Multimodal Stress Classification using Convolutional Neural Networks on Physiological Signals from Wearable Devices.

Mark Goertz[1], Nico Kuijpers[1], Petra Heck[1], and Manon Peeters-Schaap[1]

[1]Fontys University of Applied Sciences, Eindhoven, the Netherlands

January 13, 2025

Abstract

This study explores the use of one-dimensional convolutional neural networks (1D CNN) for multimodal stress classification using physiological signals from wearable devices. By integrating signals such as Electrodermal Activity (EDA), Blood Volume Pulse (BVP), skin temperature (TEMP) and accelerometry (ACC), the proposed model aims to improve the binary accuracy of stress detection. Physiological data from the WESAD dataset were pre-processed, segmented, and synchronized for optimal input to CNN. Various signal combinations were evaluated, with results showing higher performance for multimodal setups, particularly when EDA and BVP were included. The best configuration(EDA, BVP, ACC) achieved a classification accuracy of 0.96 and an F1 score of 0.88 on a test subject. The study highlights the significance of signal integration, preprocessing, and subject-independent testing while emphasizing the challenges of subject variability and the need for personalized models. Future work includes exploring transfer learning, personalization, broader datasets, and real-world testing to improve the generalizability of the model and its practical applicability.

Keywords: Multimodal; 1D Convolutional Neural Network; Deep Learning; Stress; Signal Processing; Wearable Devices.

1 Introduction

Wearable technology has emerged as a tool for stress monitoring, enabling real-time physiological data collection in everyday environments. Devices such as smartwatches (e.g., Apple Watch, Samsung Galaxy Watch), fitness trackers (e.g., Fitbit, Garmin), and biosensors (e.g., Empatica EmbracePlus, BioPatch) provide continuous monitoring of signals like heart rate, electrodermal activity (EDA), skin temperature, and movement. These devices facilitate large-scale stress monitoring without reliance on clinical settings. While consumer devices focus on general health tracking, wearables specifically targeted at research offer higher accuracy in EDA and access to raw data, making them well-suited for scientific investigations[1]-[2].

Physiological signals collected by wearables play a key role in stress assessment. EDA, for example, reflects variations in skin conductance caused by sweat gland activity, which is regulated by the sympathetic nervous system and often increases during stress [3]. Similarly, heart rate variability (HRV), a measure of fluctuations between consecutive heartbeats, decreases during stress due to sympathetic activation [4]. Accelerometry (ACC) provides additional context by detecting movement patterns, aiding in distinguishing stress responses from physical activity [5]. Skin temperature (TEMP) changes, such as short-term decreases due to stress-induced vasoconstriction, further contribute to stress evaluation [6].

With the rise of artificial intelligence (AI), machine learning (ML) and deep learning (DL) are applied to wearable-based technologies to analyze physiological data. These forms of AI offer the ability to process large and complex datasets, making it a promising approach for stress detection [7]. Convolutional Neural Networks (CNNs), specifically 1D CNNs, excel at processing time-series data by capturing both spatial and temporal patterns, making them particularly suitable for physiological signal analysis [8].

This study is part of the `Wearables and Stress project` at Fontys University of Applied Sciences. The project aims to improve the quality of life for individuals with persistent physical conditions or dementia [9]. Stress significantly affects both physical and mental health in these groups.

This research investigates the potential of 1D CNNs to detect stress by analyzing multiple physiological signals collected from a wearable device equiped with multiple sensors. By leveraging a multi-modal approach, the study aims to identify meaningful patterns and relationships between signals, to improve the accuracy and robustness of stress detection. Ultimately, this work seeks to contribute to the development of a wearable-based stress algorithm that is tailored for stress detection.

This paper is organized as follows:

- Section 2: literature Review, provides an overview of related work on stress detection using sensor technology, datasets and machine learning.
- Section 3: methodology, details the approach, including data collection, preprocessing, and the proposed multi-modal 1D CNN architecture.
- Section 4: results, presents the findings from the experimental setup, which are further analyzed in section 5: discussion.
- Finally, section 6, conclusion, summarizes the contributions of the paper and outlines directions for future work.

2 Literature Review

Several studies have explored the use of Convolutional Neural Networks (CNNs) for analyzing physiological data, highlighting both challenges and advancements in stress classification. The studies reviewed were selected based on their relevance to stress detection, their use of 1D CNN architectures, the availability of public datasets, and the thorough documentation of methodologies and results. This section examines related research, focusing on the application of 1-dimensional Convolutional Neural Networks (1D CNNs) for analyzing physiological signals and the datasets frequently employed for training and evaluating these models.

2.1 CNN and Traditional ML Models for Stress Detection studies

1-Dimensional Convolutional Neural Networks have been widely used for analyzing physiological signals due to their ability to extract meaningful features from sequential data. These models are particularly effective for stress detection because they can efficiently process time-series signals while capturing temporal dependencies and spatial hierarchies [8]. This subsection highlights key studies relevant to CNN-based and traditional stress detection models.

Li R. et al. (2020) used a 1D CNN to classify stress levels based on physiological signals from the WESAD dataset. Signals were normalized and segmented into 5-second windows to standardize input sizes. Two models were developed: a multi-class classifier for low, medium, and high stress levels, and a binary classifier distinguishing between stress and no-stress states. Both achieved high performance, with accuracy and F1-scores of 0.99 [10]. However, the study lacked transparency regarding its data splitting strategy. This raises concerns about data leakage, where overlapping data between training and testing sets may inflate performance metrics artificially.

Similarly, Vijayakumar et al. (2022) proposed a CNN model for stress detection using the DEAP dataset, analyzing multimodal signals: EDA, skin temperature, and blood volume pulse. Preprocessing involved segmenting 63-second signals into 1-second chunks with a sliding window, generating 125 frames per signal. The CNN architecture utilized multiple convolutional and max-pooling layers, combined with batch normalization and dropout regularization. The best combination of signals achieved an F1-score of 0.81 for valence detection and 0.79 for arousal classification [11].

Li J. et al. (2024) proposed three 1D CNN models for stress detection using the WESAD dataset: a personalized model, a participant-included generalized model, and a participant-excluded generalized model. The study focused on understanding the impact of individual participant data on model

performance. It addressed three-class emotion classification (neutral, stress, and amusement) using wearable biosignal data. Eight channels, representing the eight signal modalities from WESAD, were fed into an encoder network modeled after the encoder section of the U-Net architecture. The personalized model achieved an accuracy of 0.95 (SD ± 9.24), while the participant-included and participant-excluded generalized models achieved accuracies of 0.67 (SD ± 13.76) and 0.68 (SD ± 13.48), respectively. In terms of f1-scores, the personalized model scored 0.92 (SD ± 0.15), the participant-included model scored 0.58 (SD ± 0.18), and the participant-excluded model scored 0.59 (SD ± 0.17). These results highlight the trade-off between personalized performance and generalization, underscoring the challenges of creating stress detection models that perform reliably across diverse populations [12].

Opposed to CNN models, traditional machine learning models have also been used for stress detection. For instance, Schmidt et al. [13] used various machine learning models, such as Decision Tree, Random Forest, AdaBoost, Linear Discriminant Analysis, and k-Nearest Neighbors, to classify stress levels based on physiological signals from the WESAD dataset. Focused on the binary classification of stress and no-stress states using wrist data, the study achieved a highest accuracy of 0.87, and an F1-score of 0.84 was achieved using the Random Forest model. Other models in the study showed varying levels of performance. The Decision Tree model achieved an accuracy of 0.82 and a F1-score of 0.79. AdaBoost reached an accuracy of 0.83 and a F1-score of 0.80. Linear Discriminant Analysis (LDA) achieved an accuracy of 0.80 and a F1-score of 0.84, while the k-Nearest Neighbors (k-NN) model yielded an accuracy of 0.64 and a F1-score of 0.53. These results underline the varying performance of traditional machine learning models in stress detection, with Random Forest being the most effective among them.

2.2 Summary of Key Studies

The studies reviewed were selected for their relevance to stress detection, their use of 1D CNN architectures, and their provision of performance metrics such as accuracy and F1-score. Priority was given to studies that offered detailed descriptions of preprocessing techniques and datasets to ensure a comprehensive understanding of their methodologies. Table 1 summarizes the key studies that contribute to the development of CNN models for stress detection from physiological signals.

2.3 Innovations and Contributions of This Study

Inspired by previous works, this research try to improve stress detection using physiological signals and 1D CNNs, aiming to provide new insights and methodologies. The primary contributions of this study are as follows:

- **Data Split Strategy:** Previous studies on stress detection using physiological signals do not always provide explicit details regarding the methodology used to split datasets into training, validation, and test sets. To ensure clarity and reproducibility, this study employs a subject-level split approach. Specifically, two subjects are held out entirely for testing, guaranteeing that no data from these individuals is used during training or validation. Further details on this strategy are provided in the Methods section.

- **Systematic Exploration of Signal Combinations:** A systematic investigation is conducted into the influence of different combinations of physiological signals on stress detection performance. This approach identifies and highlights the effects of specific signal pairings, providing valuable insights into the contribution of each signal type.

- **Explainability Through Grad-CAM for 1D CNNs:** To enhance model explainability, this study adapts Gradient-weighted Class Activation Mapping (Grad-CAM) for 1D CNNs. This method provides a clearer understanding of which parts of the input signals contribute most significantly to model predictions.

- **Experiment Tracking for Reproducibility:** Experiment tracking tools are employed to maintain transparency and reproducibility. This ensures consistency across different experimental runs and allows for a detailed comparison of various configurations, improving the robustness of the results.

These innovations address key limitations of previous studies and contribute to the field by offering new insights into the effects of signal combinations and improving the explainability of stress detection models.

Study	Ref	Model Architecture	Pre-processing Techniques	Signals Used	Classification Task & Performance
Li R. et all. (2020)	[10]	Multi-perceptron Neural network - Input size for each sampling frequency. Higher frequency data goes through more hidden layers and will be concaternated. Each hidden layer uses ReLU activation.	- Normalization - 5 second segmentation	- EDA - BVP - X-axis ACC - Y-axis ACC - Z-axis ACC - TEMP WESAD wrist	Including ACC sensor Accuracy 99.65 F1-score: 99.42 Excluding ACC sensor Accuracy: 97.62 F1-score: 96.18
Li R. et all. (2020)	[10]	8 channel 1D CNN. - 3 convolutional layers ReLU Activation filter size [32, 7, 3] number of filters: [8, 16, 32] Sigmoid output.	- Normalization - Windowed segments of 5 seconds. - 100 epochs - Batch-size 40 - 7:3 train/test set.	- ECG - EDA - EMG - Respiration - Temperature - X-axis ACC - Y-axis ACC - Z-axis ACC WESAD chest	Including ACC sensor Accuracy: 99.80 F1-score: 99.67 Exluding ACC sensor Accuracy: 99.14 F1-score: 98.61
Vijayakumar et all. (2022)	[11]	1D CNN - 1 convolutional layer with 64 filters, kenel 5, ReLU. - Batch normalization - MaxPooling1D - Softmax	- Sliding window (50% overlap) - Tried several combinations best combination is: EDA, EMG, EOG, RSP, BVP	- EDA - Temperature - Respiration - BVP - ECG - EOG - EMG DEAP	Valence Binary classification - Accuracy: 75.2% - F1-score: 81.0% Arousal Binary classification - Accuracy: 68.4% - F1-score: 78.8%
Lisowska et all. (2021)	[14]	1D CNN - 2 convolutional layers with 16, 8 filters, kernel size of 3, follwed by max pooling layer and fully connected later with 30 nodes and output layer. Each convolutional layer has ReLU activation adn output layer with softmax Adam optimiser Batch-size: 256	Windowed data with 60 seconds. Leave-one-out (LOO) approach No further reported methods.	- BVP WESAD	Stress vs. no stress. - Population model: Mean F1: 0.813 (0.084 SD) - Personalized model: Mean F1: 0.822 (0.081 SD) Baseline vs. Stress vs. Amuse - Population model: Mean F1: 0.584 (0.176 SD) - Personalized model: Mean F1: 0.705 (0.119 SD)
Zang et all. (2022)	[15]	1D CNN - 2 convolutional layers - 2 max-pooling layers - 1 fully-connected layer - Activation: ReLU - SoftMax function	ECG segmented into different dura--tions (3s, 4s, 5s, and 6s) Butterworth Low-pass filter Notch filter Normalized Learning rate: 0.01 100 epochs	- ECG Private dataset	Binary classification - Accuracy - Recall 3-class classification - Accuracy - Recall
Li J. et all. (2024)	[12]	Feedforward head: - 8 channels - encoder network U-NET with two Conv. layers fol--lowed by a MaxPool. - output flattend. Activation: SiLU Personalized model.	- Normalized with a mean of 0 and SD of 1 - Sliding Window Algorithm (64 datapoints with 50% overlap) - Each participant has its own model trained. - 70% training, 15% validation, and 15% testing	- ECG - EDA - EMG - Respiration - Temperature - 3-axis ACC WESAD	Personalized model. Accuracy: 95.06 (9.24 SD) Participant-included Accuracy: 66.95 (13.76 SD) Participant-exclusive Accuracy: 67.65 (13.48 SD)
Islam et all. (2023)	[16]	Pre-text: 1D CNN model (SSL) Classification model: 1D CNN with weights from pre-text model Both models contains - 4 Convolutional Layers - Activation: ReLU - Output: Linear Activation	Self-supervised pre-training pre-train using forecasting technique	- EDA WESAD	SSL model outperformed supervised models in 90% of cases.
Ying et all. (2023)	[17]	1D-Convolutional Neural network with two Convo--lutional layers Activation: ReLU Classification: Softmax	IRR passband (0.5-30Hz removed) R-peak detection algorithm Normalized 7:3 split on WESAD-dataset	- ECG WESAD	Stacked Dimensions 1 dimension: Mean Accuracy: 96.12 Mean F1-score: 95.18 3 dimensions: Mean Accuracy: 98.57 Mean F1-score: 98.739 5 dimensions: Mean Accuracy: 98.37 Mean F1-score: 98.58

Table 1: Summary of Key Studies on 1D Convolutional Neural Networks for Stress Detection

2.4 Dataset Search and Selection

Selecting an suitable dataset is essential for training and evaluating the CNN model for stress detection. The criteria used to select datasets for the study are outlined in this subsection, along with the datasets identified as meeting these criteria.

2.4.1 Dataset Selection Criteria

For effective supervised learning, the alignment between stress events and physiological signals is critical. Accurate labels that are tightly coupled with the data ensure a reliable trained model. The following selection criteria were established to ensure the quality and relevance of the data used in this study:

- **Data Availability**: Only publicly accessible datasets providing raw signal data were considered. Public availability ensures the reproducibility of experiments and facilitates the preprocessing steps required for CNN-based analysis.
- **Sample Size and Diversity**: Datasets with at least 20 participants and diverse individuals were prioritized to ensure robust model training, testing, and validation across various stress conditions.
- **Annotations and Labels**: Supervised learning requires datasets with accurate labeling, providing clear ground truth for classification tasks. Reliable annotations are essential for correlating physiological signals with stress states and ensuring the validity of the model's outputs.
- **Sampling Frequency and Signal Length**: To ensure consistency in preprocessing and analysis, datasets with uniform or well-documented sampling frequencies were prioritized.
- **Documentation of Experimental Setup**: Datasets with detailed documentation of their experimental setup, including participant demographics, sensor placements, and environmental conditions, were given precedence. This information is critical for assessing data quality and understanding the context in which the physiological signals were recorded.

This structured approach to dataset selection ensures that the CNN model is trained and evaluated on high-quality data, maximizing its potential for accurate and reliable stress detection.

2.4.2 Identified Datasets

After reviewing several publicly available datasets, the following datasets were selected based on their alignment with the criteria outlined above. These datasets include a variety of physiological signals and documented information about the experimental setup which are important for detecting stress responses.
We found the following datasets: `WESAD [13]`, `AffectiveRoad [18], AMIGOS [19], and CASE [20]`. For detailed information about each dataset, including links to the datasets and used sensors, please refer to Table: 9.

Criteria	WESAD	AffectiveRoad	AMIGOS	CASE
Data Availability	✓	✓	✓	✓
Sample Size and Diversity	✗	✓	✓	✓
Annotations and Labels	✓	✗	✗	✗
Sampling Frequency and Signal Length	✓	✓	✓	✓
Documentation of Experimental Setup	✓	✓	✗	✗

Table 2: Comparison of datasets based on selection criteria.

In Table 2, several datasets are shown to provide valuable physiological signals for stress detection; however, the effectiveness of supervised learning models depends largely on the quality and accuracy of the temporal alignment between stress events and the recorded signals. The WESAD dataset stands out for its clear annotations and physiological signals, making it the most suitable choice for this research. In contrast, datasets such as `AffectiveRoad [18]`, `AMIGOS [19]`, and `CASE [20]` introduce challenges due to subjective or delayed annotations, which can negatively impact model reliability. Therefore, datasets with precise, well-aligned annotations are essential for improving the performance and generalizability of stress detection models.

3 Methodology

The methodology for this study is designed to explore the application of Convolutional Neural Networks (CNNs) for signal classification tasks using publicly available datasets. This section outlines the process from dataset selection and preprocessing to model design, training, and evaluation.

3.1 Selection of Dataset

The selection of the dataset plays a critical role, as the quality, variety, and relevance of the data significantly impact the performance and generalizability of the model. To ensure a comprehensive approach, multiple datasets containing physiological signals were considered. The final selection was based on a set of criteria to maximize the model's stress classification capabilities across diverse conditions. Based on these criteria, the following dataset was selected for this research:

`The Wearables Stress and Affect Detection (WESAD) dataset` by Schmidt et al. [21] was chosen as the primary dataset for this research. This dataset contains data from 15 graduate students in a lab environment, none of whom were pregnant, heavy smokers, or had mental disorders, chronic illnesses, or cardiovascular diseases. Their ages were 27.5 years ± 2.4 years, with twelve male and three female participants. Physiological data was measured using the Empatica E4 wristband and the RespiBAN chest sensorThe collected physiological signals, along with their corresponding sensors and sampling frequencies, are as follows:

- **Electrodermal Activity (EDA)**: Collected using the Empatica E4 wristband (sampled at 4 Hz) and the RespiBAN chest sensor (sampled at 700 Hz), measured in microsiemens (μS).
- **Electrocardiogram (ECG)**: Collected using the RespiBAN chest sensor (sampled at 700 Hz), measured in millivolts (mV).
- **Blood Volume Pulse (BVP)**: Collected using the Empatica E4 wristband (sampled at 64 Hz), measured in millivolts (mV).
- **3-axis Accelerometry (ACC)**: Collected using the Empatica E4 wristband (sampled at 32Hz) and RespiBAN chest sensor (sampled at 700Hz), measured in g (where 1g = 9.80665 m/s^2).
- **Skin Temperature (TEMP)**: Collected using the Empatica E4 wristband (sampled at 4 Hz) and RespiBAN chest sensor (sampled at 700Hz), measured in degrees Celsius (°C).
- **Respiration (RESP)**: Collected using the RespiBAN chest sensor (sampled at 700 Hz), measured in ohms (Ω).
- **Electromyogram (EMG)**: Collected using the RespiBAN chest sensor (sampled at 700 Hz), measured in microvolts (μV).

In addition to the physiological data, the participants also completed several self-report questionnaires, including the Positive and Negative Affect Schedule (PANAS), State-Trait Anxiety Inventory (STAI), Self-Assessment Manikin (SAM), and the Subjective Stress State Questionnaire (SSSQ). These questionnaires were used to assess the participants' emotional states, anxiety levels, and perceived stress, providing valuable context for the physiological data.

The dataset labels correspond to events during the study, where specific activities were designed to stimulate stress or other emotional states. The data is categorized into three emotional states: baseline, stress, and amusement, with each segment of data labeled accordingly. Each participant has a unique ID, ranging from 2 to 17 (data from participants 1 and 12 were not saved due to errors in the recording process). These labels allow for precise correlation of physiological responses to the events intended to elicit specific emotional reactions during the study. [13]

It is important to clarify that the participants were assigned to two distinct groups, with each group undergoing one of the two versions of the study protocol: Version A or Version B. The study protocol for the WESAD dataset is illustrated in Figure 1. Each participant completed only one version of the protocol. These versions were designed to capture physiological responses across different emotional states, such as baseline, stress, amusement, and meditation phases. This approach allowed for the examination of physiological changes related to stress and relaxation.

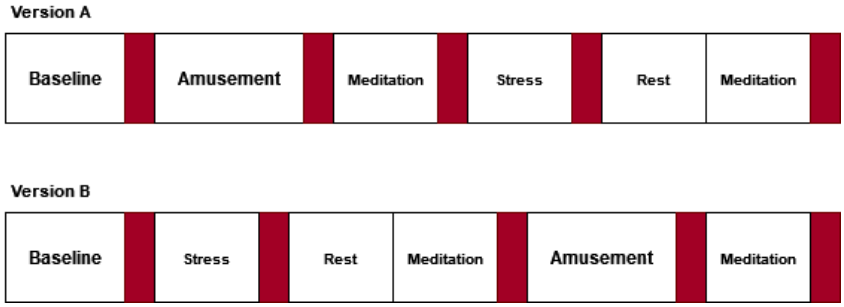

Figure 1: Study Protocol for the WESAD Dataset

3.2 Multimodal Data Integration and Pre-processing

Moving the raw data from datasets and preparing and integrating it for the model is an essential step in the process. The data from the selected datasets are preprocessed to ensure consistency and compatibility with the CNN model architecture. This section outlines the steps involved in data preprocessing, including signal synchronization, feature extraction, and label encoding.

3.2.1 Integration Strategy

The integration strategy for this study involves combining multiple physiological signals from different sensors to create a multimodal dataset. In order to effectively leverage the information from each signal to processable data for the model, a design was proposed to map the raw information to processed output. A proposed dataflow architecture for the multimodal 1D CNN model is illustrated in Figure 2. The integration strategy consists of the following steps:

- **Physiological Signals**: All raw physiological signals are collected from the WESAD dataset, including BVP, ACC, EDA, and TEMP.
- **Pre-processing**: The signals are preprocessed and normalized to ensure consistency and enhanced performance.
- **Data Segmentation**: The signals are segmented into fixed-length windows to facilitate feature extraction and classification. This segmentation step serves as input for the 1D-CNN.
- **Convolutional Neural Network**: The extracted features from each branch are concatenated to form a unified feature vector. This fusion step allows the model to leverage complementary information from different signal types.
- **Classification**: The fused feature vector is passed through fully connected layers, culminating in a softmax layer for classification. This final step outputs the probability of each class, enabling the model to distinguish between stress and non-stress states.

The model contains signal data that is preprocessed into non-overlapping windows of 8 seconds, consisting of 32 values for the lower frequencies (4Hz) and 256 values for the higher frequencies (32Hz). The model architecture is designed to process the segmented data from each modality separately, extracting features using convolutional layers before fusing the features and classifying the stress state.

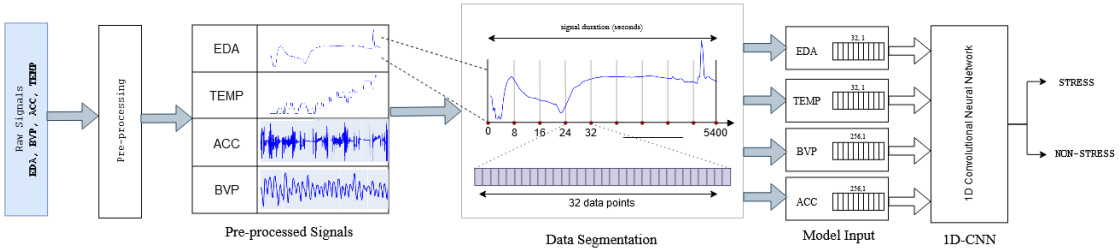

Figure 2: Proposed Multimodal 1D CNN Architecture for Stress Classification from Multimodal Physiological Signals

3.2.2 Nyquist-Shannon Sampling and Fast Fourier Transform Theorem

The Nyquist-Shannon sampling theorem establishes that to accurately represent a signal without aliasing, it is essential to sample at least twice the maximum frequency present in the signal [22]. This principle is expressed mathematically as:

$$f_s \geq 2f_{\max}$$

where f_s is the sampling frequency, and $f_{\max}$ is the highest frequency present in the signal.

In this study, we ensured the accurate analysis of the BVP signals by using a sampling rate of 32 Hz, which satisfies the Nyquist criterion for the frequency components of the BVP signal. This rate was chosen by downsampling from the original 64 Hz rate to reduce computational complexity while maintaining the physiological relevance of the data. The downsampling was implemented using the `scipy.signal.decimate` function, which applies a low-pass filter prior to downsampling, effectively mitigating the risk of aliasing [23].

To confirm that the 32 Hz sampling rate was sufficient, we analyzed the frequency content of the BVP signals using the Fast Fourier Transform (FFT). This analysis allowed us to identify the dominant frequencies within the signal, ensuring that the critical components below 10 Hz were preserved. Figures 3 and 4 show the frequency spectrum of the BVP signal sampled at both 64 Hz and 32 Hz, illustrating that the dominant frequency components are adequately captured at the reduced sampling rate.

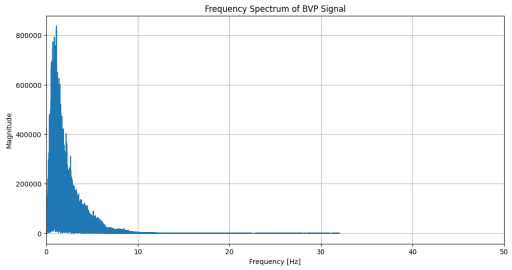

Figure 3: Frequency Spectrum of BVP signal sampled at 64 Hz

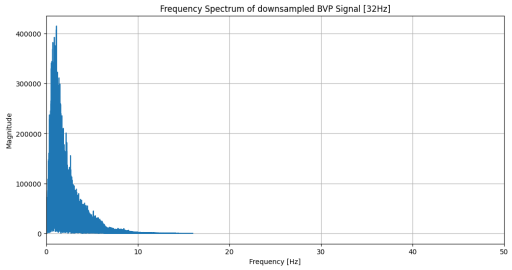

Figure 4: Frequency Spectrum of BVP signal sampled at 32 Hz

This downsampling approach ensures computational efficiency while preserving data integrity. The results, as shown in Figure 5, confirm that the dominant frequencies in the BVP signal are well represented at 32 Hz, aligning with the Nyquist theorem.

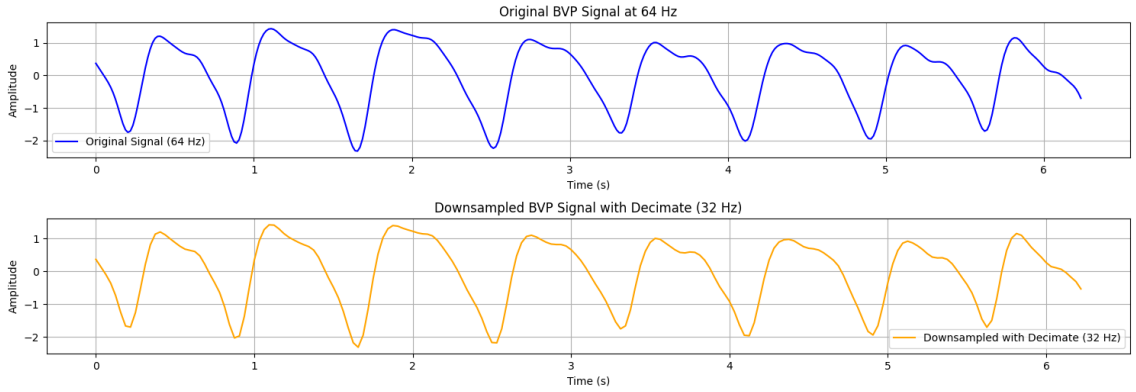

Figure 5: BVP signal sampled at 64Hz, downsampled to 32Hz with the SciPy Decimate function

3.2.3 Preprocessing Steps

Normalization:
Z-score standardization was applied to normalize the data, scaling it such that mean is 0 and standard deviation is 1. Normalization was performed on the entire signal of each subject separately to ensure consistency across all subjects [24]. The Z-score standardization formula is:

$$x_{\text{normalized}} = \frac{x - \mu}{\sigma}$$

where x is the original data, μ is the mean, and σ is the standard deviation.

Processing Acceleration Data:
The acceleration data from the wrist sensor is provided in three dimensions: X, Y, and Z. To process this data, the three dimensions were combined into a single vector by calculating the magnitude of the acceleration. The magnitude is calculated using the Euclidean norm, which is the square root of the sum of the squares of the individual components. [25]

$$a_{\text{total}} = \sqrt{a_x^2 + a_y^2 + a_z^2}$$

This approach captures the overall acceleration magnitude.

Label Encoding:
The labels in the `WESAD` dataset are correlated to the emotional states during the study. The labels are categorized into three classes:

- **Baseline**: Represents the neutral state with no emotional influence.
- **Stress**: Represents a state of increased emotional arousal or stress.
- **Amusement**: Represents a state of amusement or positive emotional response.

To streamline the classification task, the amusement and baseline states were combined into a single non-stress category. This resulted in the following label encoding:

- **No Stress (value 0)**: Combines the baseline and amusement states.
- **Stress (value 1)**: Represents the stress state.

3.2.4 Data Segmentation

The dataset used by Schmidt et al. [21] includes physiological signals from both chest and wrist sensors. The wrist sensor data was sampled at different frequencies: 4 Hz for EDA and TEMP, 32 Hz for ACC, and 64 Hz for BVP. In contrast, the chest sensor data was sampled at a higher rate of 700 Hz for ECG and RESP, with labels also provided at this 700 Hz frequency. The wrist sensor data was then segmented into 8-second windows, with the number of samples for each signal type as follows: 32 samples for EDA and TEMP, and 256 samples for ACC and BVP. The choice of 8-second windows was based on the characteristics of the EDA data, which has a slower response time compared to BVP and ACC. According to Rafiul et al. [26], the optimal window duration falls within the 5 to 10-second range.

The label for each 8-second non-overlapping window was determined by majority voting, meaning the most frequent label within the window was chosen. This approach is similar to calculating the mode of the labels for that window. The total number of available samples in a window sequence for each signal type is summarized in Table 3.

Signal Type	Original Frequency (Hz)	Used Frequency (Hz)	Sampled in a window
BVP	64	32	256
ACC(X, Y, Z)	32	32	256
EDA	4	4	32
TEMP	4	4	32

Table 3: Summary of samples in a window sequence for each signal

3.3 Model Explainability - GRADCAM

To better understand the decision-making process of the multimodal 1D CNN model, we used `Gradient-weighted Class Activation Mapping` (GRADCAM) to generate visual explanations of the model's predictions. GRADCAM calculates the weighted sum of gradients from the target class flowing into the final convolutional layers, producing a heatmap that highlights key regions of the input signals. We adapted GRADCAM from Jostein [27] and Selvaraju et al. [28] to interpret the model's predictions and identify the most influential regions in the signal.

In addition to the heatmap, we analyze two metrics to quantify interpretability: Mean intensity and its standard deviation. The Mean Intensity (with standard deviation) captures the average intensity of the gradients in the signal [28]. These metrics provide a quantitative understanding of the signal contributions to the final decision.

3.4 Model Architecture

3.4.1 Overview of the 1D CNN

The model architecture is based on a 1D-CNN designed to process time-series data from physiological signals. A 1D CNN is effective for this task as it can capture temporal patterns and features within the signal by applying convolutional filters to sequential data points [29]. The input to the model consists of time-series signals made up of preprocessed data windows from the physiological signals, each with a fixed length of x data points. The model's output is a probability distribution over the classes, indicating the likelihood that the input signal belongs to each class.

The architecture includes convolutional layers for local feature extraction, pooling layers for dimensionality reduction, and fully connected layers that aggregate features for classification. By using multiple convolutional and pooling layers, the model ultimately forms a comprehensive representation of the signals. The model design is shown in Figure 6.

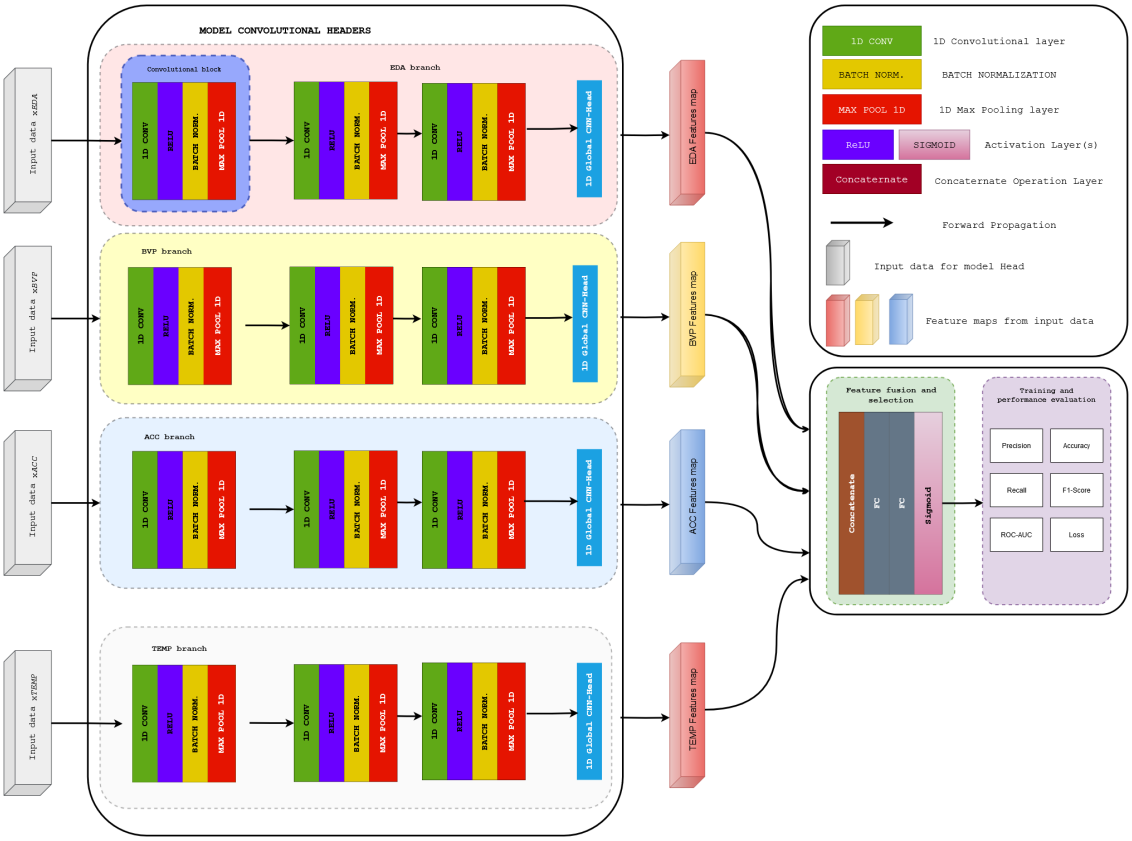

Figure 6: Multi-modal Convolutional Neural Network Architecture Configuration (visualization inspired by [30])

3.4.2 Input Layers

The input layer of the model is structured to process time-series data from multiple physiological signals, tailored specifically to the WESAD dataset. Each signal has a designated input shape based on its unique data characteristics: 32x1 for Electrodermal Activity (EDA) and Temperature (TEMP), and 256x1 for Blood Volume Pulse (BVP) and Accelerometer (ACC) data. This configuration enables the CNN to effectively learn patterns from each physiological signal using separate convolutional filters.

3.4.3 Convolutional Layers

The convolutional layers serve as the primary components of the model, tasked with identifying important patterns in the input data, ranging from simple to complex features. Three convolutional layers are employed, with filter sizes increasing progressively (32, 64, and 128), enabling the capture of a diverse range of patterns. Each convolution operates on localized regions of the input data, using a kernel size of 3, thereby emphasizing local dependencies. After each convolution, the ReLU activation function is applied, facilitating the learning of more intricate relationships within the data. The final layer utilizes a sigmoid function to output a probability between 0 and 1.

To improve generalization and mitigate overfitting, two regularization techniques are applied. Dropout with a rate of 0.3 is employed in the fully connected layers, randomly deactivating neurons during training to enhance flexibility. Additionally, L2 regularization is applied to the convolutional layers, constraining the weights to prevent excessive growth and promoting the discovery of simpler, more generalizable patterns. This layered structure allows for the extraction of hierarchical features at various levels of abstraction while minimizing the risk of overfitting.

3.4.4 Batch Normalization

Batch normalization is used after each convolution layer to keep the data steady and balanced, making the learning process smoother. It helps avoid problems like slow learning or gradients getting too small, allowing the model to learn faster. By stabilizing the data, it also helps the model perform better when faced with new, unseen data.

3.4.5 Pooling Layers

Max-pooling layers are used to reduce the size of the feature maps produced by the convolutional layers, retaining the most important features. Max-pooling with a pool size of 2 is applied after each convolutional layer, halving the feature map dimensions and highlighting the strongest activations. This helps reduce the model's computational complexity and overfitting by providing a simplified representation of the features.

3.4.6 Fully Connected Layers

The fully connected layers combine the features learned by the convolutional and pooling layers, enabling the model to classify based on patterns in the input signals. Two layers with 128 and 64 neurons are used to process these features further. ReLU activation functions are applied to each fully connected layer to introduce non-linearity. Dropout layers with a 0.3 rate are added after each fully connected layer to prevent overfitting. The final layer has one neuron with a sigmoid activation function, outputting a probability for binary classification.

3.5 Training Procedure

The training procedure for the CNN model requires steps to prepare, optimize, and fine-tune the model using measured signals as input. Here, the training process is described, including data splitting approach, loss function selection, training parameters, and regularization techniques applied to ensure the model's performance.

3.5.1 Data Preparation and Handling Class Imbalance

The dataset used in this study was preprocessed and divided into training, validation, and test sets to facilitate model training, hyperparameter tuning, and performance evaluation. Two subjects, `S16` and `S17`, were excluded from the training and validation sets to serve as the test set, ensuring that the model's evaluation is conducted on completely unseen data. This approach, displayed in

Figure: 7, allows for a fair assessment of the model's generalizability.

Preprocessing of the data, the physiological signals were divided into nonoverlapping windows of 8 seconds to form input instances. However, such preprocessing brought out an imbalanced distribution in the classes; the instances representing stress were considerably lesser as compared to instances of no-stress. Imbalancements like this may be reflected in biased model performances, with a model overhauling in favor of the majority class and/or neglecting the patterns of the minority class.

For that, `the Synthetic Minority Over-sampling Technique` (SMOTE) was utilized. SMOTE generates synthetic samples for the minority class by interpolating between existing data points. This technique increases the representation of the stress class without simple duplication and hence keeps the dataset balanced, therefore allowing the model to learn meaningful patterns across both classes. [31] [32]

Once the class imbalance was addressed, the dataset was split into training and validation sets with a 70/30 ratio. This division ensured that the majority of the data was utilized for learning, while a portion was reserved for performance monitoring during training. The random state was set to 42, and the data was shuffled to maintain a randomized and unbiased distribution. In Figure 9, the distribution of training, validation and test is shown.

- Training Set (70%): This subset served as the primary resource for model training, enabling the algorithm to discern patterns and relationships within the physiological signals. The training set was further augmented using the SMOTE technique to maintain balance in stress vs. no-stress and promote learning.
- Validation Set (30%): Used during training, this set facilitates monitoring of the model's performance, enabling adjustments to hyperparameters such as learning rate and batch size when needed.

The test set, comprising instances from subjects S16 and S17, was reserved exclusively for evaluating the model's generalizability and effectiveness on unseen data. By comparing predictions against ground truth labels, this evaluation ensured an unbiased assessment of the model. In Figure 9, the distribution of training, validation, and test sets is providing a clear overview of the data allocation strategy.

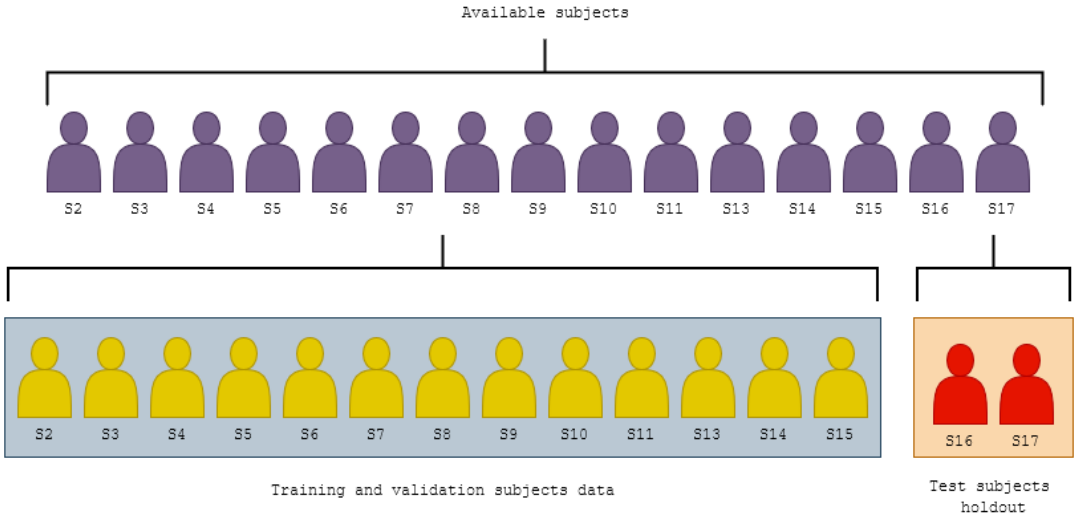

Figure 7: Subject split for training, validation, and test sets

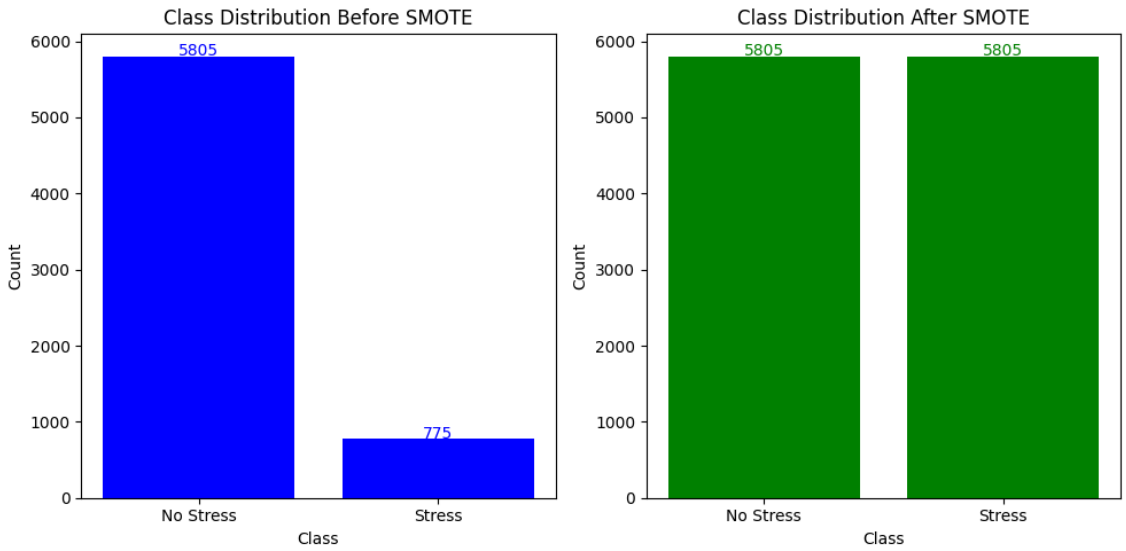

Figure 8: Data distribution of training-set: stress vs. no-stress instances

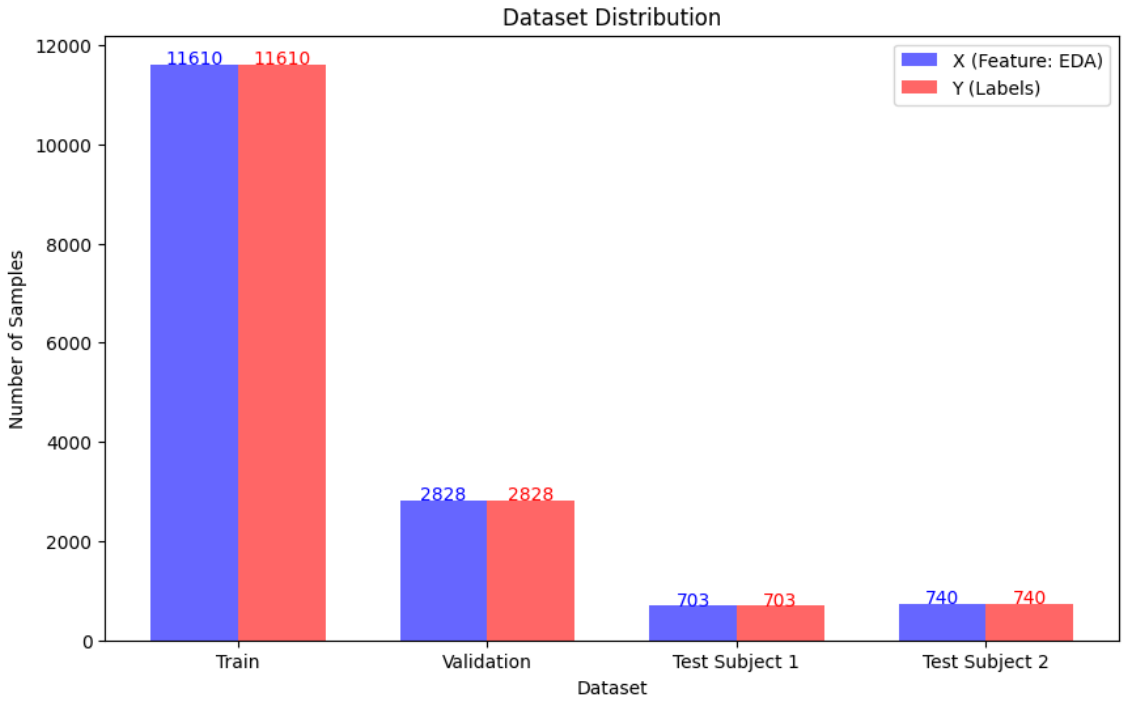

Figure 9: Data distribution: training, validation, and test sets

3.5.2 Loss Function and Optimization

To optimize the model, we utilized the `Adam optimizer`, which adjusts learning rates dynamically based on first and second moments of gradients, making it particularly effective for complex multimodal data. An initial learning rate of 0.001 was selected.

The loss function was selected to optimize the model's performance on the classification task. Given the binary nature of the classification problem, we chose the `Binary Cross-Entropy Loss` function, which is well-suited for binary classification tasks. The Binary Cross-Entropy Loss function is defined as:

$$\text{CrossEntropy-Loss} = -\sum_{i=1}^{N} y_i \log(p_i)$$

3.5.3 Training Parameters

Training parameters were tuned to optimize the balance between computational efficiency and model performance. A batch size of 64 was found to be optimal, ensuring efficient memory usage without compromising on gradient stability.

3.6 Evaluation and Experiment Tracking

3.6.1 Classification Metrics

Evaluating the effectiveness of our multimodal 1D CNN model on physiological signals requires a selection of classification metrics that can represent the accuracy and performance of the model. We adopted a range of classification metrics to evaluate the model's performance, including accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC).[33]

Binary Accuracy The binary accuracy metric measures the proportion of correctly classified samples, providing a general overview of the model's performance. Binary accuracy is defined by:

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

where True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) represent the number of correctly predicted positive samples, correctly predicted negative samples, incorrectly predicted positive samples, and incorrectly predicted negative samples, respectively.

Precision Precision measures the proportion of correctly predicted positive samples out of all samples predicted as positive. It is calculated as the ratio of true positive (TP) predictions to the sum of true positive and false positive (FP) predictions, and is defined by:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall Recall, also known as sensitivity or true positive rate, measures the proportion of correctly predicted positive samples out of all actual positive samples. It is calculated as the ratio of true positive (TP) predictions to the sum of true positive and false negative (FN) predictions, as shown in the formula below:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1-Score The F1-score is the harmonic mean of precision and recall, providing a balanced measure of the model's performance. It is calculated as the ratio of the product of precision and recall to the sum of precision and recall, as shown in the formula below:

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Area Under the Receiver Operating Characteristic Curve (AUC-ROC) The AUC-ROC metric evaluates the model's ability to distinguish between classes by measuring the area under the receiver operating characteristic curve. It provides a comprehensive assessment of the model's performance across different classification thresholds, with higher values indicating better discrimination between classes. The AUC-ROC score ranges from 0 to 1, with 0.5 representing random classification and 1 indicating perfect classification.

$$\text{AUC-ROC} = \int_0^1 \text{TPR}\, d\text{FPR}$$

Loss-score The loss-score metric evaluates the model's performance by measuring the difference between the predicted and actual values. It provides a quantitative measure of the model's accuracy, with lower values indicating better performance.

$$\text{Loss-score} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

Confusion Matrix The confusion matrix is a table used to describe the performance of a classification model by comparing the actual and predicted classifications. It provides a general view of the model's performance by showing the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The confusion matrix is defined as follows:

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

3.6.2 Experiment Tracking with DVC

To ensure consistency, reproducibility, and traceability in the training and evaluation process, we used `Data Version Control Studio (DVC)` and `Continuous Machine Learning (CML)` combined with GitHub Actions to track the experiments and the model's performance, as shown in Figure: 10. DVC is a version control system for data science and machine learning projects that making it possible to track the data, code, and models used in the experiments. This approach made it possible to reproduce the experiments, ensuring that each experiment's code and results were consistently and accurately recorded across the model development process.

DVC, a version control system and experiment tracking tool tailored for data science and machine learning, was usefull in tracking data changes and model versions, easing traceability. By managing experiment metadata, DVC enabled us to verify that identical training conditions could be reproduced, thus enhancing the validity of our model performance claims. CML, integrated with GitHub Actions, extended this by automating experiment tracking and performance logging directly in our Git repository, allowing for transparent comparisons across model iterations and ensuring accountability in the model selection process [34] - [35].

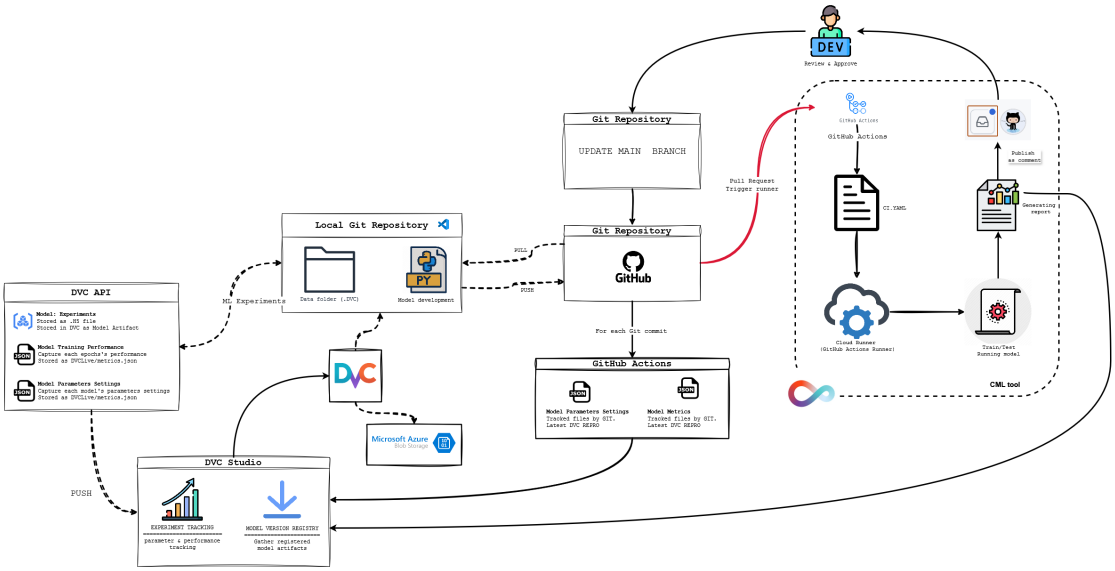

Figure 10: DVC and CML Integration

4 Results

The performance of the 1D-CNN model for stress classification was evaluated using the WESAD dataset, focusing on the raw signals from wrist-worn sensors. The model was trained, validated, and tested across all possible combinations of the four physiological signals (EDA, BVP, TEMP, ACC) to assess stress detection performance.

4.1 Model Performance

The model's performance was assessed in terms of classification metrics, including accuracy, precision, recall, F1-score, AUC, and loss.

4.1.1 Training Performance

Table 4 summarizes the model's training performance across various signal combinations. Several signal combinations achieved accuracy and F1-scores of 0.99. In contrast, the combination that included TEMP resulted in lower accuracy and F1-scores compared to the other signal combinations.

Model Training Performance on WESAD						
Signal Combinations	Binary Accuracy	Precision	Recall	F1-Score	AUC	Loss
EDA	0.92	0.90	0.95	0.92	0.97	0.22
BVP	0.97	0.95	0.98	0.97	0.99	0.16
TEMP	0.74	0.73	0.73	0.73	0.90	0.55
ACC	0.94	0.93	0.94	0.94	0.99	0.21
EDA + BVP	0.98	0.98	0.99	0.99	0.99	0.08
EDA + TEMP	0.96	0.94	0.98	0.96	0.99	0.15
EDA + ACC	0.98	0.98	0.99	0.98	0.99	0.08
BVP + TEMP	0.96	0.95	0.97	0.96	0.99	0.15
BVP + ACC	0.98	0.98	0.99	0.99	0.99	0.11
TEMP + ACC	0.98	0.97	0.98	0.98	0.99	0.12
EDA + BVP + TEMP	0.99	0.98	0.99	0.99	0.99	0.07
EDA + BVP + ACC	0.99	0.99	0.99	0.99	0.99	0.06
EDA + TEMP + ACC	0.99	0.98	0.99	0.99	0.99	0.06
BVP + TEMP + ACC	0.99	0.99	0.99	0.99	0.99	0.09
EDA + BVP + TEMP + ACC	0.99	0.98	0.99	0.95	0.99	0.05

Table 4: Training Performance of MultiModal 1D-CNN on the WESAD dataset across different wrist signal combinations.

4.1.2 Validation Performance

The validation performance of the model across different signal combinations is presented in Table 5. The highest F1-score of 0.98 was achieved by the combination of all four signals (EDA, BVP, TEMP, ACC). F1-scores above 0.92 were obtained by multimodal configurations including EDA and BVP, demonstrating better performance than single-signal models. Single-signal models, except for EDA, typically resulted in lower F1-scores compared to multimodal configurations

Model Validation Performance on WESAD						
Signal Combinations	Binary Accuracy	Precision	Recall	F1-Score	AUC	Loss
EDA	0.86	0.85	0.89	0.87	0.94	0.36
BVP	0.80	0.89	0.71	0.79	0.89	0.80
TEMP	0.73	0.72	0.74	0.73	0.78	0.56
ACC	0.80	0.84	0.74	0.79	0.88	0.64
EDA + BVP	0.92	0.95	0.88	0.92	0.96	0.35
EDA + TEMP	0.93	0.92	0.90	0.91	0.98	0.24
EDA + ACC	0.93	0.95	0.90	0.93	0.96	0.40
BVP + TEMP	0.86	0.88	0.83	0.86	0.93	0.49
BVP + ACC	0.89	0.91	0.86	0.89	0.96	0.45
TEMP + ACC	0.88	0.91	0.85	0.88	0.93	0.59
EDA + BVP + TEMP	0.93	0.97	0.90	0.93	0.98	0.27
EDA + BVP + ACC	0.96	0.97	0.94	0.96	0.98	0.22
EDA + TEMP + ACC	0.96	0.96	0.96	0.96	0.98	0.17
BVP + TEMP + ACC	0.92	0.95	0.88	0.91	0.97	0.34
EDA + BVP + TEMP + ACC	0.98	0.98	0.98	0.98	0.99	0.12

Table 5: Validation Performance of MultiModal 1D-CNN on the WESAD dataset across different wrist signal combinations.

4.1.3 Test Performance on Unseen Subjects

Tables 6 and 7 report the test performance of the MultiModal 1D-CNN on unseen subjects S16 and S17, emphasizing F1-scores as a key metric for evaluating the balance between precision and recall.

For Subject S16, (seen in Table: 6) the highest F1-score of 0.88 was achieved using the EDA + BVP + ACC signal combination, followed closely by EDA + BVP with a F1-score of 0.86, respectively. Among single-signal inputs, EDA achieved the highest F1-score of 0.70, outperforming the other single-signal modalities, while TEMP produced the lowest F1-score of 0.17.

Model Test Subject S16 Performance on WESAD						
Signal Combinations	Binary Accuracy	Precision	Recall	F1-Score	AUC	Loss
EDA	0.89	0.54	0.98	**0.70**	0.98	0.20
BVP	0.88	0.50	0.77	0.61	0.94	0.35
TEMP	0.71	0.13	0.64	**0.17**	0.40	0.65
ACC	0.85	0.42	0.26	0.51	0.80	0.66
EDA + BVP	0.96	0.75	1.0	**0.86**	0.98	0.22
EDA + TEMP	0.87	0.45	0.37	0.41	0.85	0.50
EDA + ACC	0.96	0.78	0.90	0.84	0.98	0.15
BVP + TEMP	0.87	0.46	0.74	0.57	0.89	0.43
BVP + ACC	0.90	0.59	0.76	0.66	0.93	0.43
TEMP + ACC	0.80	0.24	0.31	0.27	0.65	1.03
EDA + BVP + TEMP	0.94	0.72	0.89	0.79	0.98	0.22
EDA + BVP + ACC	0.96	0.80	0.97	**0.88**	0.99	0.17
EDA + TEMP + ACC	0.91	0.69	0.58	0.63	0.95	0.25
BVP + TEMP + ACC	0.90	0.58	0.64	0.61	0.93	0.40
EDA + BVP + TEMP + ACC	0.95	0.78	0.86	0.82	0.98	0.18

Table 6: Test Subject S16 Performance of MultiModal 1D-CNN on the WESAD dataset across different wrist signal combinations.

For Subject S17, the highest F1-score was 0.31, achieved by the BVP signal. Single-signal inputs, including EDA and TEMP, resulted in F1-scores of 0.09, while the multi-signal combination BVP + TEMP produced a higher F1-score of 0.30. The lowest F1-scores were observed across several combinations, where the metric dropped to 0.00.

Model Test Subject S17 Performance on WESAD						
Signal Combinations	Binary Accuracy	Precision	Recall	F1-Score	AUC	Loss
EDA	0.73	0.07	0.10	**0.09**	0.58	0.93
BVP	0.73	0.22	0.51	**0.31**	0.69	0.99
TEMP	0.51	0.06	0.22	**0.09**	0.45	0.74
ACC	0.81	0.23	0.24	0.24	0.52	1.28
EDA + BVP	0.86	0.36	0.16	0.22	0.59	1.35
EDA + TEMP	0.85	0.00	0.00	0.00	0.41	2.14
EDA + ACC	0.80	0.07	0.05	0.06	0.46	3.18
BVP + TEMP	0.68	0.19	0.53	0.28	0.69	1.14
BVP + ACC	0.74	0.23	0.50	**0.32**	0.70	1.09
TEMP + ACC	0.78	0.13	0.14	0.14	0.55	1.38
EDA + BVP + TEMP	0.82	0.00	0.00	0.00	0.49	2.23
EDA + BVP + ACC	0.81	0.15	0.12	0.14	0.54	1.91
EDA + TEMP + ACC	0.82	0.00	0.00	0.00	0.41	3.37
BVP + TEMP + ACC	0.77	0.24	0.41	**0.30**	0.67	1.18
EDA + BVP + TEMP + ACC	0.86	0.11	0.02	0.04	0.50	1.96

Table 7: Test Subject S17 Performance of MultiModal 1D-CNN on the WESAD dataset across different wrist signal combinations.

4.2 *Confusion Matrix of Test Subjects*

The model's performance for Test Subjects S16 and S17 is evaluated using confusion matrices for the `"Stress"` and `"No Stress"` classes. For Subject S16 (Figure 11), the model (EDA + BVP + ACC) correctly classified 599 "No Stress" (True Negatives) and 82 "Stress" (True Positives), with 20 False Positives and 2 False Negatives. For Subject S17 (Figure 12), it correctly identified 592 "No Stress" (True Negatives) and 11 "Stress" (True Positives), with 58 False Positives and 79 False Negatives.

The confusion matrixes for all subjects are available in `Section` C.

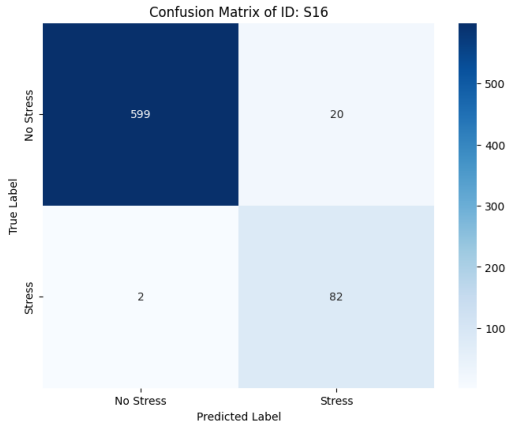

Figure 11: Confusion Matrix of the model's predictions on test subject S16.

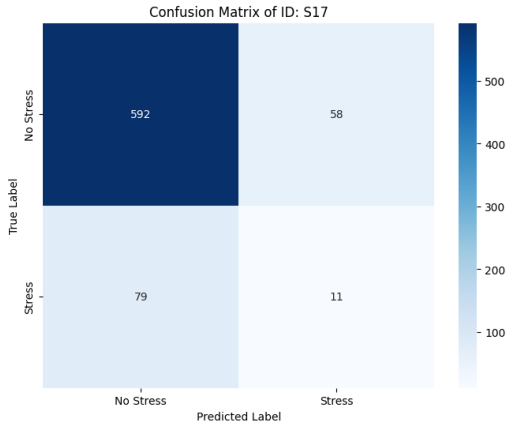

Figure 12: Confusion Matrix of the model's predictions on test subject S17.

4.3 *Plots of physiological signals predictions*

In Figures 13 and 14, the EDA signal is plotted over the test subjects S16 and S17, respectively, for the model with the signal combination EDA, BVP, and ACC. In the appendix B, the plots for the other signal combinations across all subjects are also available.

The predictions for Subject S16 show a consistent alignment with the ground truth labels, with periods of "Stress" being correctly identified and corresponding physiological signals clearly indicating the model's classification. In contrast, the predictions for Subject S17 reveal more variability, with periods of "Stress" being less accurately predicted, as reflected by more frequent misclassifications in the signal pattern.

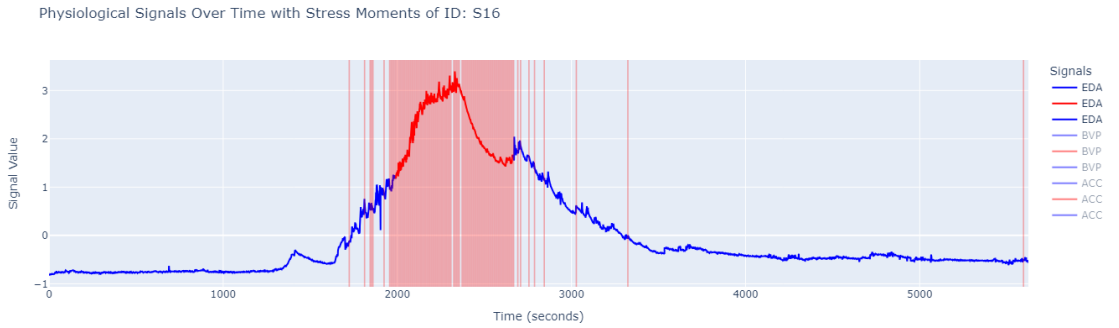

Figure 13: EDA signal for test subject S16 with signal combination: EDA, BVP, and ACC. Blue line: non-stress, red line: stress. Red background: predicted stress, white background: predicted non-stress.

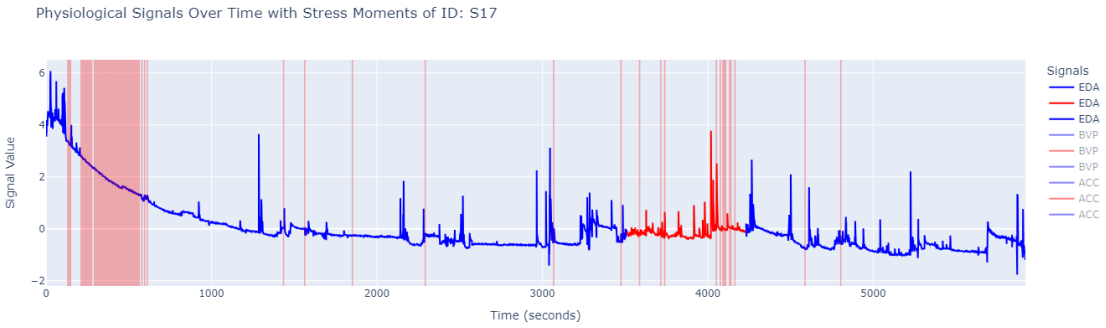

Figure 14: EDA signal plotted over test subject S17 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

4.4 GRAD-CAM Intensity Maps

The Grad-CAM intensity maps for test subject S16 are presented for True Positives, False Positives, False Negatives, and True Negatives. These maps are generated using the last convolutional layers of the model (EDA, BVP, and ACC), highlighting the regions of interest that contributed to the model's classification. The intensity maps emphasize specific areas of the EDA, BVP, and ACC signals that were crucial in the decision-making process.

In Figure 15, the points on the signal where GRAD-CAM was applied are shown for test subject S16, illustrating the locations of these intensity maps in the physiological signals.

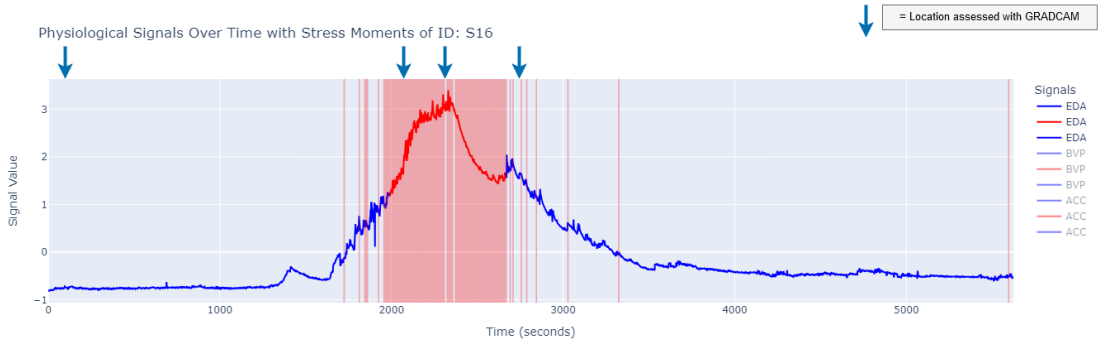

Figure 15: Locations of the GRAD-CAM intensity maps for test subject S16.

Table 8 presents the mean intensity of GRAD-CAM activations for different signals and classification outcomes. The values indicate the average contribution of each signal to the model's decision-making process for True Negatives, True Positives, False Negatives, and False Positives.

Signal	True Negative	True Positive	False Negative	False Positive
EDA	0.20	0.51	0.30	0.43
BVP	0.21	0.21	0.24	0.22
ACC	0.16	0.14	0.15	0.19

Table 8: Mean intensity of GRAD-CAM activations for different signals and classification outcomes.

4.4.1 True Negative: No-Stress - No-Stress

The Grad-CAM intensity maps for a True Negative case (No-Stress correctly classified as No-Stress) are shown using the coolwarm intensity scale for the EDA, BVP, and ACC signals. In this scale, blue areas represent lower contributions to the model's classification, while red areas indicate higher contributions.

- Figure 16 shows the Grad-CAM intensity map for the EDA signal.
- Figure 17 highlights the regions of the intensity map for the BVP signal.
- Figure 18 illustrates the intensity for the ACC signal.

To evaluate the importance of each signal in this no-stress classification, we analyzed the GRAD-CAM results. The mean importance refers to the average activation intensity across the signal, quantifying the overall contribution of different regions of the signal to the model's final prediction.

- **EDA Summary:** The EDA signal had the highest mean intensity of 0.20 (STD ±0.24) among the three signals.
- **BVP Summary:** The BVP signal showed a mean intensity 0.21 (STD ±0.22), suggesting a moderate contribution to the classification compared with EDA and ACC.
- **ACC Summary:** The ACC signal exhibited the smallest mean intensity of 0.16 (STD ±0.17) of the three signals.

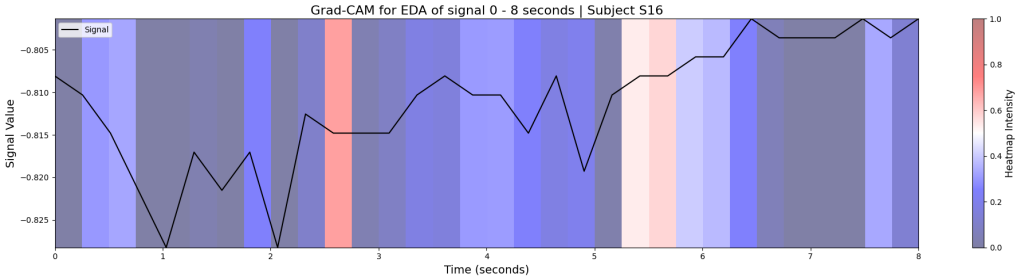

Figure 16: GRAD-CAM intensity of EDA on test subject S16 (True Negative No-Stress).

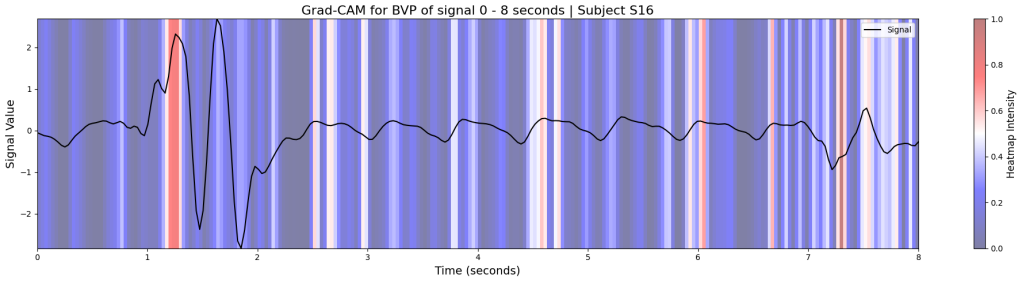

Figure 17: GRAD-CAM intensity of BVP on test subject S16 (True Negative No-Stress).

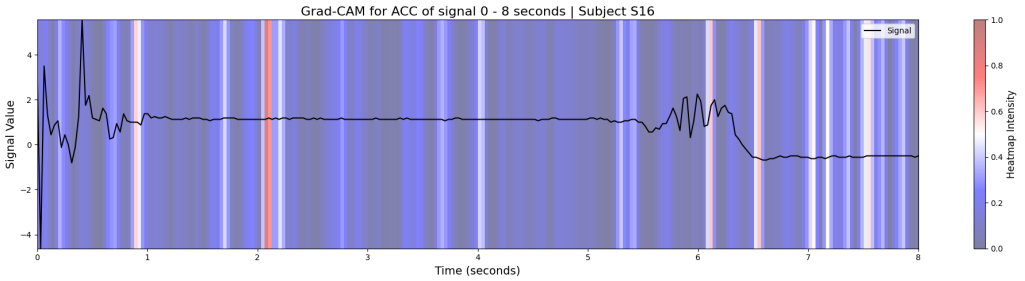

Figure 18: GRAD-CAM intensity of ACC on test subject S16 (True Negative No-Stress).

4.4.2 True Positive: Stress - Stress

The Grad-CAM intensity maps for a True Positive case (Stress correctly classified as Stress) are presented for the EDA, BVP, and ACC signals. These maps visualize the regions of the input data that contributed to the model's classification of the stress state.

- Figure 19 shows the Grad-CAM intensity map for the EDA signal, indicating the areas with the highest activation to the classification.
- Figure 20 displays the intensity map for the BVP signal.
- Figure 21 illustrates the intensity for the ACC signal.

Each map focuses on the specific regions of the signal that played a key role in the model's decision, offering valuable insights into the interpretability of the classification process. To evaluate the importance of each signal in stress classification, we analyzed the GRAD-CAM results:

- **EDA Summary:** The EDA signal had the highest mean intensity of 0.51 (STD ±0.28) among the three signals, indicating its dominant role in the model's stress classification.
- **BVP Summary:** The BVP signal showed a mean intensity of 0.21 (STD ±0.22), suggesting a moderate contribution to the classification compared with EDA and ACC.
- **ACC Summary:** The ACC signal exhibited the smallest mean intensity of 0.14 (STD ±0.24), indicating its limited contribution compared to EDA and BVP.

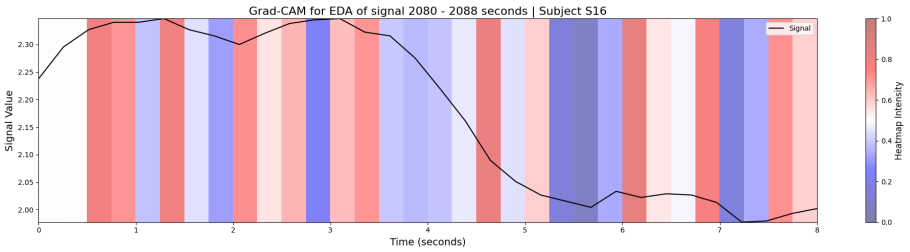

Figure 19: GRAD-CAM intensity of EDA on test subject S16 (Positive Stress).

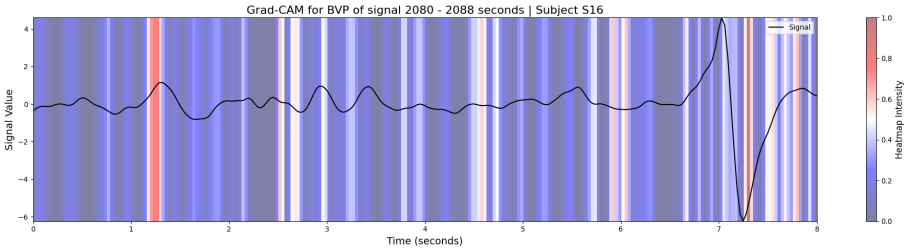

Figure 20: GRAD-CAM intensity of BVP on test subject S16 (Positive Stress).

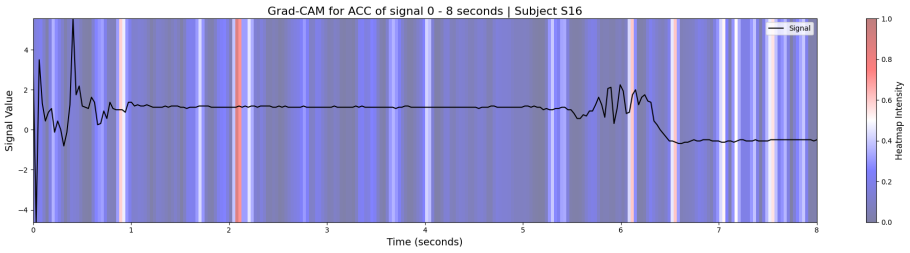

Figure 21: GRAD-CAM intensity of TEMP on test subject S16 (Positive Stress).

4.4.3 False Negative: Stress - No-Stress

The Grad-CAM intensity maps for a False Negative case (Stress incorrectly classified as No-Stress) are presented for the EDA, BVP, and ACC signals. These maps show the regions within each signal that contributed to the model's classification of the No-Stress state. The model assigned a probability of 0.023 to the no-stress classification.

- Figure 22 shows the Grad-CAM intensity map for the EDA signal, indicating the areas with activations to the classification.
- Figure 23 displays the intensity map for the BVP signal.
- Figure 24 illustrates the intensity for the ACC signal.

The Grad-CAM analysis provided the following results:

- **EDA Summary:** The mean intensity is 0.30 (STD. ±0.23).
- **BVP Summary:** The mean intensity is 0.24 (STD. ±0.26).
- **ACC Summary:** The mean intensity is 0.15 (STD. ±0.20).

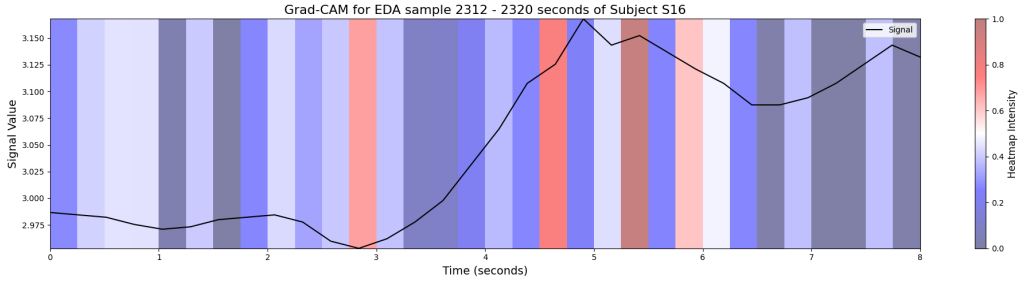

Figure 22: GRAD-CAM intensity of EDA on test subject S16 (False Negative No-Stress).

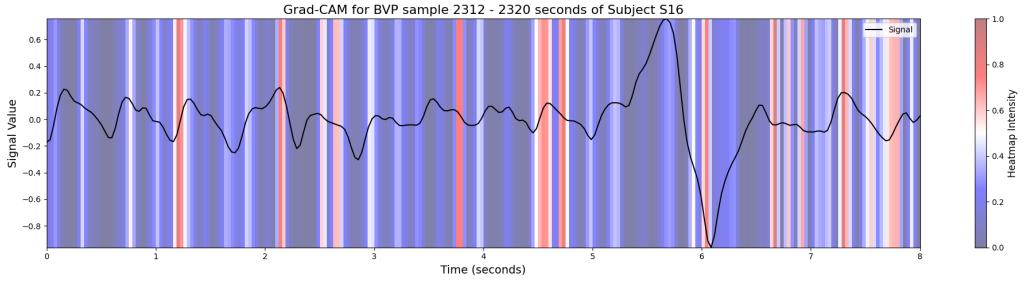

Figure 23: GRAD-CAM intensity of BVP on test subject S16 (False Negative No-Stress).

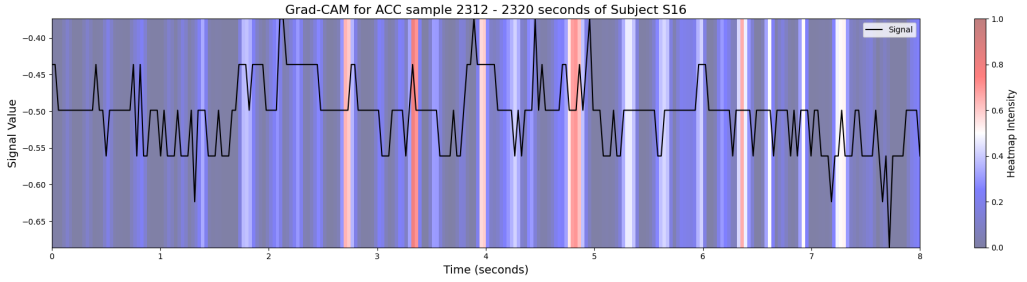

Figure 24: GRAD-CAM intensity of ACC on test subject S16 (False Negative No-Stress).

4.4.4 False Positive: No Stress - Stress

The Grad-CAM intensity maps for a False Positive case (where No-Stress is incorrectly classified as Stress) are presented for the EDA, BVP, and ACC signals. The model predicted the stress label with a probability of 0.98, despite the true label being "No-Stress." These maps highlight the regions of the input data that contributed to the incorrect classification.

- Figure 25 shows the Grad-CAM intensity map for the EDA signal, indicating the areas with activations to the classification.
- Figure 26 displays the intensity map for the BVP signal.
- Figure 27 illustrates the intensity for the ACC signal.

The Grad-CAM analysis yielded the following results:

- **EDA Summary:** The EDA signal had the highest mean intensity of 0.43 (STD ±0.28), indicating its contribution to the model's classification.
- **BVP Summary:** The BVP signal exhibited a mean intensity of 0.22 (STD ±0.22), reflecting its moderate contribution to the false positive classification.
- **ACC Summary:** The ACC signal showed the smallest mean intensity 0.19 (STD ±0.21), indicating its relatively minor contribution to the model's decision compared to the other signals.

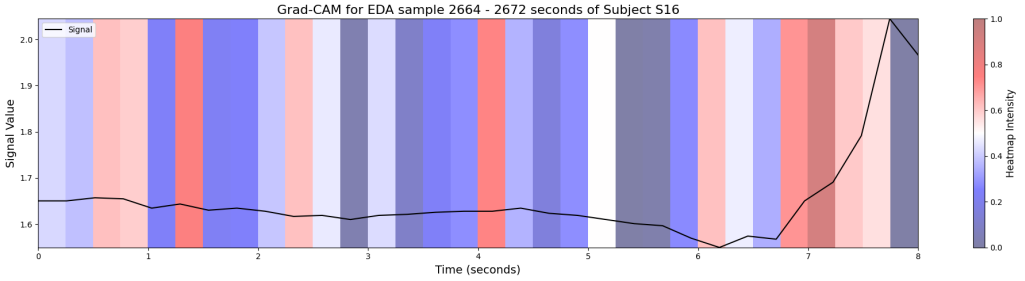

Figure 25: GRAD-CAM intensity of EDA on test subject S16 (False Positive Stress).

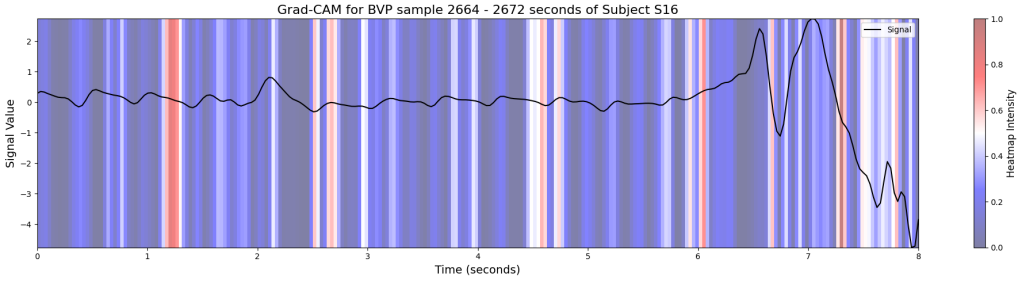

Figure 26: GRAD-CAM intensity of BVP on test subject S16 (False Positive Stress).

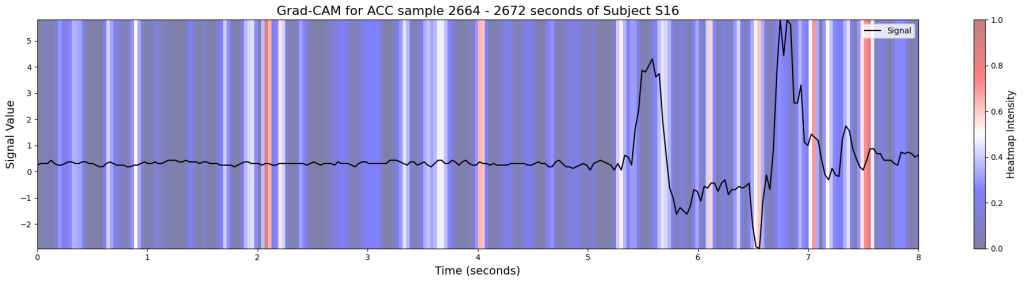

Figure 27: GRAD-CAM intensity of ACC on test subject S16 (False Positive Stress).

5 Discussion

This section provides an analysis of the key findings and observations, performance, and comparison with previous studies. We also address the impact of preprocessing and model configurations, data version control, and experiment tracking. The subject-specific performance and the comparison with previous studies are also discussed. Finally, we outline potential future work to enhance the model's performance and generalizability.

5.1 Key Findings and Observations

EDA and BVP emerged as the most informative signals for the model, aligning with their strong relation to the sympathetic nervous system. EDA, influenced by changes in sweat gland activity, directly reflects arousal, while BVP provides detailed insights into cardiovascular responses under stress. Combinations including these two modalities outperformed configurations without them, highlighting their importance in stress detection.
The best performance was achieved when EDA, BVP, and ACC signals were combined. This suggests that integrating multiple modalities allows the model to capture a broader spectrum of physiological responses to stress. The fusion of signals enhances the model's capacity to learn diverse patterns, resulting in a more robust and complete representation of stress.
While multimodal integration yielded the highest overall performance, results from individual signals also provided valuable insights. Skin temperature did not appear to contribute significantly, likely due to its slower-changing signals may not align well with the 8-seconds window size. This finding reinforces the importance of signal selection to improve classification accuracy. However, if the algorithm were adapted to incorporate 'lags,' allowing slower signals (i.e. skin temperature) to be taken care for, the model could potentially be further enriched.

The GRAD-CAM results further support the key findings of the model's behavior. By visualizing the areas of interest in the signals that the model focused on, we can confirm the relevance of EDA and BVP in stress detection. For EDA, the heatmap consistently highlighted a large part of the signal over the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) cases, confirming that EDA has a major contribution to the model's prediction outcome. Similarly, for BVP, the GRAD-CAM heatmaps showed strong attention to heartbeats, confirming the model's sensitivity to the physiological markers of stress.
Despite these strengths, the analysis encountered challenges in interpreting the model's focus in reversed scenarios—specifically, identifying no-stress features in stress-predicted signals and vice versa. While GRAD-CAM results illustrated the influence of specific signals on the predicted class, distinguishing between stress and no-stress features in the heatmaps was not possible. This limitation suggests that the use of a softmax activation function, rather than sigmoid, might have allowed clearer differentiation in these cases.

5.2 Impact of Preprocessing and Model Configurations

Data preprocessing is critical in improving the model's performance by standardizing and normalizing input signals. Techniques such as downsampling, filtering, and windowing ensure the model can effectively extract meaningful patterns from the data. However, determining the optimal preprocessing configuration is challenging due to the numerous variables and parameters involved. For instance, data augmentation techniques help balance class distributions, enabling the model to learn from a representative dataset and generalize effectively to unseen data. Another key factor is window size, which influences how much data is fed into the model at each step. Proper window sizing ensures that the model captures relevant temporal patterns in the physiological signals.
The architecture and hyperparameters of the model also play a significant role in its performance. The selection of convolutional layers, pooling layers, and fully connected layers, along with activation functions and regularization techniques, directly impacts the model's ability to extract features and learn complex patterns. Optimizing these configurations is crucial for accurately modeling the physiological responses associated with stress.

5.3 Data Version Control and Experiment Tracking

DVC (Data Version Control) was instrumental in managing the machine learning workflow by enabling systematic versioning of data, models, and code. This ensured reproducibility and traceability across all experiments.

However, integrating DVC with workflows involving cross-validation presented challenges. Cross-validation requires multiple iterations over different data splits, which conflicted with DVC's structure and added complexity to experiment tracking. After evaluating alternative tools and approaches, we determined that combining cross-validation with DVC was impractical. Consequently, the use of cross-validation was excluded in this study, and workflows with DVC's strengths were prioritized. This adjustment allowed greater focus on ensuring consistency and reproducibility in the research process.

5.4 Subject-Specific Performance

The model's performance varied significantly when evaluated on unseen subjects, highlighting challenges in generalizing across individuals. For instance, the model achieved high accuracy (0.93) and F1-score (0.80) for test subject S16, indicating its effectiveness in capturing stress patterns for this individual. In contrast, the performance for test subject S17 was notably low, with accuracies ranging from 0.54 to 0.87, and F1-scores ranging from 0.0 to 0.32 depending on the signal combination. This anomaly in performance led to a critical question: "Did Subject 17 from the WESAD dataset experience any measurable stress during the experiment?"

Physiological markers such as EDA are frequently used to detect stress, as they are indicative of sympathetic nervous system activity, which typically increases during stressful events. However, for Subject S17, the expected physiological markers were not clearly observed in this study. The EDA signal showed minimal to no amplitude, indicating a lack of the anticipated increase in skin conductance, which is commonly associated with sympathetic nervous system activation during stress.

To investigate this further, a review of the data for Subject S17 was conducted by an expert in biomedical data. It was noted that there were no prominent stress-related markers in the physiological data, making it difficult to determine whether the subject experienced stress during the experiment. Although mild fluctuations were observed in the signals later in the session, these fluctuations were less pronounced than those seen in other subjects and did not provide clear evidence of stress. In contrast to the physiological data, subject S17 self-reported experiencing stress during the experiment. In the PANAS and STAI questionnaires, the subject reported feelings of stress. Specifically, in the PANAS, stress was marked as a 5 on a 1-5 scale. However, this was not reflected in the physiological data, as the expected increase in signal amplitude was not observed. For subject S16, however, the self-reported stress from the PANAS and STAI questionnaires aligned well with the physiological data, showing clear signs of sympathetic nervous system activation, such as increased EDA amplitude.
This difference between self-reported stress and physiological data for subject S17 highlights the complexity of stress detection and suggests the need for future research to explore the factors that influence the relationship between self-reports and physiological responses, as well as the importance of considering a personalized stress detection algorithm, rather than a one-size-fits-all approach.

5.5 Comparison with Previous Studies

The findings of this study align with prior research demonstrating the effectiveness of multimodal physiological signals for stress classification in terms of accuracy. However, our study extends this field by addressing critical limitations observed in earlier works, such as data splitting strategies, preprocessing steps, and explainability.

In comparison to the work by Li et al. (2020) [10], which reports high accuracy (0.99) but lacks detailed information on the data splitting strategy, this study places greater emphasis on transparency. The data splitting process is clearly outlined, utilizing a train/validation set (13 subjects) and two separate test subjects, ensuring a robust evaluation of the model's generalizability. In contrast to Li et al., whose testing procedure was not fully specified, our study emphasizes the importance of a clear separation between training and test data.
Furthermore, our study achieved an F1-score of 0.88 for test subject S16 using a multimodal configuration (EDA, BVP, and ACC), indicating a more balanced performance in terms of precision and recall. In contrast, Li et al.'s reported F1-score of 0.99, however, may suggest overfitting, where the model could have learned patterns specific to the training data and not generalized effectively to unseen subjects.

Li J. et al. (2024) [12] explored transfer learning to improve stress detection performance by leveraging pre-trained architectures. Their study highlighted the trade-offs between personalized and generalized models, with personalized models achieving superior accuracy (0.95) and F1-score (0.91), compared to participant-exclusive generalized models (accuracy: 0.67, F1-score: 0.43).
In comparison, our study, without employing transfer learning, outperformed the best performing participant-exclusive generalized model by Li J. et al. We achieved an accuracy of 0.96 and an F1-score of 0.88 for test subject S16. These results underscore the robustness of our approach while highlighting the potential of transfer learning as a future direction.

Additionally, Schmidt et al. [36] used traditional machine learning models, including decision trees and random forests, to classify stress using physiological signals. Among the models evaluated, the random forest algorithm was the best-performing for binary classification, achieving a mean accuracy of 0.87 and a mean F1-score of 0.84.
Our approach, which employs a deep learning architecture, achieved an accuracy of 0.96 and an F1-score of 0.88 on test subject S16 using multimodal signals, including EDA, BVP, and ACC. Both approaches demonstrate the potential for stress detection using physiological signals. While traditional machine learning models, such as those used by Schmidt et al., have shown strong performance, deep learning methods may offer additional capabilities by leveraging raw data to learn complex patterns.

5.6 Future Work

This study has demonstrated the efficacy of a multi-modal 1D CNN, incorporating subject-independent normalization and experiment tracking. While the proposed approach showed promising results, it also opens several avenues for future research that could further enhance the robustness and generalizability of the model.

5.6.1 Transfer Learning

Future work can explore the use of transfer learning to improve stress detection models, particularly when dealing with limited labeled data. Transfer learning allows models to leverage knowledge from one dataset and apply it to another, reducing the need for extensive data collection. By fine-tuning pre-trained models from datasets, such as WESAD, to specific individuals, models can adapt to unique stress responses, even with small amounts of data. Incorporating transfer learning could improve the generalization of stress detection models, enabling them to perform well across diverse populations and settings with less reliance on large, labeled datasets.

5.6.2 Real-World Application and Testing

While the model has demonstrated promising results in a controlled experimental setting, future work should focus on validating the model in real-world applications. A model for wearable devices in everyday environments would allow for continuous, real-time stress monitoring in daily life. This would require addressing challenges such as environmental factors such as temperature, movement, and noise, which can impact signal quality and model performance.

To further enhance real-world applicability, a rolling window approach could be applied to the model's probability outputs to stabilize predictions over time. Instead of making predictions based solely on individual instances, the model can average probabilities across a rolling window of recent predictions. This method will be more dynamically and result in smoother stress detection.

5.6.3 Exploration of Data Preprocessing Techniques

Currently, the model uses raw physiological signals, which may contain noise and artifacts that could impact classification performance. Future work could explore how different data preprocessing techniques affect stress detection. In particular, variations in the sliding window method could reveal how different window configurations influence model accuracy. Adjusting the window length of each input to match the signal characteristics could help the model better capture important temporal patterns. Additionally, applying preprocessing techniques such as filtering, noise reduction, and feature extraction could improve the quality of the input data, leading to better model performance.

5.6.4 Personalized Stress Detection

Incorporating personalized models for stress detection is another promising direction for future work. Stress responses are highly individual, influenced by factors and environmental context. These differences mean that a generalized model may not be equally effective for all individuals. A personalized model could adapt to each user's unique physiological responses to stress, thereby improving the accuracy and reliability of stress detection. Prior research by Li et al. (2024) [12] has shown that personalized models, which are tailored to individual users, can outperform generalized models in certain settings.

6 Conclusion

The proposed 1D CNN model for multimodal stress classification has demonstrated promising results in capturing the complex temporal and spatial dependencies within physiological signals. By leveraging multimodal physiological signals such as Electrodermal Activity (EDA), Blood Volume Pulse (BVP), skin temperature (TEMP), and accelerometry (ACC), the model effectively differentiates between stress and non-stress. The findings underscore the importance of multimodal data integration in enhancing the accuracy and robustness of stress classification models. Challenges remain in optimizing the model architecture, hyperparameters, and data preprocessing techniques to achieve optimal performance. And subject variability in physiological responses to stress presents a significant challenge in generalizing the model's performance across diverse populations. Furthermore, the real-world application of the model in stress detection scenarios requires further testing to ensure its reliability and effectiveness. Addressing these limitations and challenges will be crucial for advancing the field of multimodal stress classification and developing practical solutions for stress management and well-being with wearables.

Availability of data and materials

This research uses the WESAD dataset, which was made publicly available for scientific research. We acknowledge the creators of the dataset and provide credit for their work as required by the dataset's terms of use.

Acknowledgments

This research has been co-financed by "Regieorgaan SIA", part of the "Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO)" and Fontys Kenniscentrum Applied AI for Society.

Disclosure statement:

No potential conflict of interest is reported by the authors

References

[1] M. Basza, B. Krzowski, P. Balsam, M. Grabowski, G. Opolski, and Ł. Kołtowski, "An Apple Watch a day keeps the doctor away?" *Cardiology Journal*, vol. 28, no. 6, p. 801, Dec. 31, 2021. DOI: 10.5603/CJ.2021.0140. pmid: 34985118. Accessed: Nov. 14, 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8747830/.

[2] "E4 wristband — Real-time physiological signals — Wearable PPG, EDA, Temperature, Motion sensors," Empatica, Accessed: Jan. 12, 2025. [Online]. Available: https://www.empatica.com/research/e4.

[3] W. Boucsein, *Electrodermal Activity*. Boston, MA: Springer US, 2012, ISBN: 978-1-4614-1125-3 978-1-4614-1126-0. DOI: 10.1007/978-1-4614-1126-0. Accessed: Nov. 18, 2024. [Online]. Available: https://link.springer.com/10.1007/978-1-4614-1126-0.

[4] J. E. Peabody, R. Ryznar, M. T. Ziesmann, and L. Gillman, "A Systematic Review of Heart Rate Variability as a Measure of Stress in Medical Professionals," *Cureus*, vol. 15, no. 1, e34345, Jan. 2023, ISSN: 2168-8184. DOI: 10.7759/cureus.34345. pmid: 36865953.

[5] M. Meina *et al.*, "Heart Rate Variability and Accelerometry as Classification Tools for Monitoring Perceived Stress Levels-A Pilot Study on Firefighters," *Sensors (Basel, Switzerland)*, vol. 20, no. 10, p. 2834, May 16, 2020, ISSN: 1424-8220. DOI: 10.3390/s20102834. pmid: 32429383.

[6] K. A. Herborn *et al.*, "Skin temperature reveals the intensity of acute stress," *Physiology & Behavior*, vol. 152, pp. 225–230, Pt A Dec. 1, 2015, ISSN: 1873-507X. DOI: 10.1016/j.physbeh.2015.09.032. pmid: 26434785.

[7] "Wearable Technology Applications in Healthcare: A Literature Review - ProQuest," Accessed: Nov. 14, 2024. [Online]. Available: https://www.proquest.com/openview/6c96964dfb83ca06895f330231?pq-origsite=gscholar&cbl=2034896.

[8] A. O. Ige and M. Sibiya, "State-of-the-Art in 1D Convolutional Neural Networks: A Survey," *IEEE Access*, vol. 12, pp. 144 082–144 105, 2024, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2024.3433513. Accessed: Nov. 15, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10609371/?arnumber=10609371.

[9] "Wearables & stress — Fontys Hogeschool," Accessed: Nov. 16, 2024. [Online]. Available: https://www.fontys.nl/Onderzoek/Health-Innovations-Technology-1/Wearables-en-stress-technologie-voor-gezondheid-en-welzijn.htm?bcUrl=https%3A%2F%2Fi453.fontys.nl%2Fscript.js&elementHolder=8930373&sign=d60dddb546bfafcd54ef31117d09770c.

[10] R. Li and Z. Liu, "Stress detection using deep neural networks," *BMC Medical Informatics and Decision Making*, vol. 20, no. 11, p. 285, Dec. 30, 2020, ISSN: 1472-6947. DOI: 10.1186/s12911-020-01299-4. Accessed: Sep. 11, 2024. [Online]. Available: https://doi.org/10.1186/s12911-020-01299-4.

[11] S. Vijayakumar, R. Flynn, P. Corcoran, and N. Murray, "CNN-based Emotion Recognition from Multimodal Peripheral Physiological Signals," 2022.

[12] J. Li and P. Washington, "A Comparison of Personalized and Generalized Approaches to Emotion Recognition Using Consumer Wearable Devices: Machine Learning Study," *JMIR AI*, vol. 3, no. 1, e52171, May 10, 2024. DOI: 10.2196/52171. Accessed: Nov. 14, 2024. [Online]. Available: https://ai.jmir.org/2024/1/e52171.

[13] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, *WESAD*. Accessed: Oct. 16, 2024. [Online]. Available: https://ubi29.informatik.uni-siegen.de/usi/data_wesad.html.

[14] A. Lisowska, S. Wilk, and M. Peleg, "Catching Patient's Attention at the Right Time to Help Them Undergo Behavioural Change: Stress Classification Experiment from Blood Volume Pulse," in *Artificial Intelligence in Medicine*, A. Tucker, P. Henriques Abreu, J. Cardoso, P. Pereira Rodrigues, and D. Riaño, Eds., Cham: Springer International Publishing, 2021, pp. 72–82, ISBN: 978-3-030-77211-6. DOI: 10.1007/978-3-030-77211-6_8.

[15] X. Zang, B. Li, L. Zhao, D. Yan, and L. Yang, "End-to-End Depression Recognition Based on a One-Dimensional Convolution Neural Network Model Using Two-Lead ECG Signal," *Journal of Medical and Biological Engineering*, vol. 42, no. 2, pp. 225–233, Apr. 1, 2022, ISSN: 2199-4757. DOI: 10.1007/s40846-022-00687-7. Accessed: Nov. 14, 2024. [Online]. Available: https://doi.org/10.1007/s40846-022-00687-7.

[16] T. Islam and P. Washington. "Personalization of Stress Mobile Sensing using Self-Supervised Learning." arXiv: 2308.02731 [cs], Accessed: Nov. 14, 2024. [Online]. Available: http://arxiv.org/abs/2308.02731, pre-published.

[17] G. Ying and C. Hao, "A novel one-dimensional convolutional neural network-based method for emotion recognition of electric power industry workers," *Energy Reports*, 2022 The 3rd International Conference on Power Engineering, vol. 9, pp. 763–771, Sep. 1, 2023, ISSN: 2352-4847. DOI: 10.1016/j.egyr.2023.04.297. Accessed: Nov. 14, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352484723006649.

[18] N. E. Haouij, J.-M. Poggi, S. Sevestre-Ghalila, R. Ghozi, and M. Jaïdane, "AffectiveROAD system and database to assess driver's attention," in *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, ser. SAC '18, New York, NY, USA: Association for Computing Machinery, Apr. 9, 2018, pp. 800–803, ISBN: 978-1-4503-5191-1. DOI: 10.1145/3167132.3167395. Accessed: Nov. 18, 2024. [Online]. Available: https://doi.org/10.1145/3167132.3167395.

[19] "AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups," Accessed: Nov. 18, 2024. [Online]. Available: http://www.eecs.qmul.ac.uk/mmv/datasets/amigos/readme.html.

[20] K. Sharma, C. Castellini, E. L. van den Broek, A. Albu-Schaeffer, and F. Schwenker, "A dataset of continuous affect annotations and physiological signals for emotion analysis," *Scientific Data*, vol. 6, no. 1, p. 196, Oct. 9, 2019, ISSN: 2052-4463. DOI: 10.1038/s41597-019-0209-0. Accessed: Nov. 18, 2024. [Online]. Available: https://www.nature.com/articles/s41597-019-0209-0.

[21] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, Boulder CO USA: ACM, Oct. 2, 2018, pp. 400–408, ISBN: 978-1-4503-5692-3. DOI: 10.1145/3242969.3242985. Accessed: Oct. 15, 2024. [Online]. Available: https://dl.acm.org/doi/10.1145/3242969.3242985.

[22] E. Por, M. van Kooten, and V. Sarkovic, "Nyquist–Shannon sampling theorem,"

[23] "Decimate — SciPy v1.14.1 Manual," Accessed: Oct. 15, 2024. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.signal.decimate.html.

[24] "StandardScaler," scikit-learn, Accessed: Dec. 30, 2024. [Online]. Available: https://scikit-learn/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

[25] "Euclidean Norm - an overview — ScienceDirect Topics," Accessed: Dec. 9, 2024. [Online]. Available: https://www.sciencedirect.com/topics/engineering/euclidean-norm.

[26] R. Amin and R. T. Faghih, "Physiological characterization of electrodermal activity enables scalable near real-time autonomic nervous system activation inference," *PLOS Computational Biology*, vol. 18, no. 7, e1010275, Jul. 28, 2022, ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1010275. Accessed: Dec. 9, 2024. [Online]. Available: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1010275.

[27] "WPW detection from ECG using 1D CNN," Accessed: Nov. 18, 2024. [Online]. Available: https://kaggle.com/code/bjoernjostein/wpw-detection-from-ecg-using-1d-cnn.

[28] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Feb. 2020, ISSN: 0920-5691, 1573-1405. DOI: 10.1007/s11263-019-01228-7. Accessed: Dec. 14, 2024. [Online]. Available: http://link.springer.com/10.1007/s11263-019-01228-7.

[29] J. Teuwen and N. Moriakov, "Convolutional neural networks," in *Handbook of Medical Image Computing and Computer Assisted Intervention*, Elsevier, 2020, pp. 481–501, ISBN: 978-0-12-816176-0. DOI: 10.1016/B978-0-12-816176-0.00025-9. Accessed: Oct. 29, 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/B9780128161760000259.

[30] P. Chen *et al.*, "An improved multi-input deep convolutional neural network for automatic emotion recognition," *Frontiers in Neuroscience*, vol. 16, p. 965 871, Oct. 4, 2022. DOI: 10.3389/fnins.2022.965871. pmid: 36267236. Accessed: Oct. 29, 2024. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC9577494/.

[31] L. Yu and N. Zhou. "Imbalance." arXiv: 2104.02240 [stat], Accessed: Nov. 18, 2024. [Online]. Available: http://arxiv.org/abs/2104.02240, pre-published.

[32] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 1, 2002, ISSN: 1076-9757. DOI: 10.1613/jair.953. Accessed: Nov. 20, 2024. [Online]. Available: https://www.jair.org/index.php/jair/article/view/10302.

[33] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports*, vol. 14, no. 1, p. 6086, Mar. 13, 2024, ISSN: 2045-2322. DOI: 10.1038/s41598-024-56706-x. Accessed: Nov. 4, 2024. [Online]. Available: https://www.nature.com/articles/s41598-024-56706-x.

[34] "Data Version Control · DVC," Data Version Control · DVC, Accessed: Nov. 1, 2024. [Online]. Available: https://dvc.org/.

[35] "CML · Continuous Machine Learning," CML · Continuous Machine Learning, Accessed: Nov. 1, 2024. [Online]. Available: https://cml.dev/.

[36] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, Boulder CO USA: ACM, Oct. 2, 2018, pp. 400–408, ISBN: 978-1-4503-5692-3. DOI: 10.1145/3242969.3242985. Accessed: Oct. 16, 2024. [Online]. Available: https://dl.acm.org/doi/10.1145/3242969.3242985.

[37] W. Li *et al.*, "A multimodal psychological, physiological and behavioural dataset for human emotions in driving tasks," *Scientific Data*, vol. 9, no. 1, p. 481, Aug. 6, 2022, ISSN: 2052-4463. DOI: 10.1038/s41597-022-01557-2. Accessed: Nov. 18, 2024. [Online]. Available: https://www.nature.com/articles/s41597-022-01557-2.

A Appendix: Datasets with Physiological Signals

Dataset	Sensor + Signals Available	Task/Target	Subjects	Annotations	Environment	Ref.
WESAD	Empatica E4: EDA, TEMP, ACC, BVP Respiban: ECG, EDA, RESP, EMG, ACC	Stress detection	15	Stress Baseline Amusement	Lab-based	[13]
AMIGOS	Shimmer 3 EDA, ECG Neurosky Mindwave EEG	Emotion recognition	40	Arousal Valance	Lab-based	[19]
DRIVE	Biopac MP150 ECG, EDA, EMG, RESP Vehicle sensor Eye Tracker	Stress detection during simulated driving	15	Stress	Simulated driving task	[37]
Affective Road	Empatica E4 ECG, EDA, TEMP, BVP GPS	Stress detection during driving	33	Stress Traffic-based	Real-world driving	[18]
CASE	Biopac MP150 ECG, EDA, PPG, RESP, TEMP Actigraph Accelerometer	Stress and emotion classification	50	Stress Arousal Valence	Mixed Real and Lab	[20]

Table 9: Summary of Identified Datasets with physiological signals

B Appendix: Signal Predictions with EDA, BVP and ACC

B.1 Signal Predictions for Subject S2

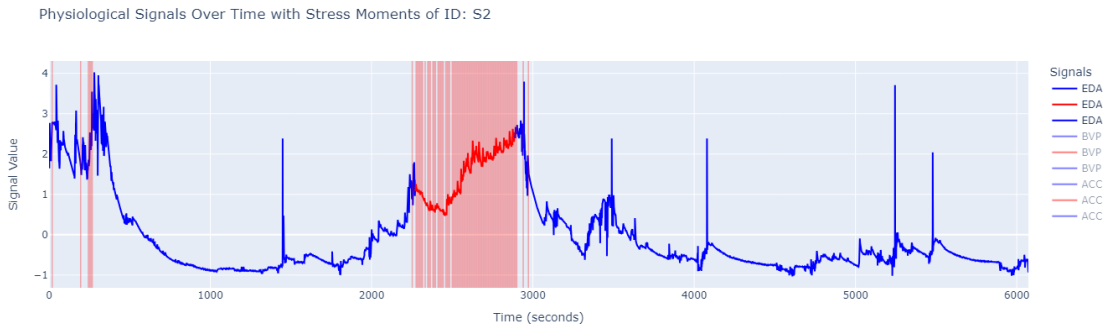

Figure 28: EDA signal plotted over subject S2 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

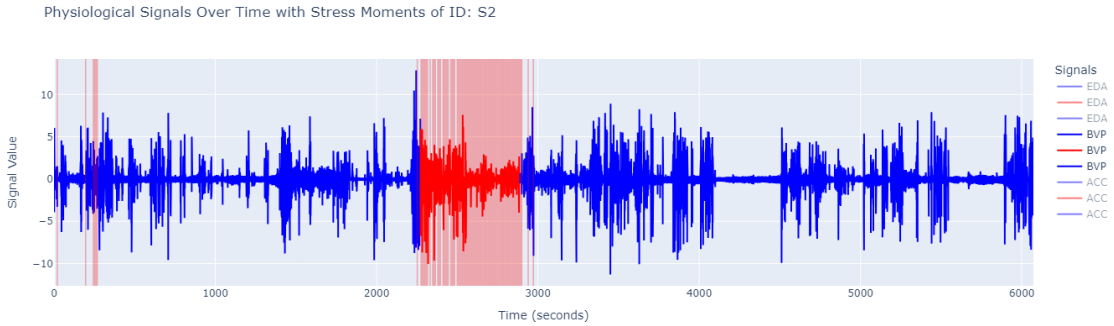

Figure 29: BVP signal plotted over subject S2 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

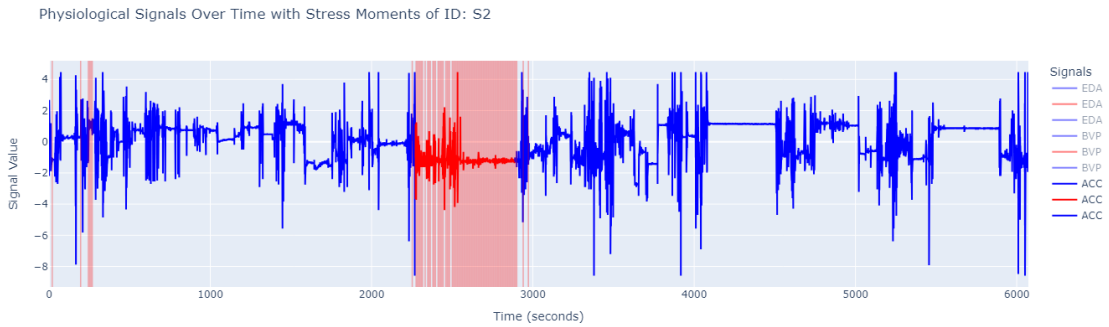

Figure 30: ACC signal plotted over subject S2 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

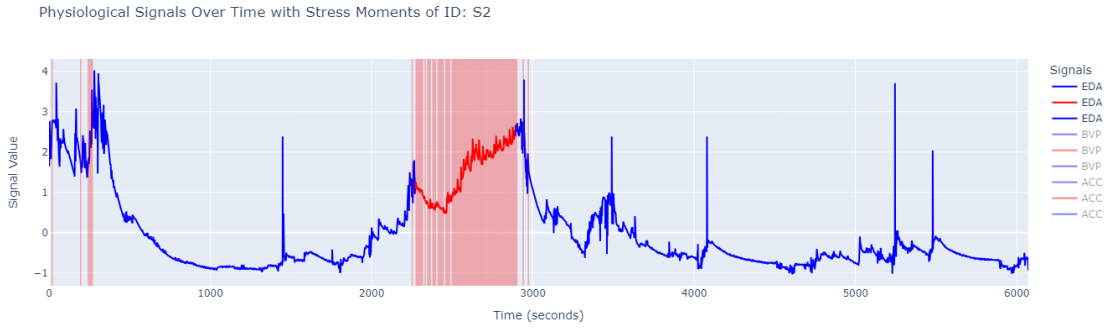

Figure 31: EDA signal plotted over subject S3 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

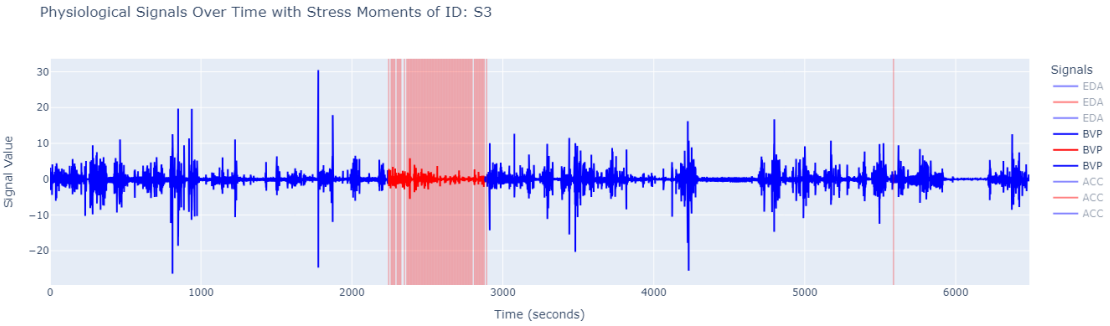

Figure 32: BVP signal plotted over subject S3 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

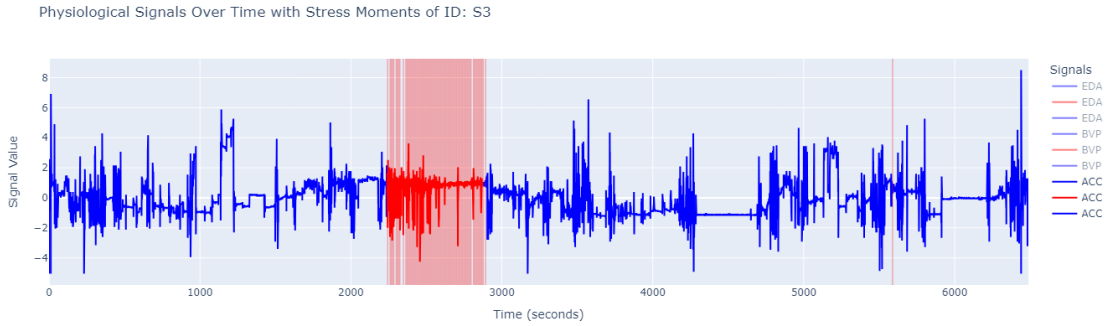

Figure 33: ACC signal plotted over subject S3 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

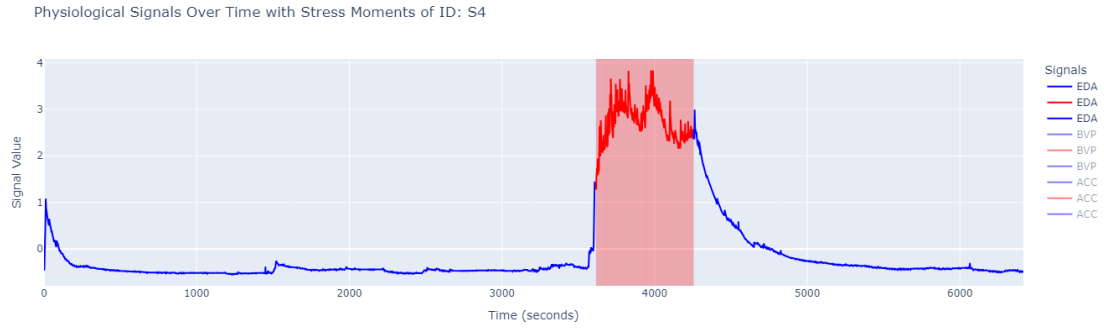

Figure 34: EDA signal plotted over subject S4 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

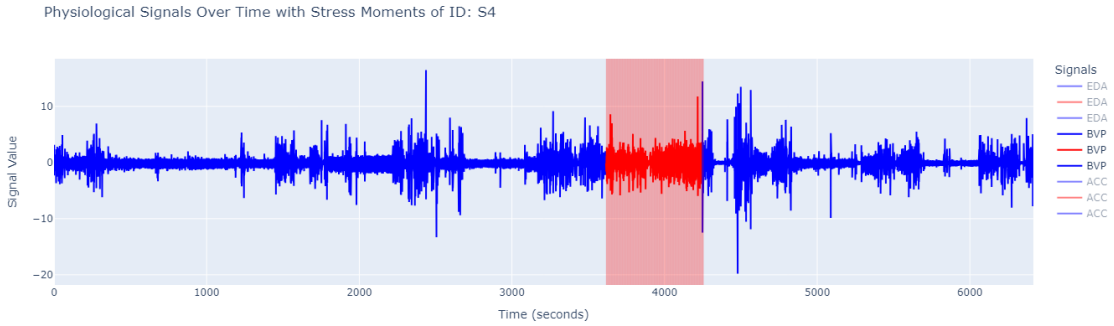

Figure 35: BVP signal plotted over subject S4 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

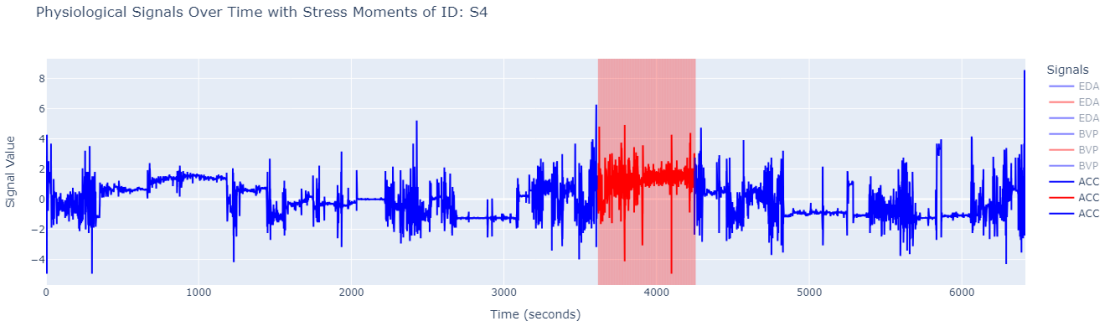

Figure 36: ACC signal plotted over subject S4 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

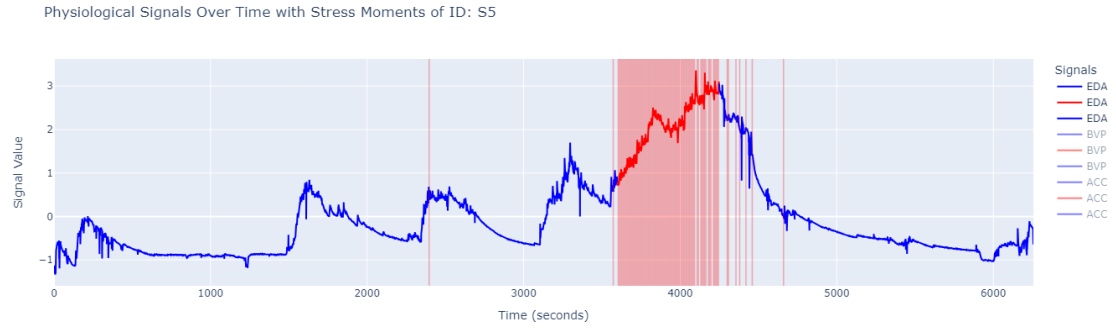

Figure 37: EDA signal plotted over subject S5 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

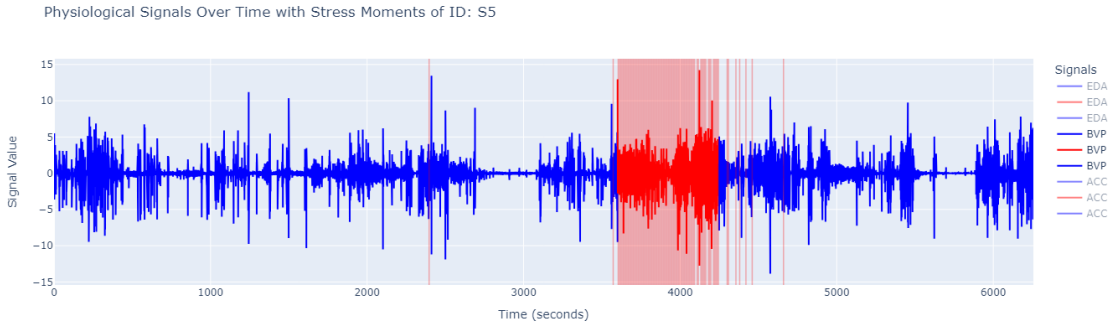

Figure 38: BVP signal plotted over subject S5 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

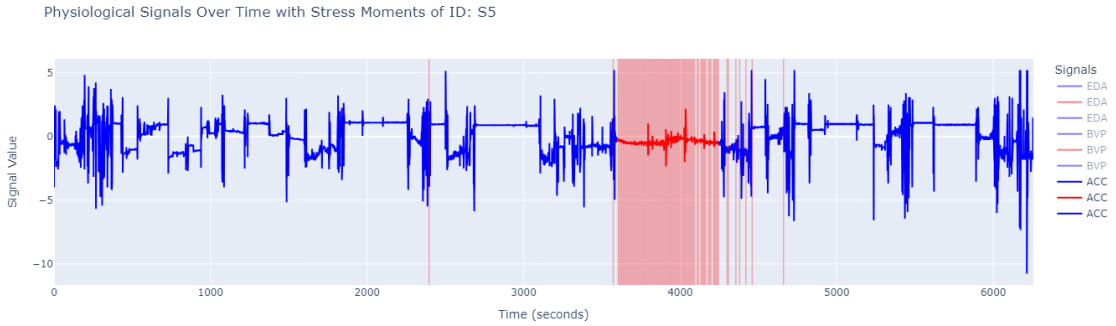

Figure 39: ACC signal plotted over subject S5 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

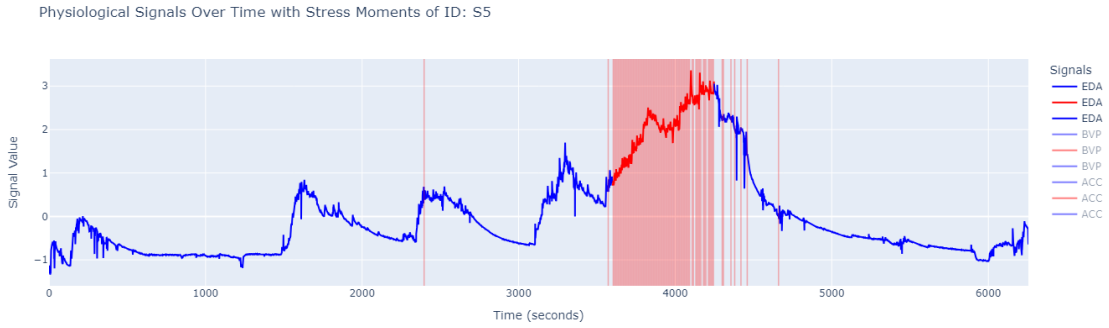

Figure 40: EDA signal plotted over subject S6 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

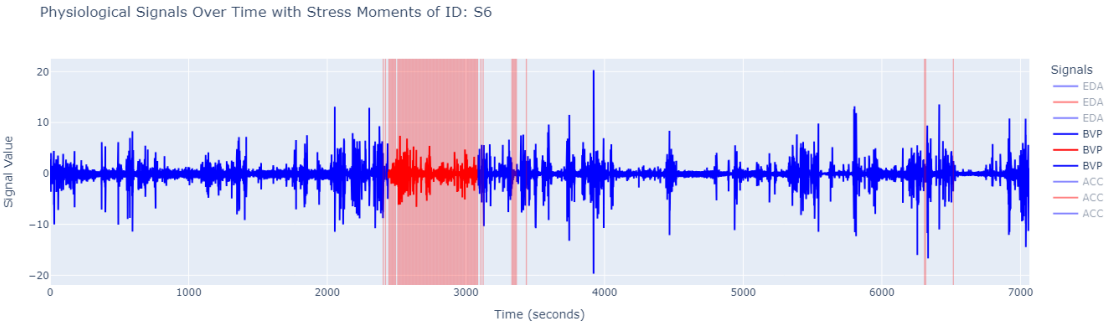

Figure 41: BVP signal plotted over subject S6 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

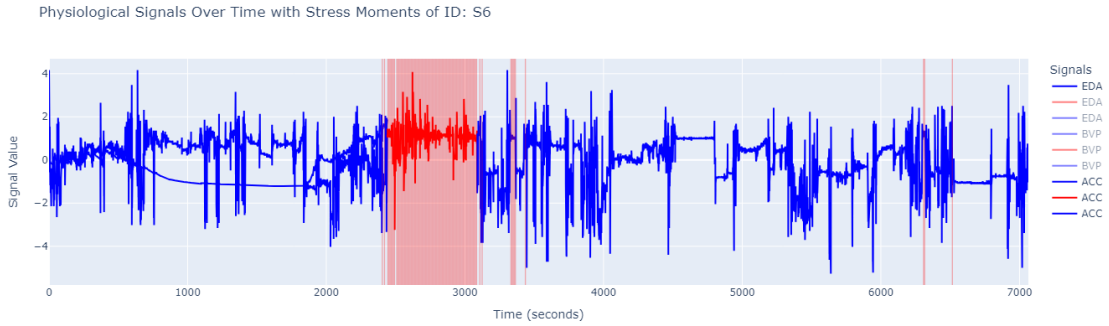

Figure 42: ACC signal plotted over subject S6 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

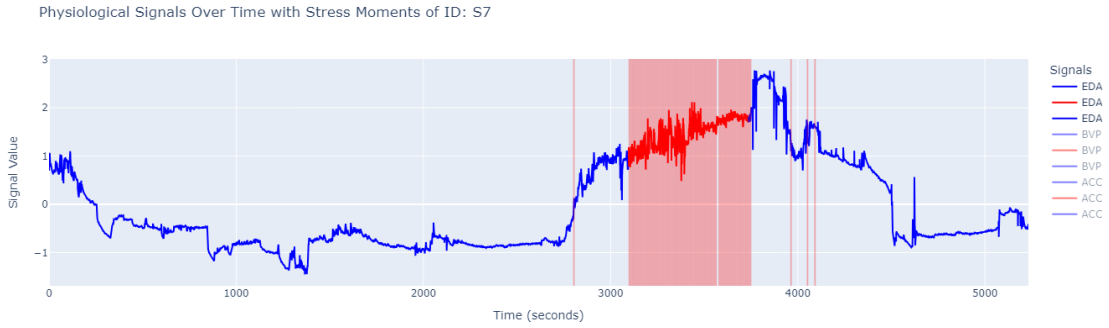

Figure 43: EDA signal plotted over subject S7 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

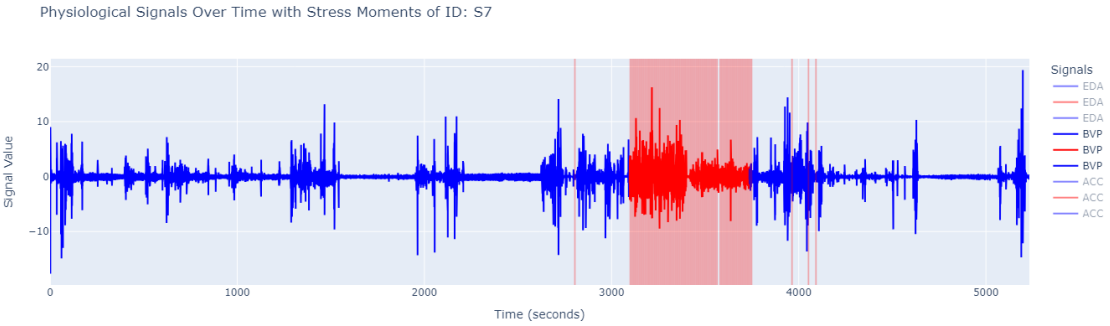

Figure 44: BVP signal plotted over subject S7 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

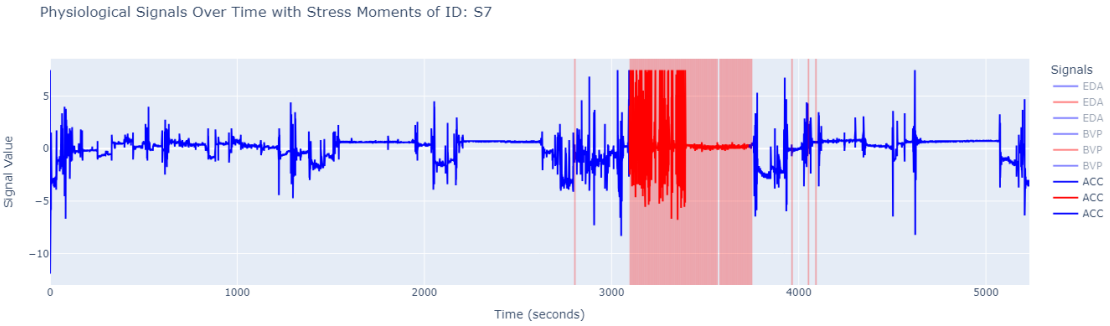

Figure 45: ACC signal plotted over subject S7 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

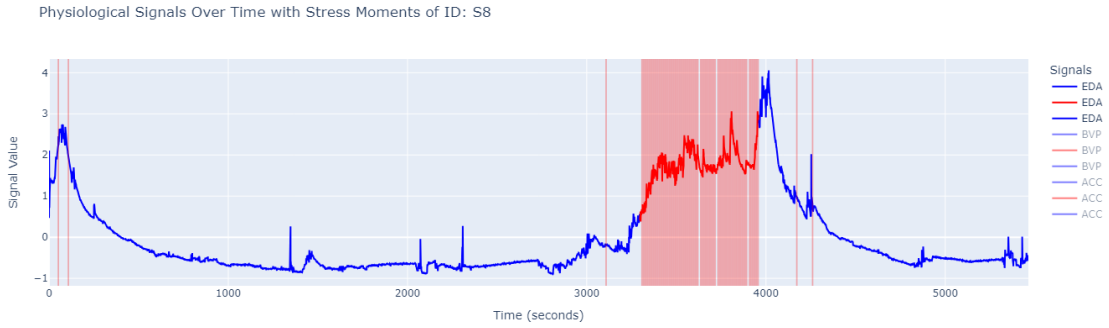

Figure 46: EDA signal plotted over subject S8 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

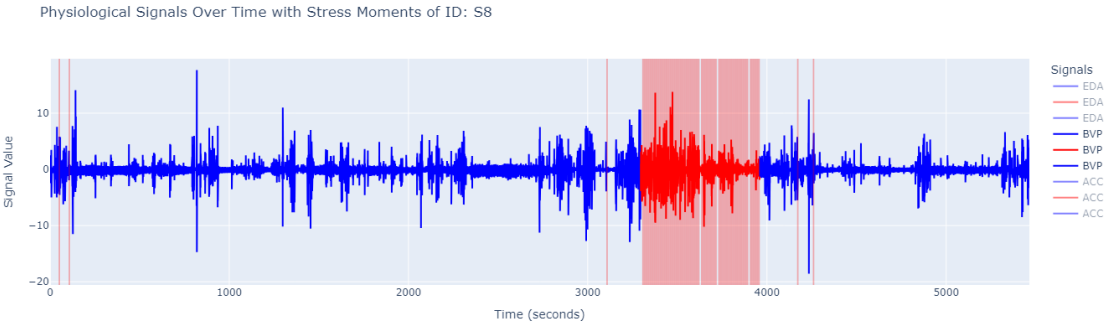

Figure 47: BVP signal plotted over subject S8 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

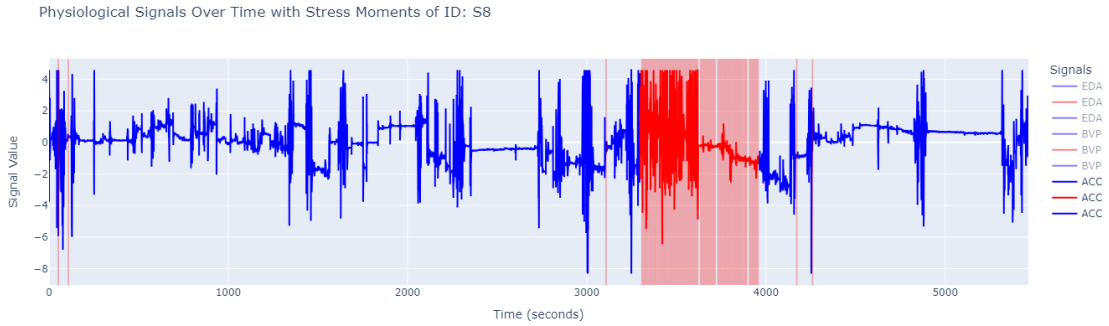

Figure 48: ACC signal plotted over subject S8 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

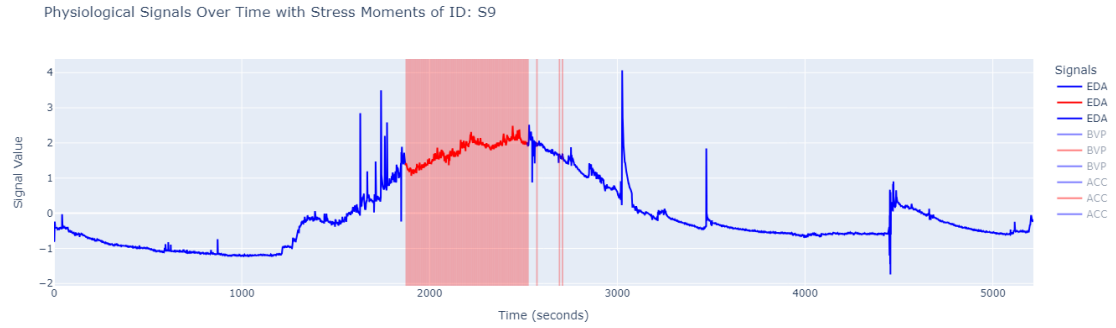

Figure 49: EDA signal plotted over subject S9 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

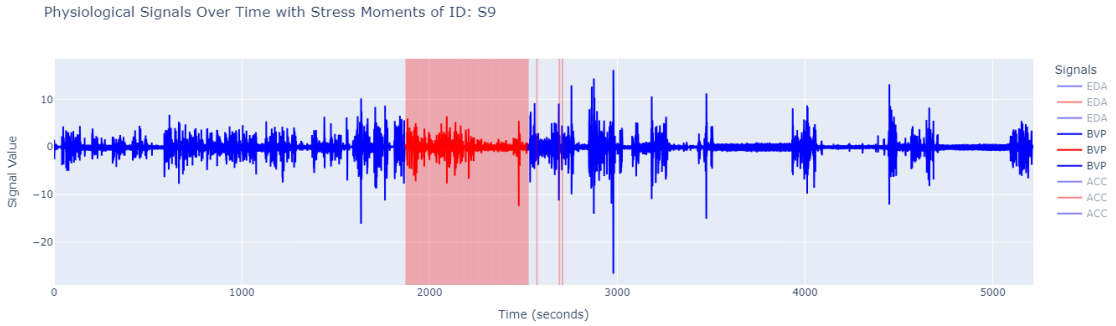

Figure 50: BVP signal plotted over subject S9 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

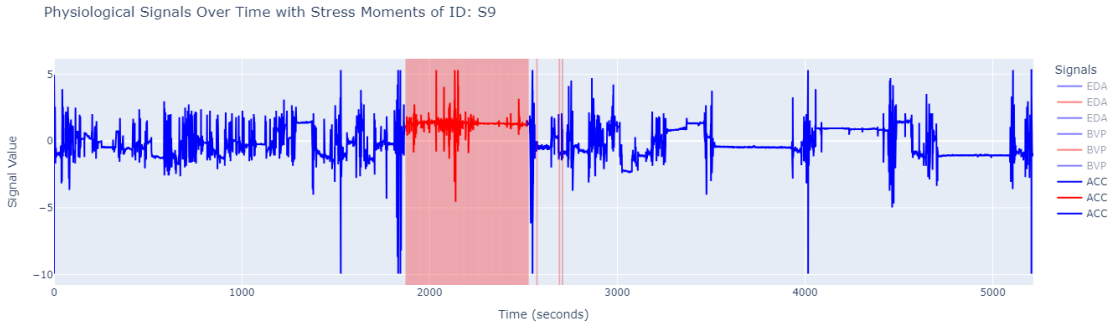

Figure 51: ACC signal plotted over subject S9 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

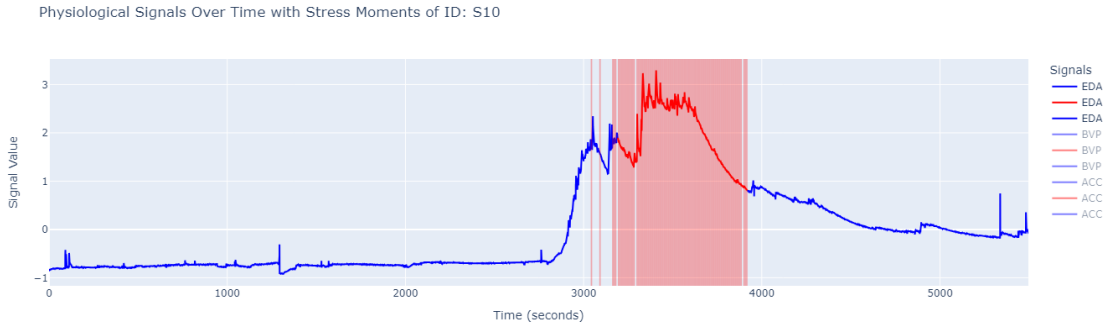

Figure 52: EDA signal plotted over subject S10 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

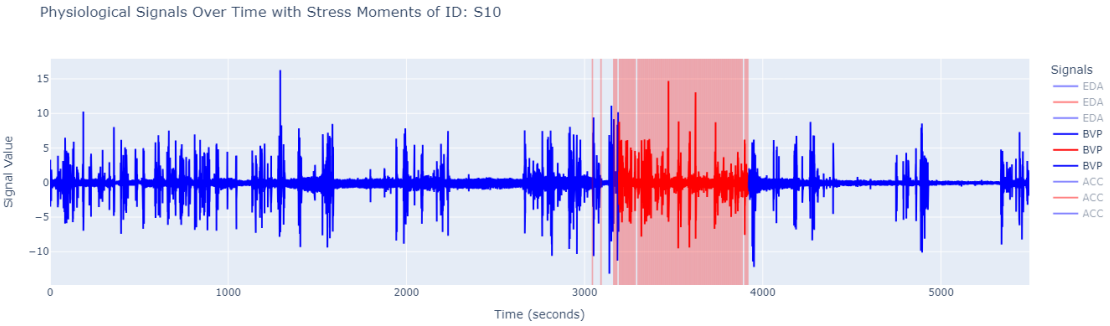

Figure 53: BVP signal plotted over subject S10 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

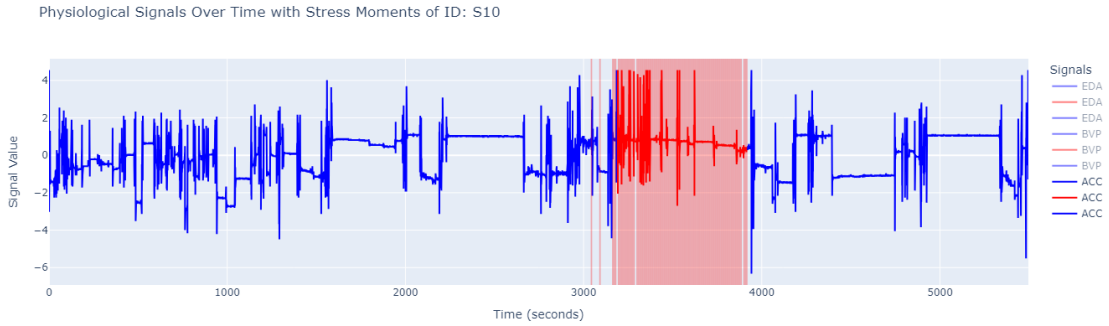

Figure 54: ACC signal plotted over subject S10 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

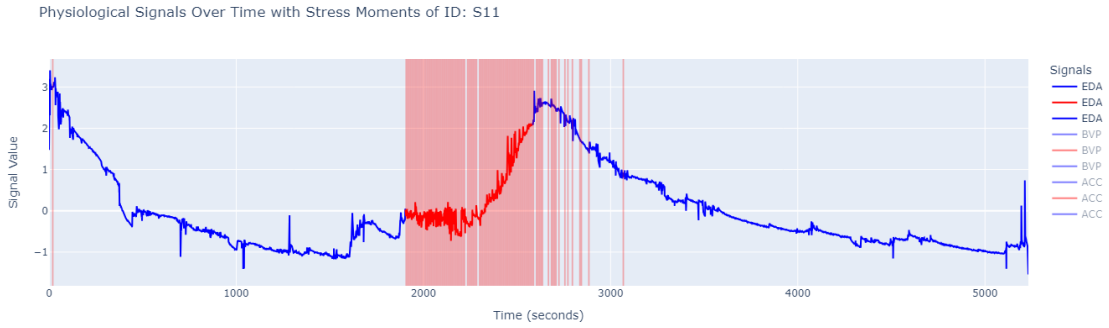

Figure 55: EDA signal plotted over subject S11 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

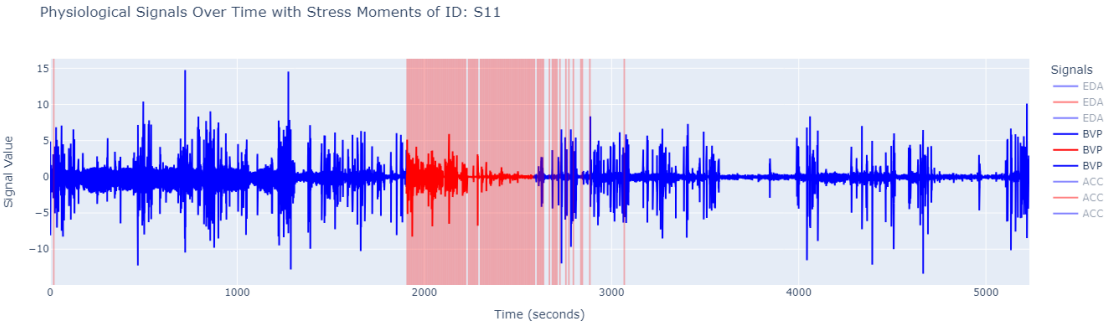

Figure 56: BVP signal plotted over subject S11 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

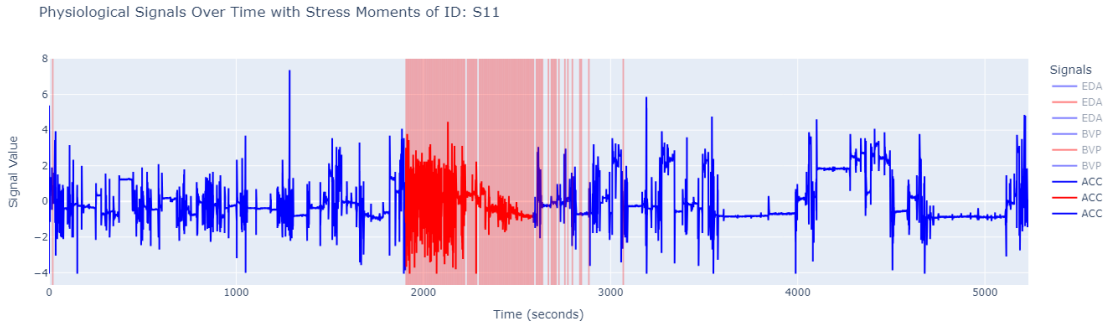

Figure 57: ACC signal plotted over subject S11 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

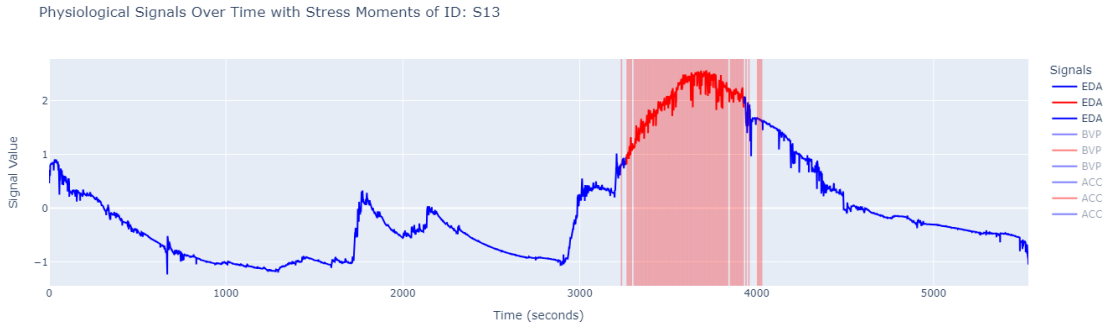

Figure 58: EDA signal plotted over subject S13 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

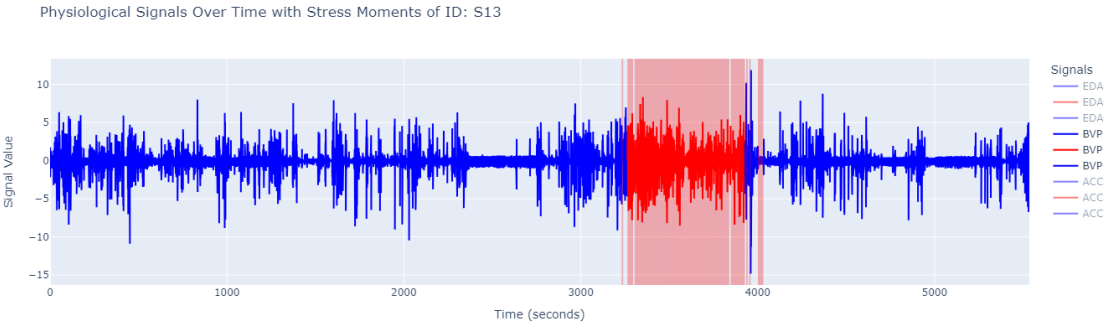

Figure 59: BVP signal plotted over subject S13 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

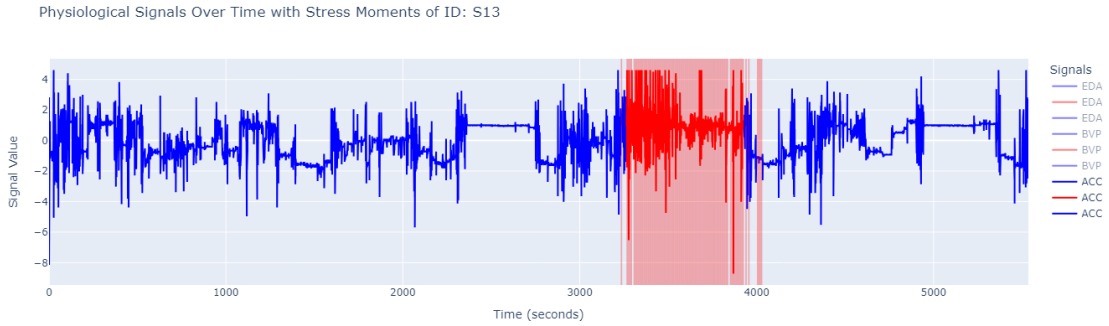

Figure 60: ACC signal plotted over subject S13 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

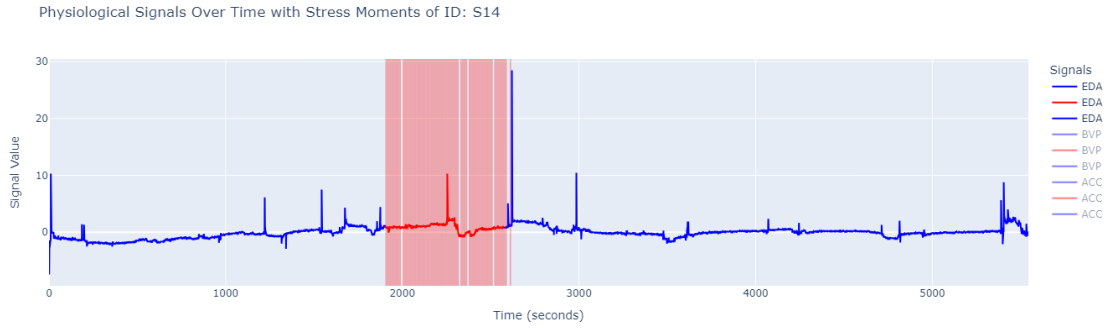

Figure 61: EDA signal plotted over subject S14 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

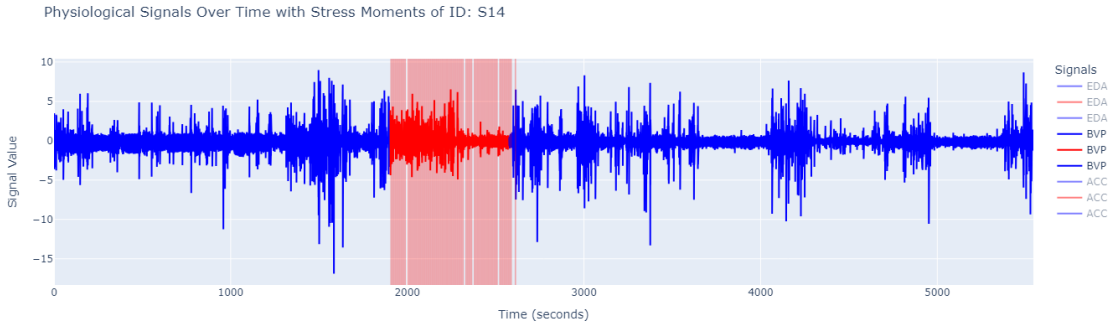

Figure 62: BVP signal plotted over subject S14 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

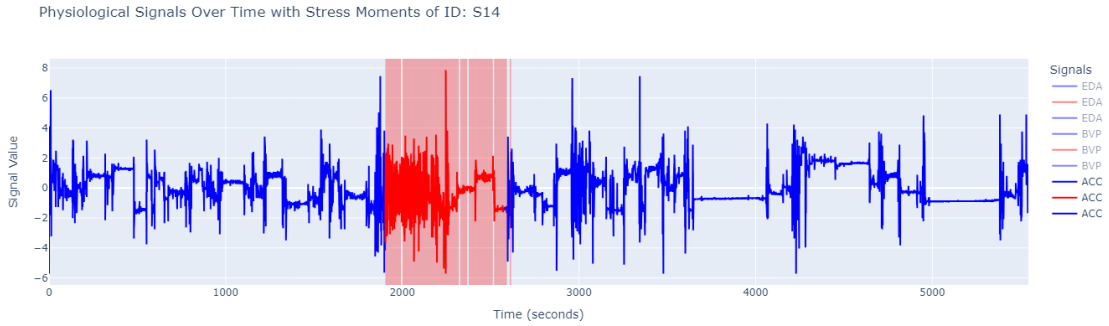

Figure 63: ACC signal plotted over subject S14 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

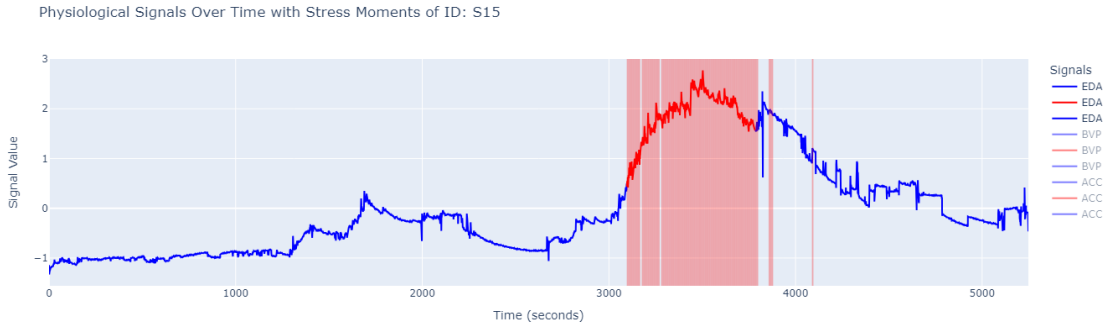

Figure 64: EDA signal plotted over subject S15 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

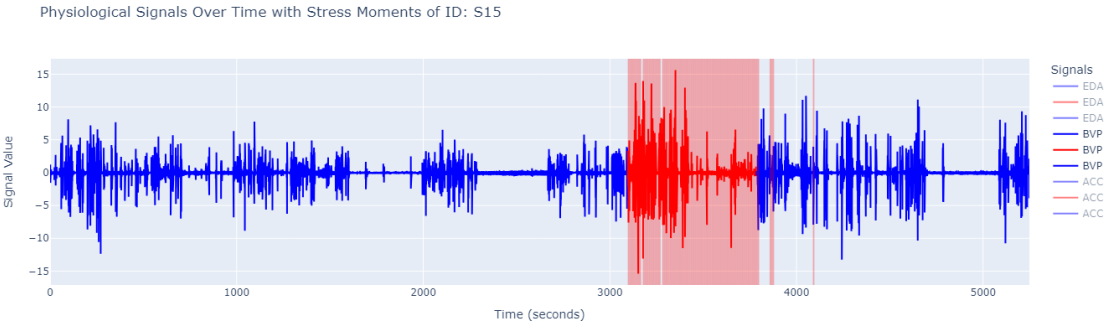

Figure 65: BVP signal plotted over subject S15 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

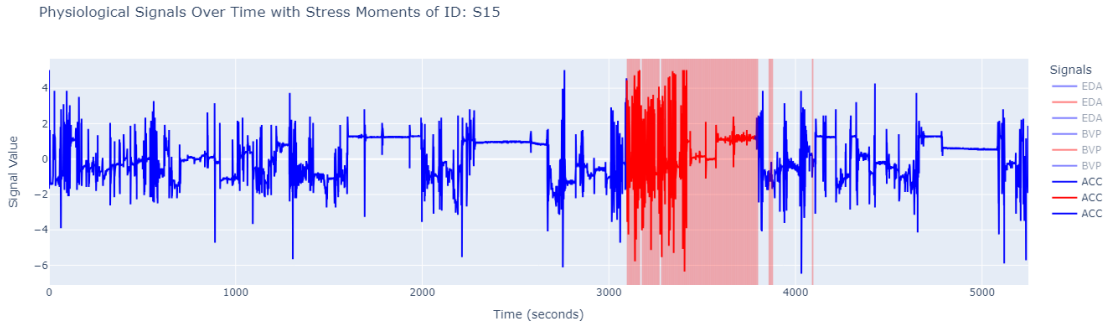

Figure 66: ACC signal plotted over subject S15 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

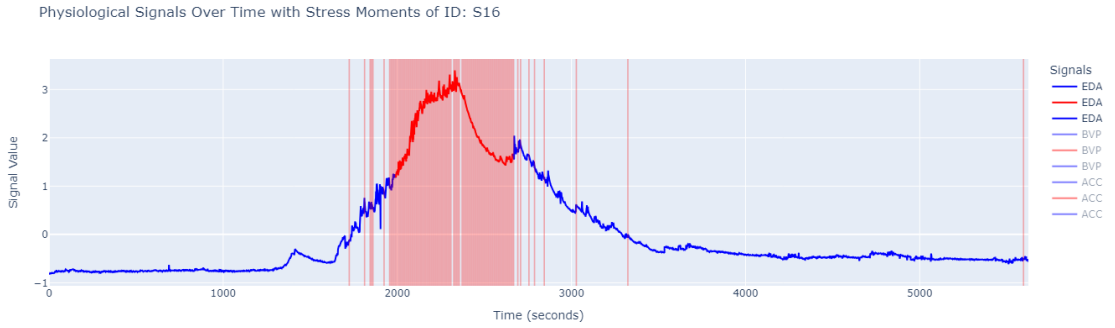

Figure 67: EDA signal plotted over test subject S16 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

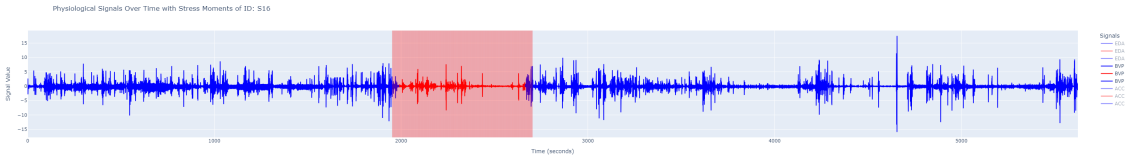

Figure 68: BVP signal plotted over test subject S16 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

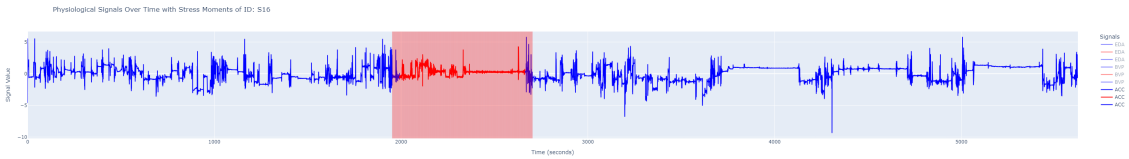

Figure 69: ACC signal plotted over test subject S16 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

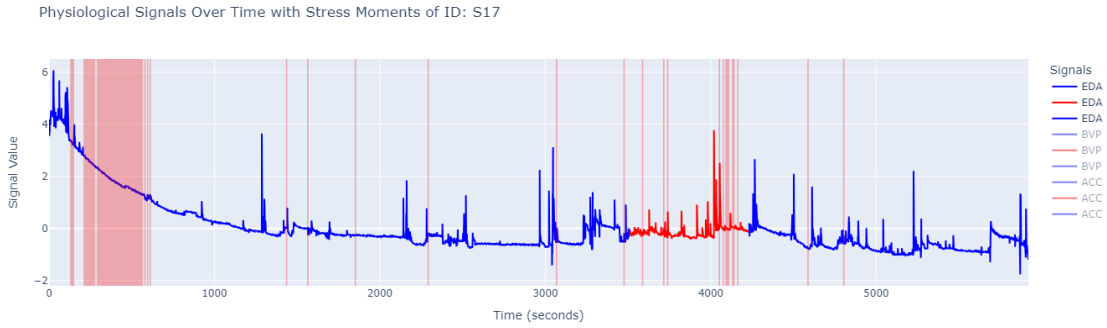

Figure 70: EDA signal plotted over test subject S17 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

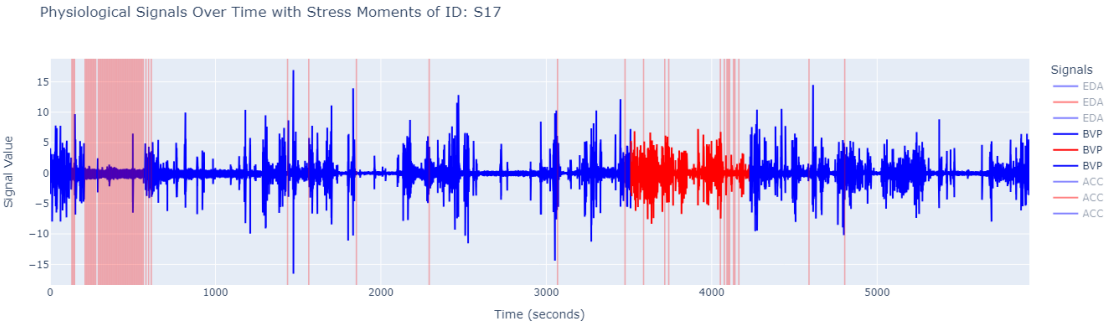

Figure 71: BVP signal plotted over test subject S17 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

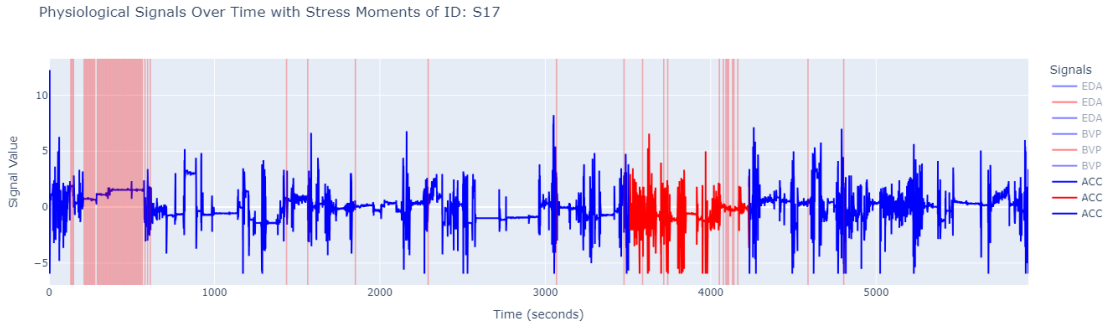

Figure 72: ACC signal plotted over test subject S17 for model with signal combination: EDA, BVP, and ACC. Blue line represents non-stress and red line represents stress. Red background represents predicted stress and white/transparent background represents predicted non-stress.

C Appendix: Confusion Matrix

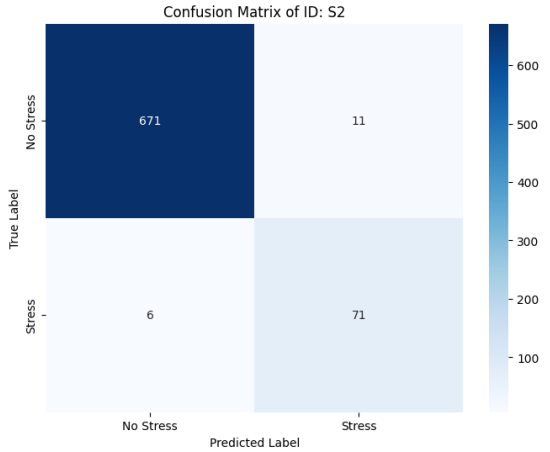

Figure 73: Confusion Matrix for Standard-Scaler Signal Predictions of Subject S2

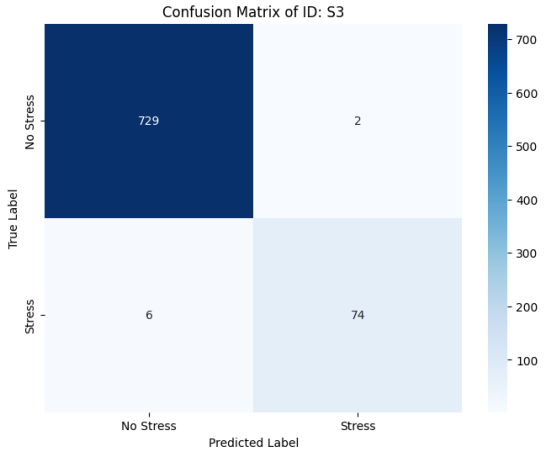

Figure 74: Confusion Matrix for Standard-Scaler Signal Predictions of Subject S3

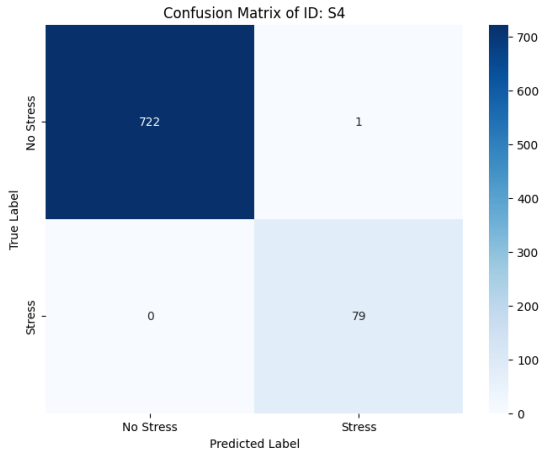

Figure 75: Confusion Matrix for Standard-Scaler Signal Predictions of Subject S4

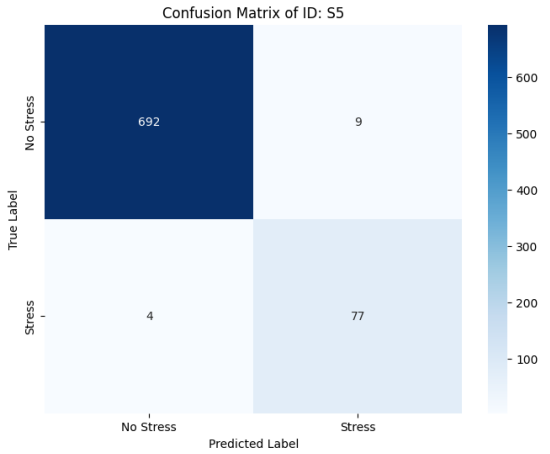

Figure 76: Confusion Matrix for Standard-Scaler Signal Predictions of Subject S5

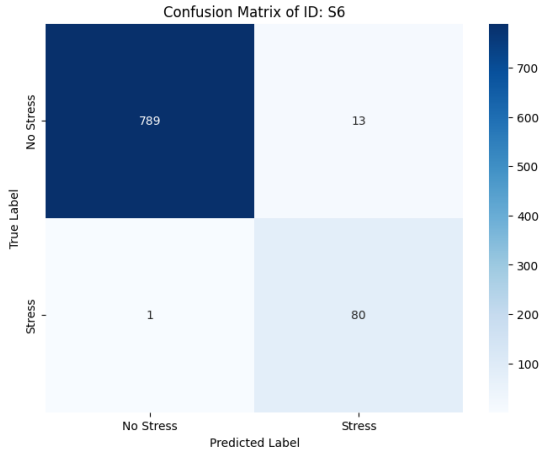

Figure 77: Confusion Matrix for Standard-Scaler Signal Predictions of Subject S6

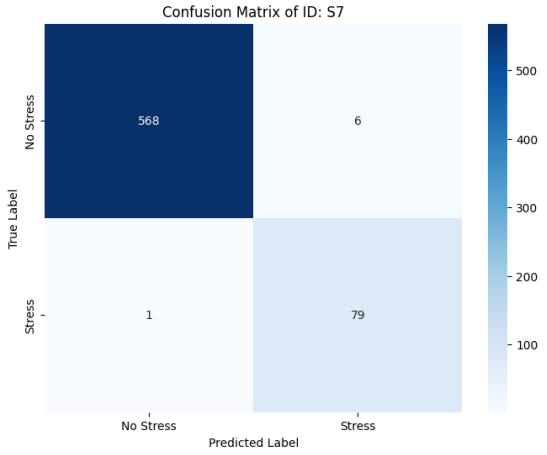

Figure 78: Confusion Matrix for Standard-Scaler Signal Predictions of Subject S7

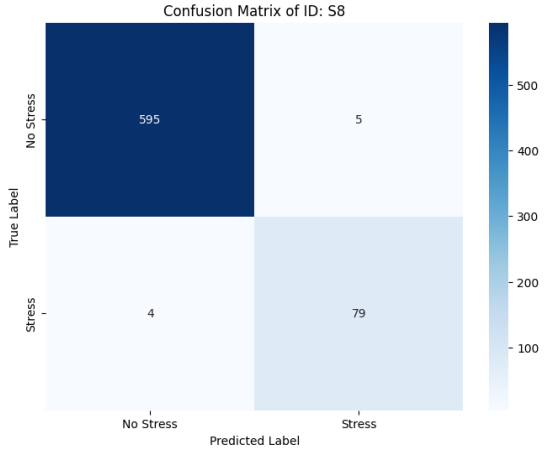

Figure 79: Confusion Matrix for Standard-Scaler Signal Predictions of Subject S8

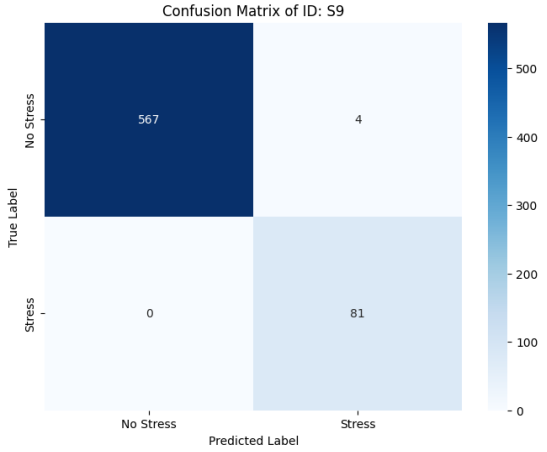

Figure 80: Confusion Matrix for Standard-Scaler Signal Predictions of Subject S9

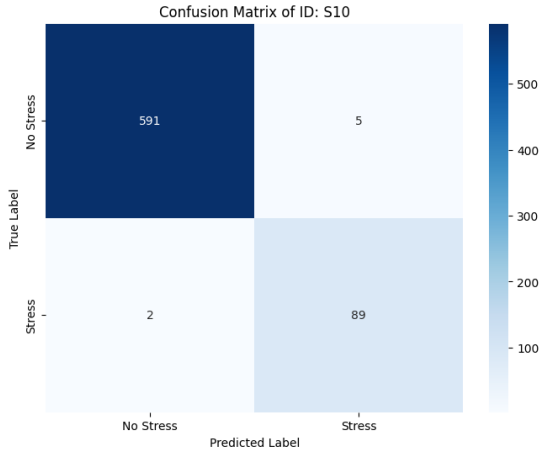

Figure 81: Confusion Matrix for Standard-Scaler Signal Predictions of Subject S10

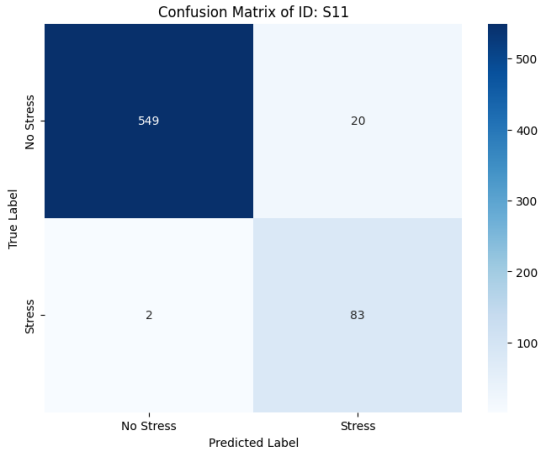

Figure 82: Confusion Matrix for Standard-Scaler Signal Predictions of Subject S11

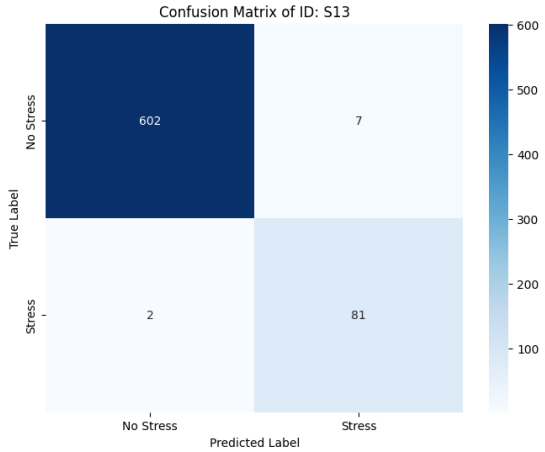

Figure 83: Confusion Matrix for Standard-Scaler Signal Predictions of Subject S13

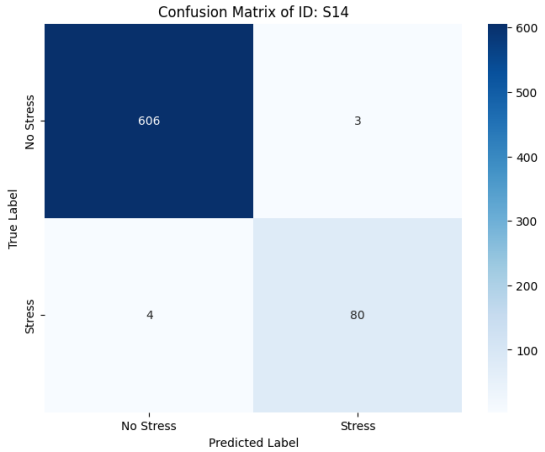

Figure 84: Confusion Matrix for Standard-Scaler Signal Predictions of Subject S14

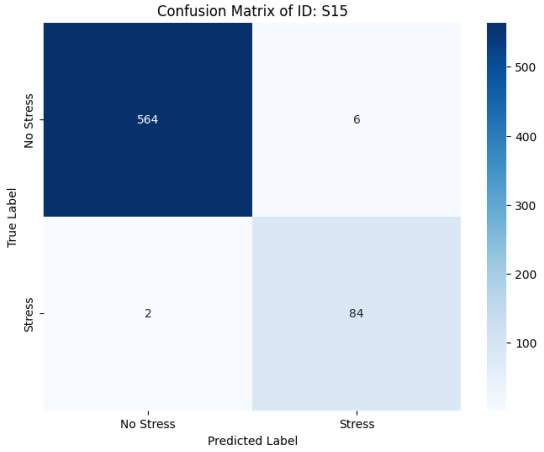

Figure 85: Confusion Matrix for Standard-Scaler Signal Predictions of Subject S15

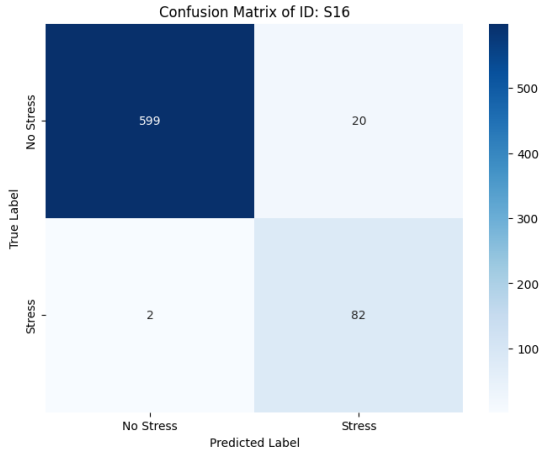

Figure 86: Confusion Matrix for Standard-Scaler Signal Predictions of Test Subject S16

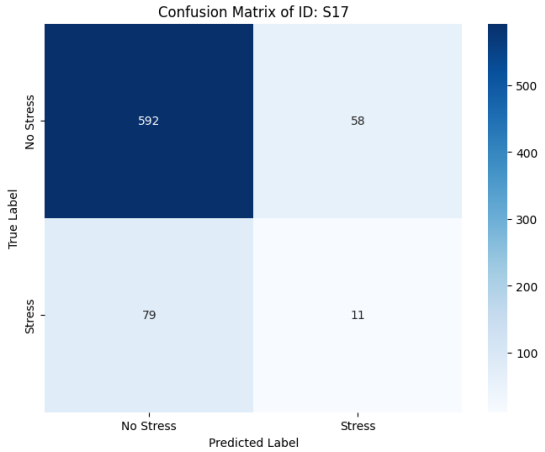

Figure 87: Confusion Matrix for Standard-Scaler Signal Predictions of Test Subject S17