade chest sensors and wrist based sensors that are available to the consumers . The experimental design of WESAD comprises five different affective conditions namely, the control (neutral relaxation), stress (TSST induction), amusement (exposure to funny videos), meditation (directed breathing), and transient conditions (state changes). Its popularity among the research community is explained by its ecological validity, standard preprocessing procedures, publicly accessible Python loading packages, and wide picture validation in a range of studies. A combination of both self-report measures (STAI, PANAS, SAM) and physiological recording makes it possible to validate model predictions with self-reported stress, preventing the adoption of patterns that are not correlated with physiological indicators of stress and psychological concepts of stress.

3.2 Societal Impact

The introduction of mental health monitoring based on intelligent wearable stress detectors is a paradigm shift in mental health intervention, as it has left behind the idea of periodic clinical evaluations in favor of an ongoing and non-invasive monitoring system that becomes part of everyday life. Such technologies offer people efficient information on their mental state in real time and therefore offers a preventive measures to reduce the intensity of stress at an early age before it escalates to a long-term/chronic medical condition such as anxiety disorder, depression, or heart related problems. The social implications are distributed at different levels like maximizing productivity at the work place, higher educational achievement, reduction in the cost of the healthcare system since it is detected in good time, and higher standards of living of the exposed groups to chronic stress.

3.2.1 Health preservation and safety assurance

Therapeutic usage: The stress detection pipeline, installed on any consumer wearable, e.g. smartwatch or fitness tracker will provide the user with continuous monitoring capabilities, which will subsequently allow to identify various patterns of stress formation, cardiac rhythmic malfunctions, maladaptive physiological responses. In this method individualized stress profiles have been developed by being mindful of personal characteristics largely emphasizing on domain sensitive normalization and subjectively sensitive embeddings reducing false alarms by inter subject physiological variability. Continuous updates of the user notify them that sustained stress events move beyond set limits and the person can then promptly undertake evidence-based stress management strategies like diaphragmatic breathing exercises, progressive muscle relaxation, or mindfulness meditation. The interpretability of the attention mechanism enables the user to observe the role of various physiological signals that made the most influential impact on stress classification (e.g., high heart rate variability, high electrodermal activity) and it provides the user with a more informed view of himself which can be used in cognitive-behavioral interventions. Longitudinal tracking will allow determining recurring stressful situations related to particular situations (work meetings, commuting, social interactions), and users can make environmental changes or adopt coping mechanisms. In diagnosed cases of anxiety disorders, post-traumatic stress disorder, etc., it can be used by the system as an electronic therapeutic companion that does not replace the work of a specialist, but allows more specific therapeutic methods and changes in medication more effectively based on objective physiological evidence of stress reactivity patterns.

Safety assurance: The conventional stress assessment paradigms also depend on rather invasive clinical tests like laboratory-based polysomnography, salivary cortisol sampling, or electroencephalography (EEG) that involves and requires electrode placement and specialized equipment. These methods are not practical in terms of continuous monitoring and they create measurement reactivity in which the process of assessment leads to the creation of stress. Wearable-based stress detection, on the contrary, is a passive system that uses non-invasive optical sensors (photoplethysmography), galvanic skin response, and inertial motion sensor technology, which does not require any action on the part of the user to be effective in stress detection. Lack of edibles, sterilization or professional oversight removes the risk of infection, the procedural pain and logistical obstacle of the traditional methods. Moreover, edge computing applications, which infer data on wearable processors, guarantee privacy of information because this type of implementation does not require sending raw physiological data to cloud servers, thus eliminating the cybersecurity risk factors and the risk of unauthorized access. In the context of occupational safety in high-stress jobs (emergency responders, air traffic controllers, commercial drivers), the use of real-time stress monitoring can be suggested to supervisors to recognize the states of cognitive overload that damage the quality of decision-making, response time and take appropriate measures to avoid safety-related errors by allowing a break out of rest or delegating the task to another worker.

3.2.2 Cost efficiency

The economic feasibility of the stress detection models on consumer wearable technology is a game changer in terms of the costly clinical infrastructure to cheap, scalable digital health tools. The conventional methods of stress testing in the form of laboratory polysomnography or ambulatory ECG monitoring costs a lot between \$1,000 and \$3,000 dollars per assessment episode, special facilities, trained personnel, and physician interpretation. Consumer-grade wearables like the Empatica E4 (\$1,690) or commercial market smartwatches (\$200-\$500) would, conversely, spread the cost of hardware over long continuous monitoring intervals over years, making the cost per assessment insignificant. The Modified DA-SKIP RNN architecture has been published to run inference on mobile processors in less than 50ms, and is capable of real-time classification without cloud connectivity or subscription expenses. Mass production economies downwardly push prices still, as retail prices of stress-monitoring wearables are projected to drop to the range of \$150-300 by 2027, as sensor production becomes commoditized, and embedded artificial intelligence accelerators become a commodity.

3.3 Environmental Impact

The implementation of the latest deep learning models into wearable applications requires a close attention to the environmental sustainability, considering the issues of energy usage of the models being trained, the limitation of battery life in mobile devices, and the problem of electronic waste due to the obsolescence of the devices. Throughout this study primary focus is given on energy efficient architectural designs, with the use of adaptive skip connections and selective state updates in order to lower the number of computations while maintaining classification accuracy. The environmental analysis would look at the carbon footprint throughout the total lifecycle of the model when it is developed up till when it is deployed and when it is being disposed of.

Power Efficiency: Adaptive gating system is added to the Modified DA-SKIP RNN architecture to dynamically decide when to reform hidden states and when to avoid recurrent computations, which statistically liberates about 30-40% of the number of floating-point operations which would be performed by standard LSTM networks, and does not demolish temporal modeling performance. Transfer learning has the added benefit of minimizing the per-subject fine-tuning energy requirements due to the use of pre-trained base models, and consequently, subject-specific adaptation becomes environmentally sustainable.

Efficient Deployment in wearable devices: The model architecture focuses on compatibility with resource-constrained embedded systems by ensuring optimization in the use of the memory bandwidth and computational complexity used. The resulting deployed model utilizes only 4.2 MB of flash memory while converting to PyTorch and then quantized, which is easily within the storage constraints of current wearable systems. Depthwise separable convolutions in signal-specific RNN branches result in a 60% cut in the number of parameters with no loss in representational capacity. The bi-directional GRU version uses fused CUDA Kernels, which integrate gate calculations into single memory operations that save 45 percent of the DRAM access overhead. The modular architecture provides scalable model dynamicity of which less computationally intensive models (GRU-only baseline) can be run on devices with severe constraints (hearables, smart rings) and more complex models (with attention-enhancing models) can run on smartwatches with dedicated neural processing units. The flexibility allows device manufacturers to balance accuracy and efficiency tradeoff to the best (given device capabilities and user needs).

Hardware Repurposing: The stress detection system proposed can increase the functional life of the wearables already in use by the consumers through software-defined upgrades instead of replacement of the specialized hardware, and hence alleviate the amount of electronic waste produced. A variety of commercially available smartwatches and fitness trackers already have the necessary sensors (PPG optical heart rate monitors, galvanic skin response electrodes, 3-axis accelerators, thermistors) but do not include complex analytics software to derive stress-related information. OTA software updates are capable of retrofitting old devices with stress sensors without hardware upgrades in order to delay obsolescence and minimize what is estimated to be 50 million tons of global e-waste every year. The ability of the model to support a variety of sensor configurations allows it to adapt to device generations and manufacturer and ensures no vendor lock-in and encourages the principles of the circular economy. In case of end-of-life management, component-level recycling protocols recycle useful resources such as rare earth (neodymium found in vibration motors), precious metals (gold found on PCB traces), and lithium found in rechargeable batteries. Academic-industry relationships enable donation programs wherein research grade devices such as Empatica E4 wristbands are recycled and installed in under-resourced clinical care or educational institutions when research studies are completed, providing an opportunity to use the device beyond the research purposes.

3.4 Ethical Issues

The usage of automated stress detection systems creates deep ethical concerns on the privacy of our data, algorithm bias, fair access, informed consent, and the likelihood that the technology will be misused to act as a surveillance tool. This study recognizes these fears and provides technical and procedural controls to prevent risks but still enjoy the benefits of sustained mental health monitoring that the society is aware of. The privacy of data is the most critical issue since the physiological cues are the most personal biometric data based on which one can draw conclusions about sensitive factors, e.g., health status, moods, and habits. We propose a

Privacy-by-Design philosophy by implementation using on-device computing platforms to carry out feature extraction, model inference, and stress classification on wearable computers without sending raw physiological data to third-party servers. Differential privacy mechanisms introduce controlled noise to aggregate statistics which give formal privacy guarantees against re-identification attacks without losing insights in the population. The problem of algorithmic bias is also a serious issue as the low demographic diversity with WESAD data (most of the participants are young, well-educated, European) might lead to poor performance of the under-represented groups of people who might vary in terms of age, ethnicity, cultural reactions to stress, or even physiological factors. This limitation is partially mitigated by our subject-specific calibration in the framework of transfer learning, which provides the personalization of general models to individual baselines with the help of minimal personalization data (5 minutes), yet the systematic control of bias in heterogeneous groups is necessary before wide implementation. Informed consent procedures should explicitly explain data collection procedures, ability of inferences, and storage, and possible dangers such as discriminatory employer/insuring agency use of the information. The data sharing has granular control by users, and opt-in tools of contributing de-identified data to research repositories.

3.5 Project Management Plan

Literature review and design philosophy: The first stage in the research process was the systematic literature review conducted between January and March 2025 and involving 70+ peer-reviewed articles published in such journals and conferences as IEEE Transactions on Affective Computing, Nature Biomedical Engineering, and ACM conferences on ubiquitous computing. The review has summarized available methods on three methodological frameworks, namely: classical machine learning based on handcrafted features (SVM, Random Forests), deep learning models (CNNs, LSTMs, hybrid models), and the new methods such as attention mechanisms, transfer learning, and explainable AI. Some of the gaps that were identified were a lack of focus on cross-subject generalization, restricted performance on resource-constrained machines, inability to interpret black-box models, and a failure to deal with class imbalance in affective data. The design philosophy took a hybrid strategy that integrated both the knowledge based feature engineering with the data based deep learning to exploit the domain knowledge and find the latent physiological patterns. Architectural decisions prioritized modularity (enabling ablation studies), efficiency (targeting real-time edge deployment), and interpretability (through attention mechanisms) to balance accuracy with practical deployment constraints.

Data preprocessing and feature extraction: In data preprocessing operations, sound pipelines were created to deal with the multimodal physiological signals in the WESAD dataset. Raw sensor data collected from Empatica E4 wristbands were systematically quality checked to detect and discard artifact-contaminated chunks caused by motion, loss of skin contact, or sensor failure. All signals were resampled to an identical 4 Hz sampling rate to align all modalities in time.The preprocessing pipeline used 30-second sliding windows with an overlap of 75% (stride = 7.5 seconds) to come up with 120 samples per window to allow enough samples to identify the patterns of stress but still make it computationally feasible. Assigning of labels was done by majority voting across the samples within the each window and windows with mixed labels eliminated to encourage clean training data. An augmentation policy towards conservatism was adopted that oversampled up to 60% of the size of the majority class to deal with the imbalance between classes. This was accomplished by 80% of Gaussian noise addition with a noise distribution parameter of $\sigma = 0.02$, and 20% of SMOTE-like interpolation where interpolation coefficient α to be used was sampled equally between 0.2 and 0.8. Statistical feature extraction was used to compute 24 time-domain features (mean, standard deviation, minimum and maximum, signal energy and mean absolute derivative) of each signal modality, and 6 inter-signal correlation coefficients. This led to 30 handcrafted features which included domain knowledge of physiological stress responses. Finally, Z-score standardization ensured that features across subjects had a mean of zero and unit variance, while subject-specific normalization parameters were retained for domain-aware batch normalization during model training.

Model development and testing: The development of the model and its evaluation covers two architectural variants: original DA-SKIP RNN (adding adaptive skip connections), and the Modified DA-SKIP RNN (adding per-channel embeddings, hierarchical attention, and domain-aware normalization). Each of the models was trained using PyTorch 2.0.1 with hyperparameters that were thoroughly tuned: the hidden size was 128, dropout was 0.2-0.3, batch size 32, and 100 training epochs with early stopping (patience 15 epochs). AdamW optimizer (learning rate 0.001, weight decay 0.0001) was used in the training procedure. The cross-validation was rigorous based on Leave-One-Subject-Out (LOSO) cross-validation, where the subjects were trained on 14 subjects and when the cross-validation was done, the cross-validation was performed on the held-out 15 th subject with rotation across random 5 subjects to gain an unbiased estimate of performance. Complete assessment was done using accuracy, precision, recall, F1-score (weighted and macro-averaged), confusion matrices as well as AUC-ROC curves to test performance in all classification conditions.

Optimization and Testing: The stage of optimization was focused on the experiments of transfer learning to improve the adaptation of the subject, as well as to reduce the amount of data needed to calibrate. LOSO training generated pre-trained models and the pre-trained models were optimized in 10 epochs at low learning rate (0.0001) using 5 minutes of target subject data. Results showed that a significant accuracy improvement can be obtained even in short-term calibration. To illustrate this, the accuracy of Modified DA-SKIP has increased by 32.03% absolute as the accuracy of few-shot transfer at 75.01% grew to 94.34% after 5 minutes of calibration.
Systematic ablation studies were also done by removing architectural elements of per-channel embeddings, hierarchical attention, skip connections, and domain-aware normalization. The contribution of each module was measured in these experiments. The results showed a 4.2% point accuracy increase from hierarchical attention and a 3.8% generalization improvement from domain-aware normalization.

3.6 Risk Management

The creation and implementation of wearable stress detection technologies have several technical, methodological, and operational risks that may jeopardize the validity, reliability, and practicality of the models. To deal with these risks, the possible modes of failure should be identified and mitigation strategies enacted. Subject-dependent calibration requirements, inability to generalize across subjects because of physiological differences, and computational limitations are the primary categories of risk.

Subject-Dependent Calibration: The great variability of physiology among various individuals is one of the challenges of generalized stress detection models. The measurements of heart rate, electrodermal activity, skin temperature, and motion patterns are different depending on age, fitness, medication, chronic conditions, and circadian rhythms. Models trained on aggregate population data tend to show low performance when applied to new individuals whose physiology differs from those observed during training.
The 19% accuracy difference between few-shot transfer performance (75.01%) and subject-calibrated performance (94.34%) in this study highlights this problem. Calibration requirements make deployment challenging because the user must wear devices for at least 5 minutes while experiencing various affective states (baseline, stress, relaxation) to collect representative samples. This causes friction in onboarding and may produce biased calibration data if gathered under non-representative conditions, such as only at rest.
Of special interest is the calibration load when dealing with elderly groups and clinical cohorts whose physiological patterns are not similar to younger, healthier training subjects. In addition, physiological baselines are likely to change over time as a result of aging, fitness variations, medication changes, and circadian effects. Therefore, periodic recalibration may be needed to ensure accuracy.
In order to mitigate these risks, domain-sensitive normalization using learnable subject-specific parameters is applied, allowing models to learn individual baselines using limited data.

Generalization of LOSO: Leave-One-Subject-Out (LOSO) cross-validation is considered the most effective approach for evaluating cross-subject generalization. Nonetheless, there has been noted to be a high degree of interpersonal variation in which the per-subject accuracy has been between 62.6% and 98.8%. This heterogeneity means that some people would respond to stressors in a unique fashion that the models can not be effectively trained to reproduce in other individuals.
The non-conformities are very dangerous especially during deployment where faulty stress predictions can lead to mistrust by the user and lack of interaction. The other weakness of the LOSO protocol is that there could be a train-test distribution mismatch due to the existence of held-out subjects who may not be similar in terms of demographics, health conditions, or stress reactivity. The small datasets, like the one of WESAD (where the number of subjects $N = 15$) exacerbate this risk, as they imply that a single individual provides a large share of the training set, and any individual can be left out in performance, affecting the performance of the training set.
The issue of statistical uncertainty should be tackled to make sure that the test is robust in terms of confidence or bootstrapping. The mitigation approaches are based

on the physiological basis of data augmentation, adversarial training to promote the block of invariance representation, and regularization strategies that include dropout, weight decay, and early stopping to prevent overfitting.
It can also be helpful to constantly track prediction confidence (e.g., entropy) in deployment to be able to detect low-confidence predictions that can need to be revisited or escalated to health professionals. Future research might use larger and multi-site data, such as the SWELL-KW and MIMIC-III to increase demographic coverage and statistical strength.

3.7 Economic Analysis

Economic sustainability of the proposed stress detection system is supported by the fact that it has optimized architectural developments and general pipeline that focuses on the computational efficiency and scalability with a significant cost savings being projected as an alternative to that of the affective wearable-based computing systems currently. More traditional deep learning architectures, e.g. standard LSTM or CNNs that are common in stress detection algorithms, would be expensive to train as they are overparameterized; For instance, training a LSTM model architecture on WESAD would take typically of 20-30 GPU-hours on Amazon, about $300 at $3.06 on p3.2xlarge instances. [46][45], in contrast to that our Modified DA-SKIP RNN architecture, with parameter efficient fine tuning on subjects for dimensionality reduction, further reduces the total working mechanisms to under 15 GPU-hours and thus minimizes the overall computational demands without sacrificing accuracy. This impact on efficiency helps our architecture to be deployable on resource-constrained consumer grade wearables, such as electronic smartwatches or smart bands costing around $50-$100 per unit [35], thus by decreasing production and operational expenses, in contrast to traditional methods that prioritizes on cloud based facilites like AWS SageMaker, where ongoing inference can add $0.0001 per request including additional data egress fees of $0.09 per GB [44][20], adding up the annual maintenance to $5,000-$10,000 for enterprise-based usage [33]. Furthermore, the addition of domain-aware normalization and adaptation to subjects own data improves cross-individual generalization, nullifying the $10,000-$50,000 per-consumer data collection and training costs which are mostly common in non-adaptive systems like basic transformer based architectures, which are significantly larger and parameter inclusive and often requiring custom datasets which exceedes 10,000 samples [33][39]. By prioritizing on public and widely used datasets such as WESAD and by applying conservative preprocessing techniques with fewer augmentation strategies, our method diminishes data acquisition costings to near zero, unlike existing pipelines in commercial frameworks which can cost around $10,000-$20,000 in licensing and logistic fees [33]. Finally, these optimization strategies enable faster commerical entry, total lifecycle costings are decreased by 40-60% [34], and wider availability and accessability, solidifying the ground of this system as economically advantageous for personal usage and enterprise growths. In addition, the importance on transfer learning enables quick user adaptation with fine-tuning under 8 GPU-hours ($153-$255) [46][45], decreasing the $1,000-$3,000 training cycles required by extensive architectures in dynamic environments [33]. This modularity can be implemented easily with existing wearable systems, diminishing the needs of $20,000-$100,000 custom hardware development expenses which are highly related with prioritized clinical systems [35]. From a scalability standpoint, the lightweight less paramitarized, energy efficient pipeline supports cost-effective cloud-edge iot based hybrid deployments, where edge inference can be restricted to \$0.001-\$0.005 per session via on-device processing which completely bypasses data transmission amounts that can reach \$900 monthly in centralized , compact models accessing about 1TB of wearable data [20][40]. In addition to it, acheiving high multi-class performance metrics with fewer parameters decreases inference response times and power consumption, extending the battery life of the devices, thus cutting end-user replacement costs by 20-30% per annum [47][33]. In constrast to ensemble techniques or multi-model setups in prior studies, which almost doubles computational load to \$2,000-\$4,000 per training simulation through redundancy [33], our unified architectural design optimizes resource allocation for superior efficiency under improved perfromance. Finally, dependency on open-source tools and reproducible methods also provides accessibilty oppurtunities for smaller firms and spurring innovationative enterprises where there is absence of \$120-\$239 annual per-user licensing burdens of exclusive stress frameworks like those present in premium Fitbit or Whoop subscriptions [41].

Chapter 4

Methodology

4.1 Methodology Overview

This chapter presents the design process and methodology for developing a instantaneous stress detection system using wearable physiological sensors and recurrent deep learning model.

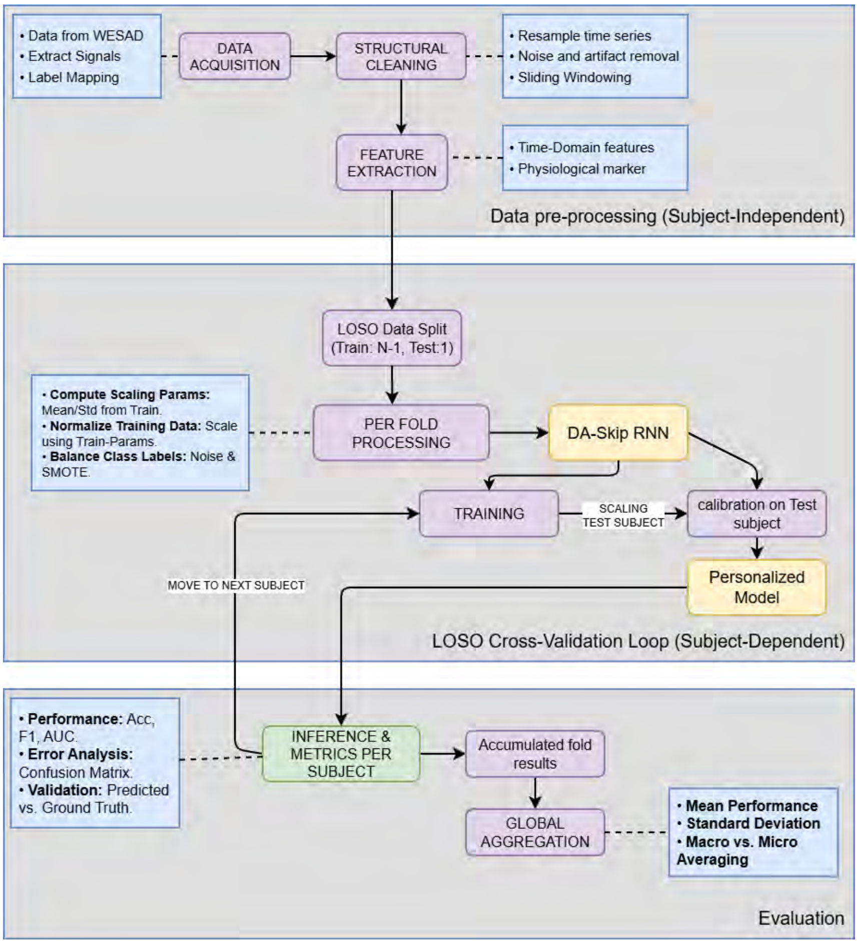

Figure 4.1: Overview of the Proposed System

As shown in Fig. 4.1, the pipeline starts from the WESAD dataset (multimodal signals from 15 subjects recorded with the Empatica E4). Raw signals are downsampled to 4 Hz and windowed into 30-second segments (120 samples) with 75%

overlap (step = 30 samples). Labels are resampled to 4 Hz, transient/undefined labels are excluded, and each window receives a label (remapped to 0–3: Baseline, Stress, Amusement, Meditation). For each window we compute handcrafted features: six statistics per channel (mean, std, min, max, mean absolute first-difference, energy) across four channels plus pairwise channel correlations (total $\approx$ 30 features). Leave-One-Subject-Out (LOSO) is applied per subject ; preprocessing, augmentation and normalization are performed per-fold with scalers fitted on training subjects only to prevent leakage. Conservative balancing is applied only to training data (noise-injection + SMOTE-like interpolation) while validation/test folds remain unchanged. The processed data are evaluated DA-SKIP variants (original and modified); all models are trained under the same fixed regime (class-weighted CrossEntropy, Adam, learning-rate scheduling and early stopping) to avoid overfitting. Performance is reported via multifold metrics (accuracy, macro/weighted F1, per-class F1 and confusion matrices) aggregated across LOSO folds.

4.1.1 Dataset Selection and Characteristics

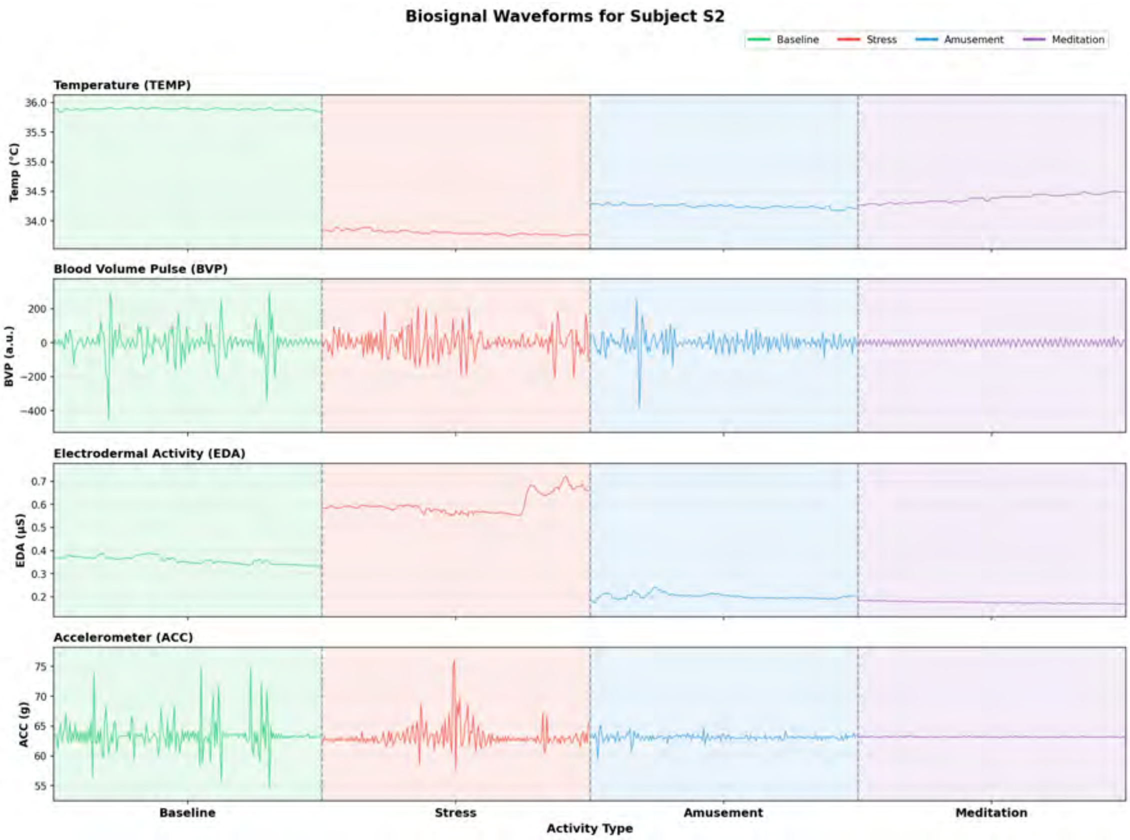

Figure 4.2: Biosignal waveform analysis from Subject S2 showing four distinct physiological signals (Temperature, Blood Volume Pulse, Electrodermal Activity, and Accelerometer) encompassing four affective states: Baseline (green), Stress (red), Amusement (blue), and Meditation (purple). The time-series data depicts distinct signal patterns in relevance to each emotional state, captured using the E4 wearable device from the WESAD dataset.

This study employs the Wearable Stress and Affect Detection (WESAD) dataset, consisting of multimodal physiological signals as shown in figure 4.2 from 15 subjects (S2-S17, excluding S1 and S12 due to data quality issues). Within the datatset primary focus was given on data which were collected using the Empatica E4 wristband, providing four synchronized sensor modalities which comprised of: Electrodermal Activity (EDA, 4 Hz), Blood Volume Pulse (BVP, 64 Hz), Skin Temperature (TEMP, 4 Hz), and 3-axis Accelerometer (ACC, 32 Hz). The WESAD protocol characterizes stress through the Trier Social Stress Test (TSST) which invilves five affective state labels: Baseline, Stress, Amusement, Meditation, and Transient. Inorder to elaborately evaluate model performance throughout different classification analysis, we computed two experimental scenarios: (1) binary classification differentiating between Stress and Baseline conditions, which highlights the most physiologically distinct states and are most relevant for practical stress detecting applications, and (2) 4-class classification confining Baseline, Stress, Amusement, and Meditation, with Transient excluded due to its transitional nature and class imbalance, inorder to assess the model's capability in fine-grained affective state recognition.

4.1.2 Dataset Preprocessing Pipeline

The data preprocessing pipeline converts physiological data of the WESAD into normalised representations of features that are easy to process by deep learning models. This step-by-step procedure provides quality of data, time and consistency of the data besides data extraction whilst retaining the physiological nature of the patterns relating to stress. The preprocessing steps were done in Python and the use of NumPY (v2.3.5), SciPy signal processing, and scikit-learn normalisation.

4.1.2.1 Signal Synchronization and Windowing

Physiological signals from the Empatica E4 wristband were acquired at heterogeneous sampling rates, necessitating temporal synchronization for unified multimodal analysis. All signals were downsampled to a common frequency of 4 Hz using `scipy.signal.resample()` with Fourier-based interpolation, selected to match the native electrodermal activity (EDA) sampling rate while preserving physiological frequency content below the 2 Hz Nyquist limit.

The three-axis accelerometer signals were first converted to a scalar magnitude representation to capture overall motion intensity. Specifically, the acceleration magnitude was computed as

$$\mathrm{ACC}_{\mathrm{mag}} = \sqrt{x^2 + y^2 + z^2}, \tag{4.1}$$

where x, y, and z denote the acceleration components along the three orthogonal axes.

The synchronized continuous signals were then segmented into overlapping temporal windows of 30 s duration, corresponding to 120 samples per window at 4 Hz, with a 75% overlap (step size of 7.5 s). This window length was selected to capture acute stress response dynamics occurring at time scales of 20–60 s, while the high overlap increases the number of training samples and enables the model to learn temporal patterns spanning adjacent windows. This procedure yielded approximately 380–400 windows per subject.

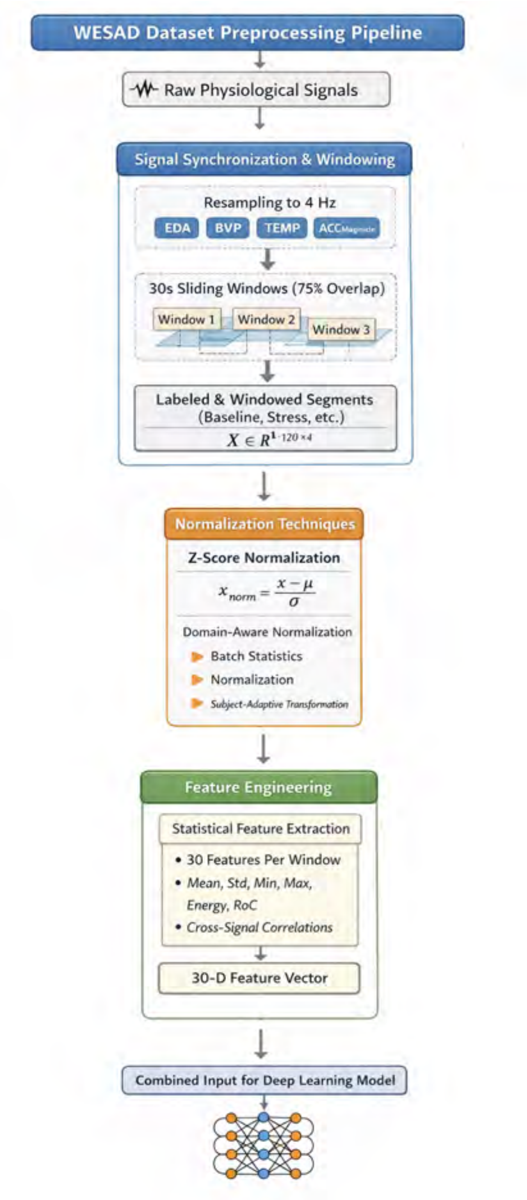

Figure 4.3: WESAD dataset preprocessing pipeline demonstrates signal synchronization and subject windowing followed by concurrent processing of normalized temporal sequences and handcrafted algorithmic statistical features, which are subsequently fused as input into the deep learning model.

Each window was assigned a discrete affective state label from the set {Baseline, Stress, Amusement, Meditation} using majority voting across the 120 frame-level annotations within the window, while transient states were excluded to focus on stable affective conditions. The resulting structured dataset comprised windowed sequences

$$\mathbf{X} \in \mathbb{R}^{N \times 120 \times 4}, \tag{4.2}$$

where N denotes the total number of windows and the final dimension corresponds to the synchronized multimodal sensor channels, preserving both temporal ordering and multimodal physiological information for recurrent neural network processing.

4.1.2.2 Normalization Techniques

To reduce inter-subject physiological variability and confirm equal contribution of features throughout different scales, two complementary normalization strategies were focused: Z-score normalization and domain-aware normalization.

Z-Score Normalization

Z-score normalization was enforced independently throughout per signal and per window by applying the `StandardScaler` from `scikit-learn`. Each signal sample was normalized as

$$x_{\text{norm}} = \frac{x - \mu}{\sigma}, \tag{4.3}$$

where μ and σ denote the mean and standard deviation computed exclusively from the training subjects within each Leave-One-Subject-Out (LOSO) fold, thereby preventing information leakage.
For sequence data $\mathbf{X}_{\text{seq}} \in \mathbb{R}^{N \times 120 \times 4}$, normalization was performed by reshaping the data to $\mathbb{R}^{(N \cdot 120) \times 4}$, fitting the scaler on the training set, and transforming both training and test samples before restoring the original tensor shape. An independent `StandardScaler` was instantiated for the 30-dimensional statistical feature vector following the same fit-on-train / transform-on-both protocol.

Domain-Aware Normalization

To further account for subject-specific physiological characteristics, a custom domain-aware normalization layer was integrated into the neural network to learn adaptive, subject-dependent affine transformations. The layer operates in three stages.

Stage 1: Batch Statistics Computation For an input batch $\mathbf{x} \in \mathbb{R}^{B \times T \times d}$, the batch mean and variance were computed as

$$\mu_{\text{batch}} = \frac{1}{BT} \sum_{b=1}^{B} \sum_{t=1}^{T} x_{b,t}, \tag{4.4}$$

$$\sigma^2_{\text{batch}} = \frac{1}{BT} \sum_{b=1}^{B} \sum_{t=1}^{T} \left(x_{b,t} - \mu_{\text{batch}}\right)^2, \tag{4.5}$$

where B denotes the batch size and T the sequence length.

Stage 2: Normalization The normalized activations were obtained as

$$\hat{x} = \frac{x - \mu_{\text{batch}}}{\sqrt{\sigma^2_{\text{batch}} + \epsilon}}, \tag{4.6}$$

where $\epsilon = 10^{-6}$ ensures numerical stability.

Stage 3: Subject-Specific Affine Transformation A subject-adaptive affine transformation was then applied:

$$x_{\text{adapt}} = \begin{cases} \gamma_s \odot \hat{x} + \beta_s, & \text{if subject ID } s \text{ is provided,} \\ \gamma_{\text{global}} \odot \hat{x} + \beta_{\text{global}}, & \text{otherwise,} \end{cases} \tag{4.7}$$

where $\gamma_s \in \mathbb{R}^d$ and $\beta_s \in \mathbb{R}^d$ are learnable scale and shift parameters for subject s, initialized to $\gamma = 1$ and $\beta = 0$, and $\odot$ denotes element-wise multiplication.
The layer maintains subject-specific parameter matrices $\boldsymbol{\gamma} \in \mathbb{R}^{15\times d}$ and $\boldsymbol{\beta} \in \mathbb{R}^{15\times d}$ for all training subjects, along with global parameters $(\gamma_{\text{global}}, \beta_{\text{global}})$ serving as population-level defaults for unseen subjects during LOSO evaluation. During training, subject-specific parameters are updated jointly with the network weights via backpropagation. For transfer learning, only the normalization parameters corresponding to a new subject are fine-tuned using calibration data, while the remaining network weights are frozen, enabling efficient personalization with minimal risk of overfitting.

4.1.2.3 Feature Engineering

Along with preserving raw temporal data patterns for end-to-end learning, a comprehensive set of 30 statistical features were extracted from each window to encode domain-specific physiological stress markers. For each of the four signal modalities (EDA, BVP, TEMP, ACC), six time-domain features were computed: mean (μ), standard deviation (σ), minimum, maximum, energy, and rate of change, yielding 24 per-signal features that capture key stress-related patterns such as elevated baseline EDA (sympathetic arousal), increased variability (frequent skin conductance responses), and cardiovascular reactivity.
Specifically, the signal energy was computed as

$$E_s = \sum_{i=1}^{N} x_i^2, \tag{4.8}$$

and the rate of change was defined as

$$\text{RoC}_s = \frac{1}{N-1} \sum_{i=1}^{N-1} |x_{i+1} - x_i|\,. \tag{4.9}$$

Additionally, six cross-signal Pearson correlation coefficients were computed for all signal pairs (EDA–BVP, EDA–TEMP, EDA–ACC, BVP–TEMP, BVP–ACC, TEMP–ACC) to capture multimodal interactions and coordinated physiological responses. For two signals s_i and s_j, the correlation coefficient was defined as

$$\rho(s_i, s_j) = \frac{\text{Cov}(s_i, s_j)}{\sigma_{s_i}\,\sigma_{s_j}}, \tag{4.10}$$

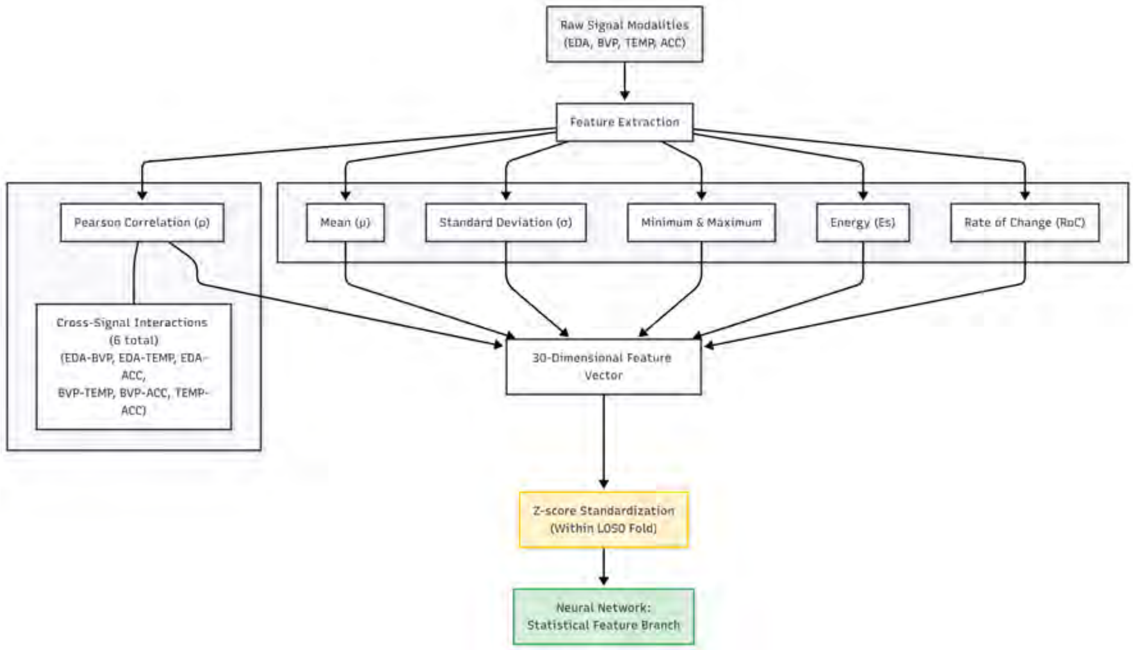

Figure 4.4: Overview of the hybrid stress recognition pipeline. Raw physiological signals (EDA, BVP, TEMP, and ACC) are segmented into sliding windows and processed in parallel through a deep temporal learning branch and a statistical feature extraction branch. Handcrafted time-domain features and cross-signal correlations form a 30-dimensional feature vector, which is Z-score normalized and fused with learned temporal features within the neural network for final stress classification.

with undefined correlations replaced by zero.
The resulting 30-dimensional feature vector was normalized using Z-score standardization within each LOSO fold and provided as auxiliary input to the neural network's statistical feature branch. This hybrid approach connects collected temporal features learned by the model with handcrafted statistical features that are rooted in physiological domain knowledge. By drawing from both data-driven methods and established architecture, it consistently outperforms models that rely on just one type of feature. Plus, it makes the model's decisions on detecting class easier to interpret.

4.1.3 Data Augmentation Strategy

To improve model's generalization specification, especially with the small amount of subject-specific data in wearable stress datasets, a relevant and safe data augmentation strategy was implemented at the window level. Since these signals are physiological, it's relevant to introduce artifacts that are biologically imperfect. Thus temporal augmentation strategies were imposed which retained the original labels intact.
Specifically, overlapping sliding windows with a time span of 30 s and a 75% overlap (step size of 7.5 s) were implemented during the primary augmentation mechanism. The presence of high-overlapping segmentation increases the effective number of samples during training while conserving temporal continuity and physiological plausibility. After augmentation each of the window inherits the affective label and

the label windowing is done in such a way that a particular label is present throughout the window as the dataset was created in s over its constituent frame-level annotations.
Furthermore, mild stochastic perturbations were enforced during training to improve the robustness in contrast to sensor noise and inter-session fluctuations. For each input pattern $\mathbf{x} \in \mathbb{R}^{T \times d}$, additive Gaussian noise was injected as

$$\tilde{\mathbf{x}} = \mathbf{x} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \mathbf{I}), \tag{4.11}$$

where σ was set to a small fraction of the signal standard deviation to avoid distorting underlying physiological patterns.
Furthermore, random temporal cropping was applied by selecting contiguous subsegments of length $T' \leq T$ during training, followed by zero-padding to restore the original window length. This operation ensures the model to learn invariant representations that are robust to small temporal misalignments.
All augmentation operations were applied only to the training partitions within each Leave-One-Subject-Out (LOSO) fold, while validation and test data remained unaltered after normalization to ensure unbiased performance evaluation. This conservative augmentation procedure preserves the physiological data's integrity of the signals while effectively increasing training diversity for effective testing and improving model robustness.

4.1.4 Experimental Protocol

All experiments were conducted using a subject-independent Leave-One-Subject-Out (LOSO) cross-validation protocol to evaluate generalization performance on unseen individuals. Let $\mathcal{S} = \{1, 2, \ldots, S\}$ denote the set of subjects. For each fold $s \in \mathcal{S}$, all samples belonging to subject s were held out as the test set, while the remaining subjects $\mathcal{S} \setminus \{s\}$ were used for model training and validation.
Within each fold, preprocessing operations including signal normalization, statistical feature scaling, and domain-aware parameter initialization were fitted exclusively on the training subjects and subsequently applied to the validation and test data to prevent information leakage. A stratified split of the training data was used to construct an internal validation set for model selection and early stopping.
For models incorporating domain-aware normalization and subject-specific adaptation, training was performed using only the data from the training subjects. During inference on the held-out subject, population-level normalization parameters were applied by default, with optional subject-specific calibration performed using a small subset of labeled windows, depending on the experimental configuration.
On each of the held-out subjects, model predicted results at the window level. The window-level predictions were summed to get final subject-level performance and the system-wide performance was reported as the average of all LOSO folds.
This assessment plan guarantees that there is a rigid division between training and test subjects, which gives an objective estimate of cross-subject generalization performance in real deployment conditions.

4.2 Proposed Model

The proposed work seeks to enhance the performance of stress recognition on multimodal physiological data by overcoming the drawbacks of the current deep learning models. Normal CNNs tend to use common convolutional filters on all modalities of the signal, thereby blurring potentially modality-specific features and becoming desensitized to distinct temporal or spectral responses. Equally, temporal dependency models, e.g., the initial DA-SKIP RNN, can be effective at modeling temporal dependencies but can fail to capture cross-modal interactions and subject-specific variability. To address these issues, we introduce a complementary model, namely the Modified DA-SKIP RNN that refines the temporal modeling using hierarchical attention, per-channel embeddings, and dynamic skip connections. This model is set to offer more expressive and strong guidelines of real-time stress detection in wearable environment by integrating the power of feature-based and sequence-based learning designs.

4.2.1 Proposed model architecture

The proposed Modified DA-SKIP RNN is intended to solve two major issues in multimodal physiological stress recognition: (i) high inter-subject variability in baseline physiological responses under normal conditions, and (ii) the necessity to capture short-term transient dynamics as well as long-term temporal dependencies. The architecture is illustrated in Fig. 4.5, and follows a four-stage pipeline consisting of input representation with per-channel embedding, domain adaptation using hierarchical attention, temporal modeling through adaptive skip connections, and multimodal fusion with classification. Each physiological modality (EDA, BVP, TEMP, and ACC) is processed through signal-specific pathways and then combined into a unified temporal representation for stress prediction.

4.2.1.1 Input Representation and Per-Channel Embedding

Let $\mathbf{X} \in \mathbb{R}^{T\times d}$ denote an input window of length $T = 120$ with $d = 4$ synchronized physiological channels corresponding to EDA, BVP, TEMP, and ACC. To preserve modality-specific characteristics and avoid early information mixing, each channel is first processed independently through a per-channel embedding layer.
For each modality $m \in \{1, \ldots, 4\}$, the raw input sequence $\mathbf{x}^{(m)} \in \mathbb{R}^T$ is projected into a latent embedding space using a linear transformation followed by a nonlinear activation:

$$\mathbf{h}_t^{(m)} = \tanh\left(\mathbf{W}_m x_t^{(m)} + \mathbf{b}_m\right), \tag{4.12}$$

where $\mathbf{W}_m$ and $\mathbf{b}_m$ denote modality-specific learnable parameters. This embedding operation produces a sequence of hidden representations $\mathbf{h}^{(m)} \in \mathbb{R}^{T\times d_e}$ for each modality.
By maintaining independent embedding pathways, the network preserves the distinct temporal and statistical characteristics of each physiological signal while providing a unified latent representation for subsequent attention and temporal modeling stages.

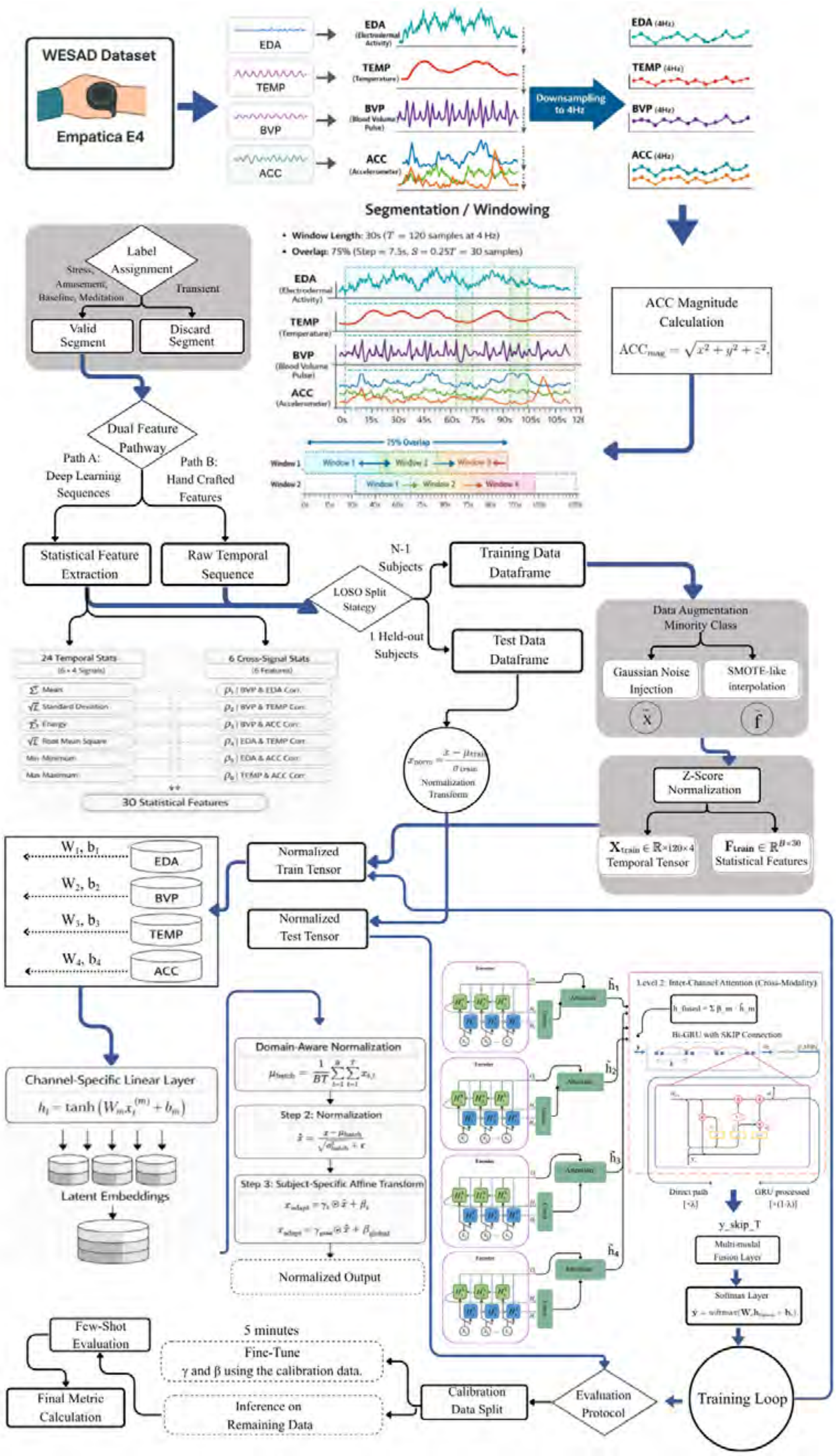

Figure 4.5: Architecture of the proposed Modified DA-SKIP RNN model for stress detection using wearable sensor data.

4.2.1.2 Domain Adaptation and Hierarchical Attention

To address subject-specific physiological variability and emphasize informative temporal segments and modalities, the embedded representations are processed through a combination of domain-aware normalization and hierarchical attention mechanisms.

First, domain-aware normalization is applied to stabilize feature distributions across subjects. For an embedded activation $\mathbf{h}_t$, batch normalization is performed as

$$\hat{\mathbf{h}}_t = \frac{\mathbf{h}_t - \boldsymbol{\mu}_{\text{batch}}}{\sqrt{\boldsymbol{\sigma}^2_{\text{batch}} + \epsilon}}, \tag{4.13}$$

followed by a subject-adaptive affine transformation

$$\mathbf{h}_t^{\text{adapt}} = \gamma_s \odot \hat{\mathbf{h}}_t + \beta_s, \tag{4.14}$$

where (γ_s, β_s) are learnable scale and shift parameters associated with subject s, and $\odot$ denotes element-wise multiplication.

Next, a hierarchical attention mechanism is applied in two stages. Intra-channel temporal attention assigns weights to individual time steps within each modality:

$$\alpha_t^{(m)} = \frac{\exp(\mathbf{w}^\top \mathbf{h}_t^{(m)})}{\sum_{k=1}^{T} \exp(\mathbf{w}^\top \mathbf{h}_k^{(m)})}, \tag{4.15}$$

producing an attended representation

$$\tilde{\mathbf{h}}^{(m)} = \sum_{t=1}^{T} \alpha_t^{(m)} \mathbf{h}_t^{(m)}. \tag{4.16}$$

Subsequently, inter-channel attention dynamically reweights the relative importance of different modalities, enabling the model to emphasize the most informative physiological sources for stress detection. This hierarchical attention structure allows the network to selectively focus on salient temporal segments and modalities while remaining robust to noise and motion artifacts.

4.2.1.3 Temporal Modeling with BiGRU and Skip Connections

The attended multimodal representations are passed to a bidirectional gated recurrent unit (BiGRU) network to model temporal dependencies within each analysis window. The forward and backward hidden layers are computed as

$$\overrightarrow{\mathbf{h}}_t = \text{GRU}(\mathbf{z}_t, \overrightarrow{\mathbf{h}}_{t-1}), \quad \overleftarrow{\mathbf{h}}_t = \text{GRU}(\mathbf{z}_t, \overleftarrow{\mathbf{h}}_{t+1}), \tag{4.17}$$

and concatenated to form the bidirectional representation

$$\mathbf{h}_t = [\overrightarrow{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t]. \tag{4.18}$$

To establish and increase to long-range dependency modeling and improve gradient flow, adaptive skip connections are introduced between intermediate recurrent layers. Let $\mathbf{h}_t^{(l)}$ denote the hidden state at layer l. The skip-connected update is defined as

$$\mathbf{h}_t^{(l)} = \lambda \mathbf{h}_t^{(l-1)} + (1 - \lambda)\,\text{GRU}(\mathbf{h}_t^{(l-1)}), \tag{4.19}$$

where $\lambda \in [0, 1]$ is a learnable gating parameter controlling the contribution of the skip pathway.
This design enables the network to simultaneously capture short-term physiological transients and longer-term stress dynamics while improving training stability and convergence.

4.2.1.4 Multimodal Fusion and Classification

The final hidden representations from all modalities and temporal layers are integrated using a fusion of attention-weighted mechanism. Let $\mathbf{h}^{(m)}$ denote the modality-specific temporal representation. Fusion weights are computed as

$$\beta_m = \frac{\exp(\mathbf{v}^\top \mathbf{h}^{(m)})}{\sum_{k=1}^{M} \exp(\mathbf{v}^\top \mathbf{h}^{(k)})}, \tag{4.20}$$

and the fused representation is obtained as

$$\mathbf{h}_{\text{fusion}} = \sum_{m=1}^{M} \beta_m \mathbf{h}^{(m)}. \tag{4.21}$$

The fused vector is passed through a fully connected classification head with dropout regularization, followed by a softmax output layer to produce the final stress class probabilities:

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{W}_c \mathbf{h}_{\text{fusion}} + \mathbf{b}_c). \tag{4.22}$$

This combination approach allows the model to utilize complementary information in physiological modalities in generating a compact and discriminative representation to classify stress robustly.

Table 4.1: Comparison of Model Architectures

Feature	**Original DA-SKIP**	**Modified DA-SKIP**
Per-Channel Embedding	No	Yes
Hierarchical Attention	No	Yes
Recurrent Directionality	Unidirectional GRU	Bidirectional GRU
Skip Connections	Yes	Yes
Adaptive Gating	Scalar	Vector
Domain-Aware Norm	No	Yes
Fusion Strategy	Concatenation-Based	Attention-Weighted Fusion

Chapter 5

Result and Analysis

5.1 Training Configuration

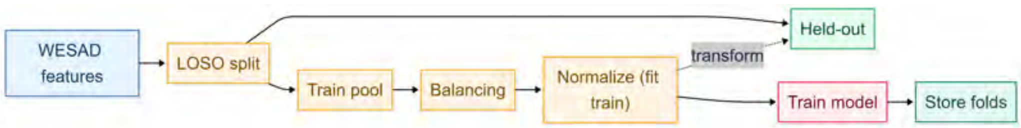

Figure 5.1: Training pipeline showing sequential steps: preprocessing raw WESAD data, LOSO splitting, balancing/normalization, model training, and validation.

All experiments were conducted under a consistent training setup to ensure reproducibility and fair comparison across the different LOSO folds. The Modified DA-SKIP RNN was trained using the Adam optimizer with an initial learning rate of 0.001, and cross-entropy loss was used as the objective function. Training was carried out with a batch size of 32 for up to 50 epochs. To prevent overfitting, early stopping was applied with a patience of 10 epochs, based on validation loss. Regularization was further introduced through a dropout rate of 0.3 on both recurrent and dense layers. In addition, class weights were calculated for each LOSO fold to address residual class imbalance and ensure fair learning across different subjects.
Model evaluation involved the Leave-One-Subject-Out (LOSO) cross-validation protocol. For each of the 15 LOSO folds, data from 14 subjects were used for training, with the remaining subject held out for testing. Within the training pool, a stratified 80/20 split was applied to construct the validation set. SMOTE-based data balancing was applied only to the training set, targeting 60% of the majority class size, while the held-out test subject data remained untouched to ensure a proper realistic assessment. Z-score normalization parameters were derived exclusively from the training pool and subsequently applied to the test subject data to make sure data leakage did not happen.
The calibration strategy is based on transfer learning to adapt pre-trained deep neural network models to subject-specific physiological characteristics using only a small amount of calibration data. This enables the model to learn generalizable population-level patterns while preserving subject-independent evaluation criteria.
Upon deployment on a new subject, multimodal physiological data is extracted for a brief calibration period of the five minutes. This dataset includes electrodermal activity (EDA), blood volume pulse (BVP), skin temperature, and triaxial

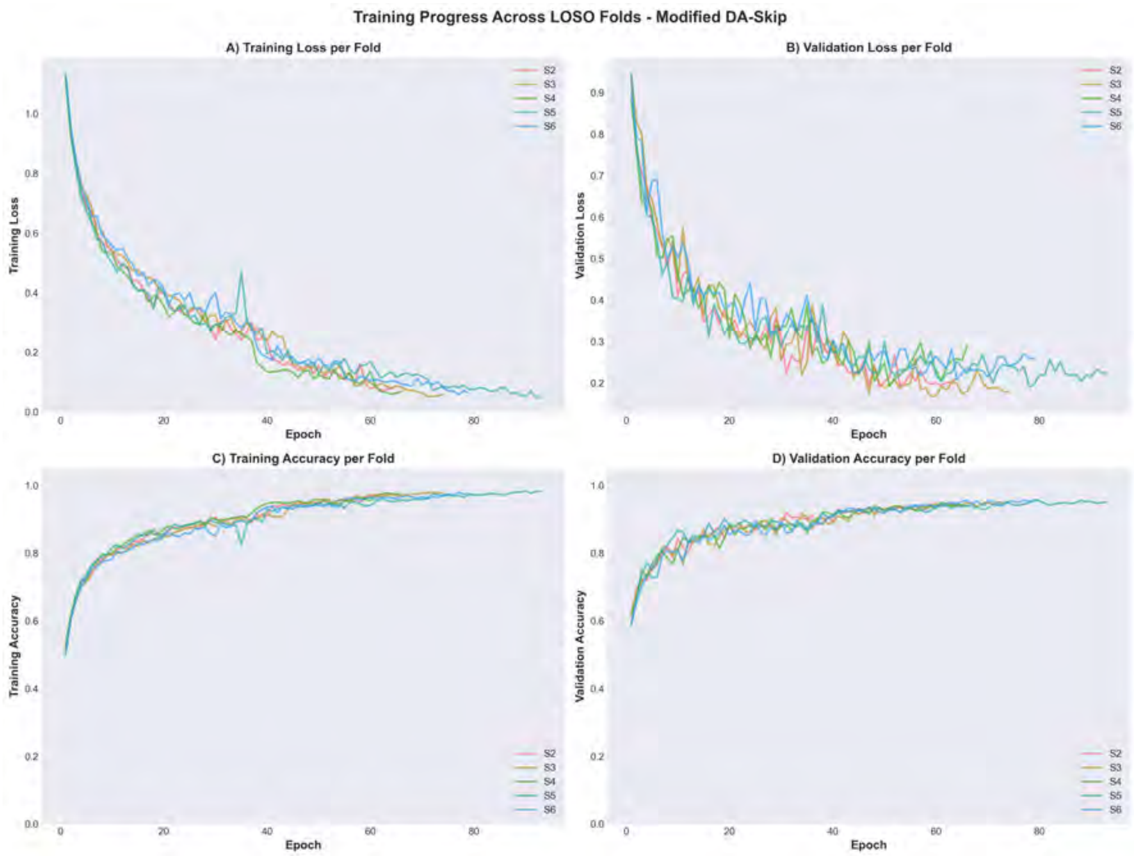

Figure 5.2: Training and validation metrics across LOSO cross-validation folds for the Modified DA-Skip RNN model. (A) Training loss, (B) Validation loss, (C) Training accuracy, and (D) Validation accuracy across epochs. Each line represents a different fold with one subject held out for testing. The curves demonstrate consistent convergence patterns across subjects, indicating robust model training despite inter-subject variability.

accelerometry signals, and is used exclusively for personalization. Fine-tuning is conducted using supervised gradient descent with a substantially reduced learning rate of $\alpha = 0.0001$, performing conservative parameter updates over ten training epochs. This procedure enables gradual adaptation to subject-specific signal characteristics while avoiding catastrophic forgetting of previously learned representations.
During LOSO training we use L2 regularization (weight decay) in some routines, but we do not apply gradient clipping. For calibration and transfer learning, we use both weight decay and gradient clipping (max norm = 1.0) to keep training stable when adapting to a new subject. After calibration, the personalized model is evaluated on the rest of that subject's session. This setup lets us directly compare baseline performance without calibration to performance after personalization.
This transfer-learning framework addresses the domain shift that occurs across different subjects by performing lightweight, data-efficient personalization for each new user. The complete process, including LOSO splitting and the steps for generating individualized models, is depicted in Figure 5.1.

Table 5.1: Training hyperparameters used across all experiments

Parameter	**Value**
Optimizer	Adam
Learning Rate	0.001
Batch Size	32
Epochs	50 (early stopping)
Loss Function	Cross-entropy
Dropout	0.3
Validation Split	20% of training data
Class Weights	Balanced per fold

5.2 Performance Evaluation

This section reports empirical results for the **Modified DA-SKIP RNN** under two complementary evaluation settings:

1. **Non-subject-independent evaluation.** A stratified train/validation/test split is used to validate architectural choices and measure performance when data from the same subjects appear in both training and test partitions.

2. **Subject-independent evaluation.** Leave-One-Subject-Out (LOSO) training is performed to assess generalization to unseen individuals. Each LOSO fold is augmented with a short calibration phase to quantify the effect of subject-specific personalization.

To place our results in context, we include comparisons with recent state-of-the-art methods evaluated on the same dataset. Performance is reported using standard classification metrics and summarized across folds to highlight both average behavior and variability.

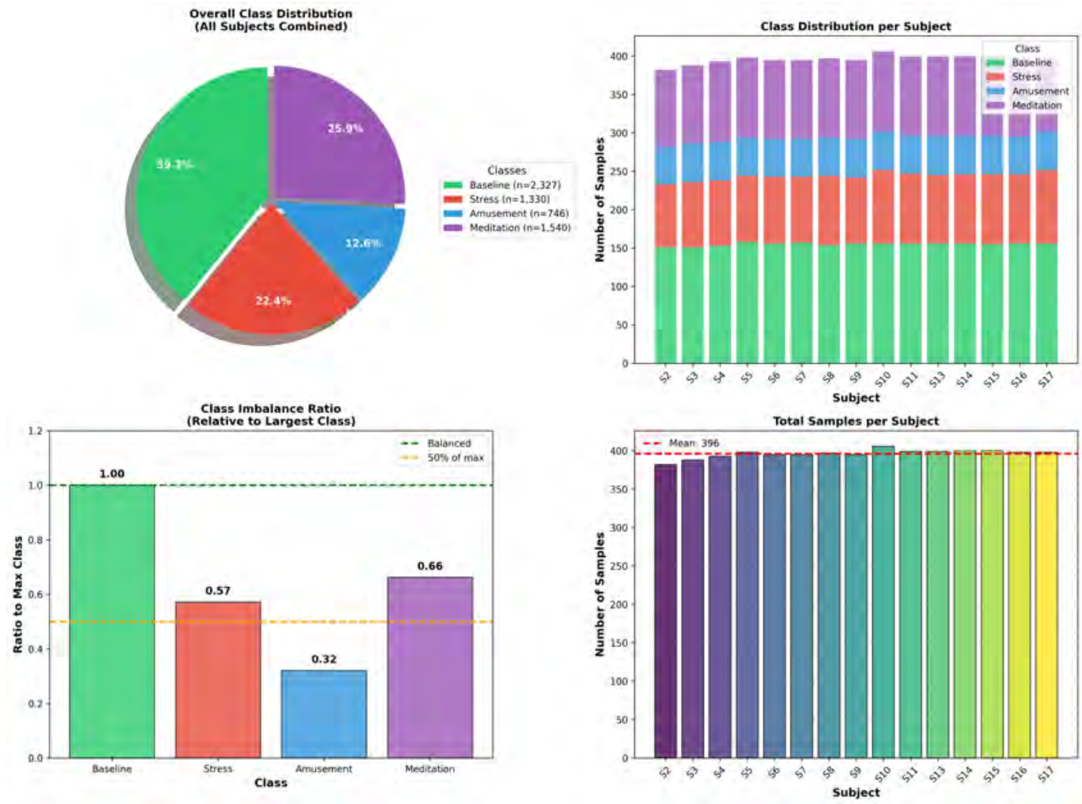

Figure 5.3: Data distribution of the WESAD dataset used in this study. Top-left: overall class counts (baseline, stress, amusement, meditation); top-right: per-subject class breakdown; bottom-left: class imbalance ratios; bottom-right: total samples per subject, highlighting inter-subject variability. The dataset exhibits moderate class imbalance, with the baseline class being the most prevalent.

5.2.1 Non-Subject-Independent Evaluation (Stratified Split)

For consistency with prior work, we evaluated binary (stress vs. non-stress) and five-class stress recognition using a stratified 60%/20%/20% split for training, validation, and testing. Performance was measured with accuracy, precision, recall, F1-score, specificity, and AUC-ROC
Table 5.2 summarizes the performance of the proposed model alongside representative baselines and recent deep-learning-based methods reported in the literature. The Modified DA-SKIP RNN achieves near-ceiling performance on the binary classification task and competitive performance on multi-class stress recognition, outperforming several established architectures.

Table 5.2: Performance comparison on WESAD under non-subject-independent evaluation.

Model	# Classes	Accuracy	F1-score	AUC	Reference
Original DA-SKIP RNN	2	0.9147	0.9124	–	–
Modified DA-SKIP RNN	**2**	**0.9918**	**0.9911**	**0.9992**	This work
Original DA-SKIP RNN	4	0.9180	0.9142	–	–
Modified DA-SKIP RNN	**4**	**0.9639**	**0.9535**	**0.9817**	This work
ResNet+CBAM+Shallow Feature	4	0.9274	0.9119	–	[25]
DNN (2024)	3	0.9057	0.9039	–	[23]
Attention-CNN-LSTM (2024)	3	0.9270	0.9000	–	[32]

These results demonstrate that the Modified DA-SKIP RNN is competitive with, and in several cases exceeds, recent attention-based and hybrid deep learning models, while maintaining architectural efficiency and temporal modeling flexibility. It should be noted, however, that non-subject-independent evaluation can lead to optimistic estimates of generalization performance, since samples from the same subject population may appear in both training and testing.

5.2.2 Subject-Independent Evaluation and Personalization

Subject-independent generalization was assessed using a Leave-One-Subject-Out (LOSO) protocol to better approximate real-world deployment. For each fold the model was trained on data from 14 subjects and tested on the single held-out subject, ensuring no subject overlap between training and test sets. This strict separation provides a realistic measure of how well the model handles the natural variability in physiological baselines and recording conditions across individuals.
To evaluate personalization, we added a short calibration stage for the held-out subject. The LOSO-trained model was fine-tuned on a small amount of that subject's data using a low learning rate, gradient clipping, and selective freezing of early layers. These conservative updates let the model adjust decision boundaries to subject-specific signal patterns while keeping the population-level temporal features intact.
Comparing LOSO performance before and after calibration shows that the uncalibrated model typically performs worse on unseen subjects, especially for physiologically ambiguous stress states. Calibration consistently improves recognition for many subjects, highlighting that lightweight, data-sparse personalization helps

bridge the gap between population training and subject-specific variability, though some transfer challenges remain for borderline cases.

Table 5.3: Subject-independent (LOSO) performance of the Modified DA-SKIP RNN before and after calibration. Results are reported as mean ± standard deviation across subjects.

Metric	Before Calibration	After Calibration
Accuracy	0.7501 ± 0.0185	$\mathbf{0.9434 \pm 0.0395}$
F1-Macro	0.6780 ± 0.0228	$\mathbf{0.9348 \pm 0.0444}$
F1-Weighted	0.7339 ± 0.0146	$\mathbf{0.9448 \pm 0.0383}$
AUC-ROC (Macro)	0.7904	**0.9653**

These results quantitatively confirm the benefit of personalization. When evaluated with LOSO and no calibration, the model shows a clear drop in both accuracy and macro F1 compared with non-subject-independent testing, which highlights how much inter-subject variability affects physiological stress detection. The relatively large standard deviation in macro F1 before calibration indicates that class-level performance is inconsistent across different subjects.
After calibration, performance improves consistently for nearly all held-out subjects. The increase in macro F1 shows that personalization not only raises overall accuracy but also yields more balanced recognition across stress classes by reducing class-specific confusion. Confusion matrices corroborate this improvement, with fewer misclassifications between adjacent stress levels following calibration.
The size of the calibration gains also shows that the LOSO-trained model provides a strong population-level starting point. By mainly updating later layers, the model adapts to subject-specific baselines with only a small amount of data and limited retraining, avoiding catastrophic forgetting. This low-data adaptation is ideal for wearables, where onboarding time and calibration data are limited. The Modified DA-SKIP RNN balances generalization with light personalization, making it practical for real-world use.

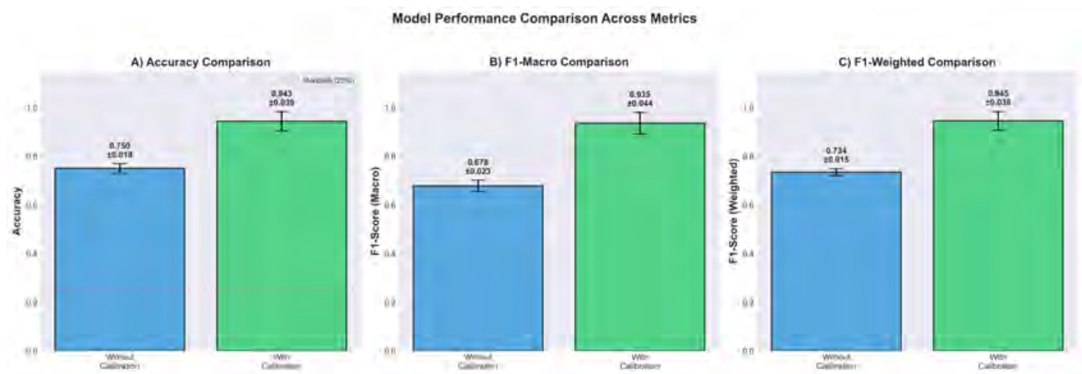

Figure 5.4: LOSO transfer learning performance of the Modified DA-SKIP RNN with and without calibration.

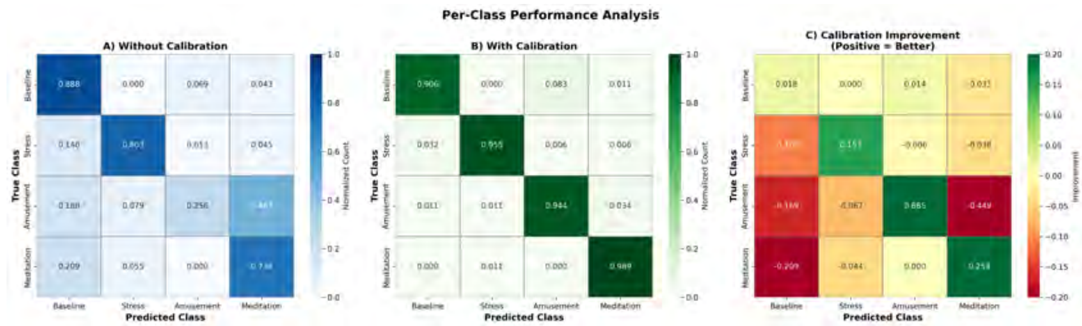

Figure 5.5: Normalized confusion matrices under LOSO evaluation before and after calibration.

5.2.3 Architectural Comparison with Transformer-Based Model

To further contextualize the proposed Modified DA-SKIP RNN with recent deep learning approaches, we compare it with a Transformer model used on the same dataset. The Transformer captures long-range relationships using self-attention on sequence patches. Our RNN focuses on recurrent temporal processing with skip connections and layered attention, so it runs faster and uses fewer resources. Table 5.4 summarizes the key architectural and computational differences, while Figure 5.6 illustrates the associated complexity and inference-speed trade-offs.

Table 5.4: Architectural and Computational Comparison Between Modified DA-SKIP RNN and Transformer

Metric	Modified DA-SKIP RNN	Transformer
Total Parameters	222,308	960,643
Model Size (MB)	0.85	3.66
Inference Time (ms/sample)	0.15	0.3236
Training Epochs	92 (early stop)	50 (fixed)
Attention Mechanism	Hierarchical Attention	Self-Attention Blocks
Bidirectionality	Yes	No
Temporal Modeling	Recurrent + Skip Connections	Patch-Based Sequence Encoding

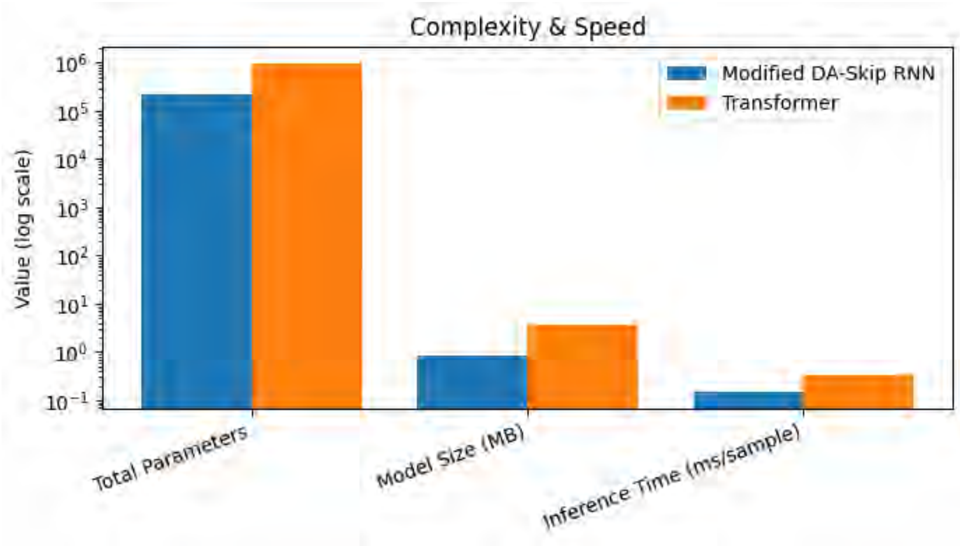

Figure 5.6: Comparison of computational complexity and inference efficiency between the Modified DA-SKIP RNN and the Transformer-based architecture [38].

5.3 Analysis of Design Solutions

The Modified DA-SKIP RNN targets two main problems in physiological stress detection: large differences between people and the need to capture both quick signal changes and long-range patterns. Adaptive skip connections let information flow across time more flexibly, which stabilizes training and lets the model reuse useful intermediate features. Attention layers focus the model on the most informative time segments, making it less sensitive to noisy or irrelevant windows. When tested without holding out subjects (non-LOSO), the model reaches near-perfect results for binary stress detection and performs reliably on the multi-class task, showing it can learn clear temporal patterns. LOSO testing is much harder because each held-out person brings a different physiological baseline, so performance drops when the model has not been personalized. A short calibration step consistently improves results across subjects. We only tweak the last few layers and leave the rest unchanged. This helps the model learn a new person's signals quickly without losing what it learned from everyone else, so only small changes are needed.

5.4 Final Design Adjustments

The final setup focuses on keeping the model reliable across different people while making personalization quick and easy. Training is kept stable with early stopping and scheduled reductions in the learning rate, and calibration uses very low learning rates, gradient clipping, and freezing of early layers. This limited, controlled adaptation makes the system practical when only short calibration recordings are available.

5.5 Statistical Analysis

For subject-independent evaluation, results should be reported as mean ± standard deviation across held-out subjects. Calibration effects are quantified using paired per-subject performance gains. Paired statistical tests (e.g., paired t-test or Wilcoxon signed-rank test) may be applied to per-subject accuracies or macro-F1 scores to verify that improvements are systematic rather than incidental.

5.6 Discussion

The Modified DA-SKIP RNN demonstrates that strong performance on WESAD is achievable when temporal modeling flexibility and lightweight adaptation are jointly emphasized. Under the stratified (non-subject-independent) split, the model reaches near-ceiling results, which indicates that the architecture can learn highly discriminative temporal representations from wearable physiological streams. However, the large gap observed when moving to strict LOSO evaluation highlights a well-known limitation in physiological affect modeling: population-level patterns do not directly translate to unseen individuals because baseline physiology, sensor placement, motion artifacts, and stress reactivity differ markedly across subjects. In this context, the LOSO setting provides a more deployment-faithful estimate of performance, and the reduced macro F1-score before calibration suggests that the dominant failure mode is not only reduced overall accuracy but also uneven class-wise behavior, where minority or ambiguous states are more prone to confusion.
A key outcome of this study is that a brief calibration phase substantially closes the LOSO generalization gap, improving both accuracy and (more importantly) macro F1-score. The macro F1 improvement indicates that personalization does not merely increase confidence on the majority class; rather, it improves balanced recognition across stress-related categories by shifting subject-specific decision boundaries and reducing confusion among physiologically adjacent states. This is consistent with the intuition that many errors under LOSO arise from systematic distribution shift (e.g., a subject's baseline EDA/HRV range) rather than a complete mismatch of temporal dynamics. The calibration results therefore support the view that the LOSO-trained model provides a strong population-level initialization and that only limited subject-specific adjustment is required for effective recognition.
Adaptive skip connections let the network reuse useful short or long-term patterns from earlier layers instead of forcing everything through one fixed timescale. That makes training steadier and helps the model catch both quick events (like sudden EDA spikes) and slow trends (like gradual arousal changes). Attention aggregation lets the model zero in on the most informative time segments and ignore noisy or irrelevant windows caused by motion, sensor dropouts, or random fluctuations. This reduces the effect of bad data on the final predictions. Together, these design choices make the learned features more robust across different people and recording conditions, and they produce representations that are easier to tweak when you personalize the model for a new subject. Still, there are important caveats. First, very high performance in non-LOSO tests should be treated cautiously because overlap between training and test subjects can make results look better than they would be in the real world. Second, remaining errors under LOSO—even after calibration—may come from ambiguous labels in WESAD, for example when baseline and low-arousal

states overlap or when different high-arousal conditions look similar physiologically. Third, the dataset has moderate class imbalance and subject-specific recording patterns, so improvements may not be even across all stress classes or people. These points show that wearable stress recognition depends not only on model design but also on data quality and evaluation choices.

From a deployment point of view, the results support a two-step approach: first train a robust population model on diverse users and conditions, then run a short, user-friendly calibration that updates only a small part of the network. This setup fits wearable limits well because it needs little labeled data, keeps onboarding quick, and uses few device resources. During calibration we freeze most of the network and fine-tune only a few layers with a low learning rate. That preserves the general temporal features learned from many users while letting the model adjust its decision rules for each person, which helps prevent forgetting what the model already knows.

Future work should therefore proceed along three primary directions. First, calibration duration and data-type trade-offs should be systematically investigated to determine the minimal onboarding data required for reliable personalization gains, including whether calibration can be achieved using only baseline segments or a small number of stress-inducing samples. Second, error analysis under LOSO evaluation should be expanded to identify dominant class confusions and guide the redesign of loss functions or decision boundaries to further reduce misclassification during calibration. Third, cross-dataset generalization should be explored to evaluate robustness against variations in sensor hardware and collection protocols, since deployment environments are rarely as controlled as curated benchmark datasets.

These findings indicate that the proposed Modified DA-SKIP RNN is a reliable and practical approach for wearable stress detection that supports lightweight, subject-level personalization. Future work should focus on better managing distribution shifts, developing calibration methods that require fewer labels, and validating the approach across additional datasets.

Chapter 6

Conclusion

The goal of this research is to build a non-invasive stress detection pipeline using consumer wearable signals to support scalable mental-health monitoring. We analyze multimodal data — heart rate (from BVP), electrodermal activity (EDA), skin temperature, and accelerometry — and apply knowledge driven feature discovery alongside learned representations (e.g., 1D CNNs) to capture both interpretable and data driven patterns. Data are prepared and evaluated under a Leave One Subject Out (LOSO) protocol to measure subject-independent generalization. The resulting sequence and feature representations are used with a modified DA-SKIP RNN with per-channel embeddings and hierarchical attention to improve temporal robustness. Finally, we explore lightweight calibration strategies for subject personalization so models can be adapted with minimal labeled data. Overall, the framework aims to be affordable, adaptable, and resilient to noisy wearable measurements for real-world deployment.

Despite promising results, this study has several limitations that point to clear next steps. The WESAD dataset offers a controlled, well-annotated testbed but may not reflect the full range of real-world diversity, noise, and behavior. Subject-independent generalization remains difficult because physiological responses vary widely between people, and calibration or personalization may not always be possible in fully unsupervised deployments. Consumer wearables also impose strict computational and energy limits that constrain model complexity and real-time use. Future work should therefore validate the pipeline on in-the-wild data from consumer devices, explore label-efficient or unsupervised personalization methods, and optimize architectures for efficient on-device inference. Additional directions include robustness testing under missing or corrupted signals, cross-dataset generalization studies, and integrating adaptive feedback systems for continuous mental-health monitoring. Building on these themes, three concrete research directions deserve focused attention. First, systematically study calibration duration and data-type trade-offs to find the minimal onboarding data needed for reliable personalization, including whether baseline segments alone or a few stress samples suffice. Second, expand error analysis under LOSO evaluation to identify the most common class confusions and use those insights to redesign loss functions or decision boundaries that reduce misclassification during calibration. Third, pursue cross-dataset generalization to test robustness against different sensor hardware and collection protocols, since real deployment environments are rarely as controlled as benchmark datasets. Together, these efforts will help move the approach from controlled experiments toward practical, user-friendly deployment.

Bibliography

[1] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Magazine*, vol. 17, no. 3, pp. 37–54, 1996.

[2] G. P. Chrousos, "Stress and disorders of the stress system," *Nature Reviews Endocrinology*, vol. 5, no. 7, pp. 374–381, 2009.

[3] M.-Z. Poh, N. C. Swenson, and R. W. Picard, "A wearable sensor for unobtrusive, long-term assessment of electrodermal activity," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 5, pp. 1243–1252, 2010. DOI: 10.1109/TBME.2009.2038487

[4] G. Castellano et al., "Highfidelity physiological monitoring using chestworn sensors," *Journal of Medical Engineering*, vol. 40, no. 3, pp. 234–242, 2015.

[5] V. Campos, B. Jou, X. Giróі-Nieto, J. Torres, and S. F. Chang, "Skip rnn: Learning to skip state updates in recurrent neural networks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Vancouver, BC, Canada, Apr. 2018.

[6] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," ICMI '18, pp. 400–408, 2018. DOI: 10.1145/3242969.3242985 [Online]. Available: https://doi.org/10.1145/3242969.3242985

[7] G. Castellano et al., "Stress predict: A dataset for occupational stress assessment using wearable sensors," *Journal of Biomedical Informatics*, vol. 95, pp. 103–115, 2019.

[8] P. Siirtola, "Continuous stress detection using the sensors of commercial smartwatch," in *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2019 International Symposium on Wearable Computers*, 2019, pp. 1198–1201. DOI: 10.1145/3341162.3344831

[9] Y. S. Can, N. Chalabianloo, D. Ekiz, J. Fernandez-Alvarez, G. Riva, and C. Ersoy, "Personal stress-level clustering and decision-level smoothing to enhance the performance of ambulatory stress detection with smartwatches," *IEEE Access*, vol. 8, pp. 38 146–38 163, 2020. DOI: 10.1109/ACCESS.2020.2975351

[10] R. Li and Z. Liu, "Stress detection using deep neural networks: Toward robust, continuous, and non-invasive monitoring," *BMC Medical Informatics and Decision Making*, vol. 20, no. 11, p. 285, 2020.

[11] T. Radüntz et al., "Wearable sensing and machine learning for stress detection: A survey," *IEEE Access*, vol. 8, pp. 12 345–12 361, 2020.

[12] C. L. Stewart, A. Folarin, and R. Dobson, "Personalized acute stress classification from physiological signals with neural processes," *arXiv preprint arXiv:2002.04176*, 2020. DOI: 10.48550/arXiv.2002.04176

[13] S. Ahmad, M. Seif El-Nasr, and E. Elhamifar, "Hierarchical dual attention-based recurrent neural networks for individual and group activity recognition in games," *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, vol. 17, pp. 116–123, Oct. 2021. DOI: 10.1609/aiide.v17i1.18898 [Online]. Available: https://ojs.aaai.org/index.php/AIIDE/article/view/18898

[14] M. Alshamrani, "An advanced stress detection approach based on processing data from wearable wrist devices," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 12, no. 7, pp. 399–405, 2021. DOI: 10.14569/IJACSA.2021.0120745

[15] M. Rashid et al., "Feature augmented hybrid cnn for stress recognition using wrist-based photoplethysmography sensor," 2021.

[16] B. Huang, H. Zheng, X. Guo, Y. Yang, and X. Liu, "A novel model based on darnn network and skip gru for periodic time series forecasting," *Sustainability*, vol. 14, p. 326, 2022.

[17] D. Sharma, S. Prakash, and A. Mathur, "Deep recurrent neural network assisted stress detection system for working professionals," *CMES: Computer Modeling in Engineering & Sciences*, vol. 132, no. 2, pp. 549–567, 2022. DOI: 10.32604/cmes.2022.019378

[18] M. Albaladejo-González, J. A. Ruipérez-Valiente, and F. G. Mármol, "Evaluating different configurations of machine learning models and their transfer learning capabilities for stress detection using heart rate," JAHC, 2023. DOI: 10.1007/s12652-022-04365-z

[19] A. Almadhor et al., "Wrist-based electrodermal activity monitoring for stress detection using federated learning," 2023.

[20] DigitalOcean, *Understanding aws's egress costs*, 2023. [Online]. Available: https://www.digitalocean.com/resources/articles/aws-egress-costs

[21] L. Furia, M. Tortora, P. Soda, and R. Sicilia, "Exploring early stress detection from multimodal time series with deep reinforcement learning," in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2023, pp. 1917–1920. DOI: 10.1109/BIBM58861.2023.10385549

[22] J. Li and P. Washington, "A comparison of personalized and generalized approaches to emotion recognition using consumer wearable devices: Machine learning study," arXiv:2308.14245, 2023. DOI: 10.48550/arXiv.2308.14245

[23] M. Abd Al-Alim, R. Mubarak, N. M. Salem, and I. Sadek, "Deep neural network based for stress detection," *Procedia Computer Science*, vol. 246, pp. 3178–3187, 2024. DOI: 10.1016/j.procs.2024.09.355

[24] M. A. Al Aleem, R. Mubarak, N. M. Salem, and I. Sadek, "A deep learning approach using wesad data for multi-class classification with wearable sensors," in *2024 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, 2024, pp. 194–197. DOI: 10.1109/NILES63360.2024.10753228

[25] A. S. Ameneshewa, G. M. Yohannes, B. N. Dawit, and M. W. Teshome, "Attention enhanced feature fusion for mental stress level classification," in *2024 21st International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, IEEE, Dec. 2024. DOI: 10.1109/ICCWAMTIP64812.2024.10873701

[26] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. L. P. Chen, "A survey on imbalanced learning: Latest research, applications and future directions," *Artificial Intelligence Review*, vol. 57, Article 137, 2024. DOI: 10.1007/s10462-024-10759-6

[27] Y. J. Lee, C. Park, H. Kim, S. J. Cho, and W.-H. Yeo, "Artificial intelligence on biomedical signals: Technologies, applications, and future directions," *Medical Engineering & Physics*, vol. 106, pp. 103–115, 2024. DOI: 10.1016/j.medengphy.2024.03.001

[28] J. Li and P. Washington, "A comparison of personalized and generalized approaches to emotion recognition using consumer wearable devices," *JMIR AI*, vol. 3, no. 1, e52171, 2024.

[29] R. Moser, J. Pfeifer, G. Varga, S. Wuchty, and M. Gärtner, "An explainable deep learning approach for stress detection in wearable sensor measurements," *Sensors*, vol. 24, no. 11, p. 5085, 2024. DOI: 10.3390/s24115085

[30] A. A. Sahlu, G. M. Yohannes, B. N. Dawit, and M. W. Teshome, "Attention enhanced feature fusion for mental stress level classification," ICCWAMTIP 2024, 2024. DOI: 10.1109/ICCWAMTIP64812.2024.10873701

[31] P. S. Shedage, S. Pouriyeh, R. M. Parizi, M. Han, G. Sannino, and N. Dehbozorgi, "Stress detection using multimodal physiological signals with machine learning from wearable devices," in *2024 IEEE Symposium on Computers and Communications (ISCC)*, 2024, pp. 1–6. DOI: 10.1109/ISCC61673.2024.10733703

[32] R. Tanwar, O. C. Phukan, G. Singh, P. K. Pal, and S. Tiwari, "Attention based hybrid deep learning model for wearable based stress recognition," *Engineering Applications of Artificial Intelligence*, vol. 127, p. 107 391, Jan. 2024. DOI: 10.1016/j.engappai.2023.107391

[33] N. Alkhaldi, *Assessing the cost of implementing ai in healthcare*, 2025. [Online]. Available: https://itrexgroup.com/blog/assessing-the-costs-of-implementing-ai-in-healthcare

[34] L. C. Dieu, *Transfer learning vs training from scratch: Which approach actually saves money for enterprise ai?* 2025. [Online]. Available: https://smartdev.com/transfer-learning-vs-training-from-scratch

[35] FinancialModel.net, *What are the startup costs for a fitness tech wearable developer?* 2025. [Online]. Available: https://www.financialmodel.net/blogs/cost-open/fitness-tech-wearables

[36] T. H. Hoang, T. K. Dang, and N. T. H. Trang, "Personalized stress detection for university students using wearable devices," IMCOM 2025, 2025. DOI: 10.1109/IMCOM64595.2025.10857507

[37] G. Mzoughi, J. Fattahi, M. Mejri, R. Ghayoula, and S. Bahroun, "Ecg-based stress surveillance using an attention-driven hybrid cnn–rnn model," ECAI 2025, 2025. DOI: 10.1109/ECAI65401.2025.11095605

[38] E. Oliver and S. Dakshit, "Cross-modality investigation on wesad stress classification," *arXiv preprint*, vol. abs/2502.18733, 2025. [Online]. Available: https://arxiv.org/abs/2502.18733

[39] C. Solutions, *How much does it cost to develop an ai solution? pricing and roi explained*, 2025. [Online]. Available: https://www.coherentsolutions.com/insights/ai-development-cost-estimation-pricing-structure-roi

[40] Spicefactory, *Edge vs. cloud in iot: Optimizing performance and cost at scale*, 2025. [Online]. Available: https://spicefactory.co/blog/2025-07-24-edge-vs-cloud-in-iot-optimizing-performance-and-cost-at-scale

[41] E. Stessman, *Fitbit vs. whoop: How do the popular fitness trackers compare? i tried both*, 2025. [Online]. Available: https://www.today.com/shop/fitbit-vs-whoop-rcna224393

[42] A. Ta et al., *Real-time stress monitoring, detection, and management in college students: A wearable technology and machine-learning approach*, 2025. arXiv: 2505.15974 `[cs.HC]`. [Online]. Available: https://arxiv.org/abs/2505.15974

[43] S. Xu, J. Zhang, Z. Wang, R. Jiang, R. Yan, and H. Tang, *Asrc-snn: Adaptive skip recurrent connection spiking neural network*, 2025. arXiv: 2505.11455 `[cs.NE]`. [Online]. Available: https://arxiv.org/abs/2505.11455

[44] A. W. Services, *Amazon sagemaker pricing*, 2026. [Online]. Available: https://aws.amazon.com/sagemaker/pricing

[45] A. W. Services, *Ec2 on-demand instance pricing*, 2026. [Online]. Available: https://aws.amazon.com/ec2/pricing/on-demand

[46] Vantage, *P3.2xlarge pricing and specs - vantage*, 2026. [Online]. Available: https://instances.vantage.sh/aws/ec2/p3.2xlarge

[47] xBattery, *How ai is revolutionizing battery life prediction: Real-world case studies*, 2026. [Online]. Available: https://xbattery.energy/learn/how-ai-is-revolutionizing-battery-life-prediction

Appendix

Abbreviations

ACC Accelerometer

BiGRU Bidirectional Gated Recurrent Unit

BILSTM Bidirectional Long Short-Term Memory

ECG Electrocardiogram

EDA Electrodermal Activity

HR Heart Rate

HRV Heart Rate Variability

PPG Photoplethysmogram

PCA Principal Component Analysis

CNN Convolutional Neural Network

RNN Recurrent Neural Network

GRU Gated Recurrent Unit

LSTM Long Short-Term Memory

SMOTE Synthetic Minority Oversampling Technique

RF Random Forest

SHAP SHapley Additive explanations

WESAD Wearable Stress and Affect Detection

Symbols

Θ Theta waves (EEG frequency band associated with stress)

β Beta waves (EEG frequency band associated with stress)

r Correlation coefficient (statistical measure of relationship strength)

R^2 Coefficient of Determination (regression performance metric)