Evaluating the Potential of Wearable Technology in Early Stress Detection: A Multimodal Approach

Basil A. Darwish[1], Nancy M. Salem[1], Ghada Kareem[2], Lamees N. Mahmoud[1], Ibrahim Sadek[1]

[1]Biomedical Engineering Department, Faculty of Engineering, Helwan University

[2]Department of Biomedical Engineering, Higher Technological Institute, 10th Ramadan City, Egypt

Abstract

Stress can adversely impact health, leading to issues like high blood pressure, heart diseases, and a compromised immune system. Consequently, using wearable devices to monitor stress is essential for prompt intervention and effective management. This study investigates the efficacy of wearable devices in the early detection of psychological stress, employing both binary and five-class classification models. Significant correlations were observed between stress levels and physiological signals, including Electrocardiogram (ECG), Electrodermal Activity (EDA), and Respiration (RESP), establishing these modalities as reliable biomarkers for stress detection. Utilizing the publicly available Wearable Stress and Affect Detection (WESAD) dataset, we employed two ensemble methods, Majority Voting (MV) and Weighted Averaging (WA), to integrate these signals, achieving maximum accuracies of 99.96% for binary classification and 99.59% for five-class classification. This integration significantly enhances the accuracy and robustness of the stress detection system. Furthermore, ten different classifiers were evaluated, and hyperparameter optimization and K-fold cross-validation ranging from 3-fold to 10-fold were applied. Both time-domain and frequency-domain features were examined separately. A review of commercially available wearable devices supporting these modalities was also conducted, resulting in recommendations for optimal configurations for practical applications. Our findings highlight the potential of multimodal wearable devices in advancing the early detection and continuous monitoring of psychological stress, with significant implications for future research and the development of improved stress detection systems.

Keywords: Wearable devices, psychological stress detection, Electrocardiogram (ECG), Electrodermal Activity (EDA), Respiration (RESP), ensemble methods, multimodal integration.

1. Introduction

Stress, a condition of mental strain or pressure due to upsetting conditions, is a major contributor to human physiology and pathophysiology. It has been linked to several conditions, such as autoimmune diseases, metabolic syndrome, sleep disorders, and suicidal thoughts and inclination [1] Over 70% of Americans regularly experience stress. Chronic stress can severely impact physical, mental, and social behaviors, potentially leading to numerous serious human disorders [2]. It is associated with the development of cancer, cardiovascular disease, depression, and diabetes, and thus is deeply detrimental to physiological health and psychological wellbeing [3], [4].

The response to stressful stimuli is based on a complex brain network, which requires well-tuned, functional neuroanatomical processing to detect and interpret the event as a potential threat to humans. However, the difficulty with the diagnosis and treatment of stress disorders lies in the complexity of the system and in the fact that stressors trigger different structures of the brain network [1]. Traditional diagnostic methods, such as self-reported questionnaires and physiological measurements, often lack the accuracy and objectivity needed for effective stress management [5]. Furthermore, these methods provide spontaneous measures, and they are not best suited for long-term monitoring.

Research indicates stress significantly contributes to cardiovascular disease by triggering pathophysiological changes such as increased sympathetic activation, heightened blood pressure, and inflammatory responses. These stress-induced mechanisms are particularly critical in individuals with pre-existing cardiovascular conditions, emphasizing the need for targeted preventive measures [6].

Given the prevalence and impact of stress, developing robust methods for the rapid and accurate detection of human stress is of paramount importance. This is where Artificial Intelligence (AI) and Machine Learning (ML) come into play. AI and ML have been able to predict stress and detect the brain's normal states vs. abnormal states (notably, in post-traumatic stress disorder (PTSD)) with an accuracy around 90% [1], [7]. Recent advancements in

ML techniques have further enhanced their predictive capabilities, making them valuable tools for stress detection [8].

Despite these advancements, several challenges and limitations exist in implementing ML for stress detection. Issues such as data privacy, the requirement for large and diverse datasets, and potential biases in algorithms need to be addressed [9]. Given that physiological signals have been demonstrated to be dependable indicators of stress [10], integrating multi-modal data (e.g., physiological signals, behavioral data, and environmental factors) could improve the robustness and accuracy of ML models [11].

The potential benefits of accurate stress detection using ML are extensive, including early intervention, personalized treatment plans, and improved mental health outcomes. To fully realize these benefits, future research should focus on refining ML algorithms, addressing ethical concerns, and developing user-friendly applications for both clinicians and patients [12]. In summary, while the application of ML in stress detection is promising, continuous research and development are crucial to overcome existing limitations and fully leverage these technologies to improve mental health care [13].

Given the critical importance of accurate stress detection and the potential of AI and ML technologies, it is essential to review the existing literature on these topics. The following section will provide a comprehensive examination of current research, highlighting key findings, methodological approaches, and the evolving landscape of AI and ML applications in stress detection. This review will also identify gaps in the literature and propose directions for future research, setting the stage for a deeper understanding of how these technologies can be harnessed to address the complexities of stress and its impact on human health.

2. Literature Review

Smets et al. (2016) [14] explores the application of machine learning techniques in detecting psychophysiological stress by analyzing various physiological signals. The research, conducted in a controlled laboratory setting, examines Electrocardiogram (ECG), Galvanic Skin Response (GSR), temperature, and respiration during a stress test. By comparing six different machine learning techniques, the study identifies personalized dynamic Bayesian networks as the most effective, achieving a prediction accuracy of 84.6%. The study's approach of using both general and personal models to enhance classification accuracy is noteworthy. However, the research is confined to a controlled environment, which may not accurately represent real-world stress conditions. Furthermore, the study does not investigate the impact of stressor intensity or the potential for real-time stress monitoring outside the laboratory setting. Despite these limitations, the findings underscore the potential of machine learning in improving the precision of stress detection tools and contributing to the development of responsive stress management strategies.

Gjoreski et al. (2017) [15] explored a novel method for detecting stress using a wrist-worn device in real-life settings, combining machine learning with a context-based approach. This study employed multiple classifiers, including a Support Vector Machine (SVM) classifier, which achieved a 71% accuracy within a six-minute window. It tested three-class classification and utilized leave-one-subject-out (LOSO) cross-validation to ensure robustness. The research integrated data from a laboratory stress detector, an activity recognizer, and a context-based stress detector, aiming to provide continuous and unobtrusive stress monitoring that adapts to various activities and environmental factors. Despite its promising approach, the study acknowledges the challenges of accurately detecting stress in diverse real-life conditions and suggests the need for further refinement and broader testing.

Schmidt et al. (2018) [16] provide a significant contribution to affect recognition by offering a publicly available dataset specifically tailored for wearable stress detection. This multimodal dataset includes physiological and motion data collected from both wrist- and chest-worn devices during a controlled lab study involving 15 subjects. Key modalities encompassed include blood volume pulse (BVP), electrocardiogram (ECG), electrodermal activity (EDA), electromyogram (EMG), respiration (RESP), body temperature (TEMP), and three-axis acceleration (ACCE). Notably, the dataset was designed to bridge gaps in available standard datasets by including multiple affective states, such as neutral, stress, and amusement, and supports the development of automated stress monitoring systems. To validate their approach, the researchers employed the LOSO cross-validation method. They report a classification accuracy of 93.12% using a linear discriminant analysis classifier (LDA) for binary stress detection. This high level of accuracy highlights the robustness of the dataset and the effectiveness of the

LDA classifier in stress and affect detection. However, the study does present certain limitations. It does not explore the impact of varying window sizes for data analysis, which could affect the performance and applicability of the classification algorithms in real-world scenarios where stress levels can fluctuate over different periods.

Can et al. (2019) [17] developed a portable stress detection system that utilizes physiological data gathered from discreet smart wearables. The system incorporates techniques for artifact removal and feature extraction tailored for real-world applications. They gathered physiological signals and questionnaire data from 21 participants using Samsung Gear S, S2, and Empatica E4 sensors. The analysis, employing a 10-fold cross-validation for robustness, revealed that the Multilayer Perceptron (MLP) algorithm yielded the highest accuracy of 92.19% using heart rate (HR) and ACCE data from the Empatica E4. However, a limitation of this study is its focus on binary stress classification, rather than employing a multi-class classification approach that could potentially offer a more nuanced understanding of stress levels.

Siirtola et al. (2020) [18] utilizes the publicly available AffectiveROAD dataset, which includes data collected via the Empatica E4 sensor. In their experiments, the integration of features from blood volume pulse (BVP) and skin temperature (SKT) yielded the most favorable results. Using LOSO cross-validation, the study achieved an average accuracy of 82.3% with a bagged tree-based ensemble. However, a notable limitation of the study is the absence of class-balancing techniques, which could potentially affect the reliability and generalizability of the findings.

Kaczor et al. (2020) [19] address physician stress in emergency medicine through the objective monitoring of physiological indicators and digital biomarkers using wearable sensors. Their innovative approach deviates from traditional subjective self-reports by employing machine learning algorithms, to determine stress episodes. The study achieves a prediction accuracy of 64.5% using a Naive Bayes classifier within a 20-minute pre-stress episode window. Notably, the research employs robust validation techniques, including 10-fold cross-validation and Receiver Operating Characteristic (ROC) curve analysis, and tests multiple classifiers to ensure reliability. However, the study is limited by its exclusive use of binary classification, failing to account for varying stress levels. It does not explore the effects of different time window sizes on prediction accuracy. Despite these limitations, the findings highlight the potential of wearable sensors in facilitating real-time stress interventions and improving the well-being of high-stress professionals.

Iqbal et al. (2021) [20] investigated the utility of wearable sensors for the real-time monitoring of stress. The study utilized the WESAD dataset. The study applied logistic regression to assess features derived from electromyography (EMG), electrodermal activity (EDA), respiration (RESP), and heart rate (HR). The logistic regression model achieved an accuracy of 85.71% in binary classification of stress states. This high level of accuracy demonstrates the potential of wearable devices in effectively identifying stress. Additionally, the validity of the approach was confirmed through 14-fold cross-validation, underlining the efficacy of logistic regression in processing complex bio physiological data for stress detection.

Iqbal et al. (2022) [21] delve into advanced strategies for continuous and real-time stress monitoring. Their research emphasizes the use of wearable sensor technology, with a particular focus on heart rate (HR) as a vital physiological parameter. Employing a Random Forest classifier integrated with 10-fold cross-validation, they analyze the SWELL-KW dataset. This technique not only enhances the robustness of their findings but also facilitates a significant binary classification of stress, achieving an accuracy of 75%.

Ehrhart et al. (2022) [22] highlight the significant advancements in human-centered applications leveraging wearable sensors and machine learning, particularly deep learning, for stress detection through physiological signals. They specifically utilized a stacked Long Short-Term Memory (LSTM) and a Fully Convolutional Network (FCN) classifier, employing features based on Galvanic Skin Response (GSR) and Skin Temperature (SKT), to achieve an impressive binary classification accuracy of 86.76% using tests on unseen cross-validation. The authors discuss the challenges of acquiring large, labeled datasets in this domain, which often leads to imbalanced data for training robust models. To address these limitations, they explored the use of a Conditional Generative Adversarial Network (cGAN) to augment the dataset, effectively enhancing its volume and diversity. This approach not only mitigated data imbalance but also significantly improved classifier performance, demonstrating that the synthetic data are indistinguishable from real data in their application. Despite these advancements, the study is limited by the absence of multi-class classification capabilities, which could potentially enhance the applicability of the stress detection models further.

Kuttala et al. (2023) [23] conducted a pivotal study on binary stress classification utilizing advanced deep learning techniques. The research spotlighted a critical issue: individuals experiencing stress often fail to recognize their stress levels, thus underlining the necessity for early and precise stress detection mechanisms. In their innovative approach, they harnessed multimodal hierarchical CNN feature fusion, significantly enhancing stress detection capabilities. This technique involved the integration of low, mid, and high-level features from Convolutional Neural Networks (CNNs), with concatenated multi-level CNN features for each of two key physiological signals: Electrodermal Activity (EDA) and Electrocardiogram (ECG). These features were then synergistically fused using the Multimodal Transfer Module (MMTM). Their comprehensive analysis spanned both raw frequency domain data and targeted frequency band features to ascertain the model's effectiveness. The empirical testing of the model across four benchmark datasets—ASCERTAIN, CLAS, MAUS, and WAUC—involved an initial training phase with 36, 18, 43, and 42 subject samples, respectively, followed by a testing phase comprising 9, 4, 16, and 16 subject samples from each dataset. The results were impressive, demonstrating high accuracies of 97.61% on ASCERTAIN, 95.94% on CLAS, 88.75% on MAUS, and 83.96% on WAUC. Despite these promising outcomes, the study acknowledged the need to expand the studies on hierarchical feature fusion and different multi-modal fusion techniques on hierarchical features.

Kalra et al. (2023) [24] conducted a study on pulse rate variability (PRV) using photoplethysmography (PPG) to monitor 15 subjects across five cognitive states: relaxation, deep breathing, and three varying levels of stress-related tasks. They discovered 18 significant features, split evenly between the time and frequency domains, which all showed statistical significance ($p < 0.05$) according to the Friedman test. Initially, a multi-layer perceptron (MLP) model was employed, achieving a classification accuracy of 85.1%±1.1%. This was further improved to 91%±1.1% when deep neural networks (DNN) were applied. A potential limitation is its broader focus on multiple cognitive states rather than exclusively on stress. This may dilute the specific insights and nuances related to stress detection and analysis, potentially impacting the specificity of the findings related to stress-related pulse rate variability.

Greco et al. (2023) [25] developed a novel methodology for detecting acute stress using only electrodermal activity (EDA) signals. This approach utilized a Support Vector Machine with a Recursive Feature Elimination algorithm (SVM-RFE) for classifying stress at an individual level. Employing a single-sensor system, the method demonstrated robustness to noise, incorporated rigorous phasic decomposition, and implemented unbiased multiclass classification. The methodology was tested on 65 volunteers subjected to various acute stress stimuli through a modified Trier Social Stress Test. For binary classification, the authors reported successful stress detection with an average accuracy of 94.62%. Furthermore, they proposed a four-class pattern recognition system capable of distinguishing between non-stressed states and three different stress conditions, achieving an average accuracy of 75% using leave-one-subject-out (LOSO) cross-validation. These results, obtained under controlled conditions, lay the groundwork for future applications in more ecological settings.

Richer et al. (2024) [26] explored the association between acute psychosocial stress and body movements using inertial measurement unit (IMU)-based motion capture suits. Data were gathered from 59 participants over two studies, in which participants experienced both the Trier Social Stress Test (TSST) and a control condition (friendly-TSST; f-TSST) in a randomized sequence. The research revealed a consistent freezing behavior in response to acute stress, marked by decreased overall movement and longer periods of immobility. Utilizing a Random Forest (RF) classifier with five-fold cross-validation, the study achieved a 73.4% accuracy in identifying acute stress from movement data. However, the study was limited by its binary classification approach, which did not address multiple stress levels. This study demonstrates the potential of using body posture and movement analysis as reliable indicators of acute psychosocial stress, presenting an alternative to conventional stress assessment methods.

In light of the gaps identified in the current literature, this paper aims to investigate the validity of using wearable devices for the early detection of psychological stress in both binary and five-class classifications. We employ machine learning techniques, testing various classifiers with optimized hyperparameters and cross-validation, using ECG, EDA, and RESP biometrics individually and in ensemble. Our study uniquely contributes by exploring five-class stress classification and an ensemble system using multiple biometrics, aiming to improve quality of life through more reliable and nuanced stress monitoring.

3. Methodology

To address the research gaps identified earlier, this study aims to explore the feasibility of robust psychological stress detection employing systematic signal segmentation and feature extraction to comprehensively characterize physiological responses, a rigorous machine learning pipeline involving feature selection, diverse classifiers, and hyperparameter optimization, a meticulous performance evaluation utilizing K-fold cross-validation (K-CV) as well as binary and multi-class detection. A comprehensive overview of our workflow is presented in Fig. 1.

3.1. Data Acquisition

This study employs the Wearable Stress and Affect Detection (WESAD) dataset, a publicly available resource for stress classification research (https://ubicomp.eti.uni-siegen.de/home/datasets/icmi18/) [16]. Within WESAD, data for 15 subjects was collected across three experimental conditions designed to elicit varying stress levels: baseline, stress, and amusement. This study utilizes an electrocardiogram (ECG), electrodermal activity (EDA), and respiration (RESP) collected from the chest-worn RespiBAN Professional device, all sampled at 700 Hz. The labels used in this study were from the Positive and Negative Affect Schedule questionnaire (PANAS) available in WESAD, more specifically the 21st item (Stressed) with its five possible responses (1 = Not at all, 2 = A little bit, 3 = Somewhat, 4 = Very much, 5 = Extremely) in case of multi-class. For binary classification, responses of 'Not at all' were considered class 0 (no stress) with remaining response levels considered class 1 (stressed).

3.2. Pre-processing

We utilized the BioSPPy library, an open-source tool for biosignal processing. This library provided us with robust and efficient algorithms for the analysis and filtering of Electrocardiogram (ECG), Electrodermal Activity (EDA), and Respiration (RESP) signals. For this study, we used the default parameters provided by the library. For additional information, documentation, and code examples, we recommend visiting the official BioSPPy GitHub repository (https://github.com/PIA-Group/BioSPPy).

To thoroughly investigate the impact of temporal signal length on stress classification, this study employed a strategic segmentation approach inspired by previous research [16], [27]. Window sizes were tested in increments, exploring durations of 60, 120, 210, 300, and 390 seconds. Additionally, to examine the effect of overlap, shifts of 10, 20, 30, 60, 120, 210, 300, and 390 seconds were applied to each window size appropriately to a total of 31 combinations, including the original unsegmented signal.

To ensure feature compatibility and improve machine learning model performance, the extracted features were normalized using Z-score. This process involved subtracting the mean and dividing by the standard deviation of each feature.

3.3. Feature Extraction

Feature extraction targeted both time-domain and frequency-domain characteristics across modalities. The statistical results were obtained directly from preprocessed signal segments. To analyze frequency patterns, a Fast Fourier Transform (FFT) and power spectral density (PSD) calculations were performed. The same set of statistical features was then derived from the power spectrum to enable comparative analysis across domains (Table 1)

Table 1: Extracted Features, including time and frequency domain features.

Modality	Time Domain Features	Frequency Domain Features
ECG/ EDA/ RESP	Mean	PSD Mean
	Variance	PSD Variance
	Standard Deviation	PSD Standard Deviation
	Median	PSD Median
	Maximum	PSD Maximum
	Minimum	PSD Minimum
	First Quartile	PSD First Quartile
	Third Quartile	PSD Third Quartile
	Skewness	PSD Skewness
	Kurtosis	PSD Kurtosis

3.4. Feature Selection

In our study, feature selection was performed using the Select From Model (SFM) method [28], which employs a Random Forest classifier as a meta-transformer. Specifically, we utilized a Random Forest with n=100 trees to determine feature importance scores, following a methodology analogous to that described in [26].

3.5. Classifiers and Hyperparameter Optimization

To ensure a comprehensive evaluation and the potential to uncover unexpected relationships; we tested ten different classifiers: Random Forest (RF), Extreme Gradient Boosting (XGB), k-nearest Neighbors (kNN), Logistic Regression (LR), Decision Tree (DT), AdaBoost (AB), Extra Trees (ET), Bagging (BAG), Quadratic Discriminant Analysis (QDA), and Linear Discriminant Analysis (LDA). We also applied hyperparameter optimization using grid search.

3.6. Class Distribution Balancing

Addressing class imbalance within the dataset was essential in mitigating potential classifier bias, Generative Adversarial Networks (GANs) have demonstrated effectiveness in creating novel data samples, making them a valuable technique for augmenting datasets. This augmentation is particularly beneficial for enhancing classifier performance on datasets that are small or imbalanced [22], [29]. Hence, we employed a multi-step Generative Adversarial Network (GAN) based data augmentation strategy. First, class representation was calculated, and classes exceeding the mean representation were truncated through random subsampling. Under-represented classes were targeted for augmentation with dedicated GANs [30]. These GANs were trained to learn underlying feature distributions, enabling the generation of synthetic samples mimicking the real data's characteristics. Augmentation continued until each under-represented class matched the mean class frequency, creating a more balanced dataset for training.

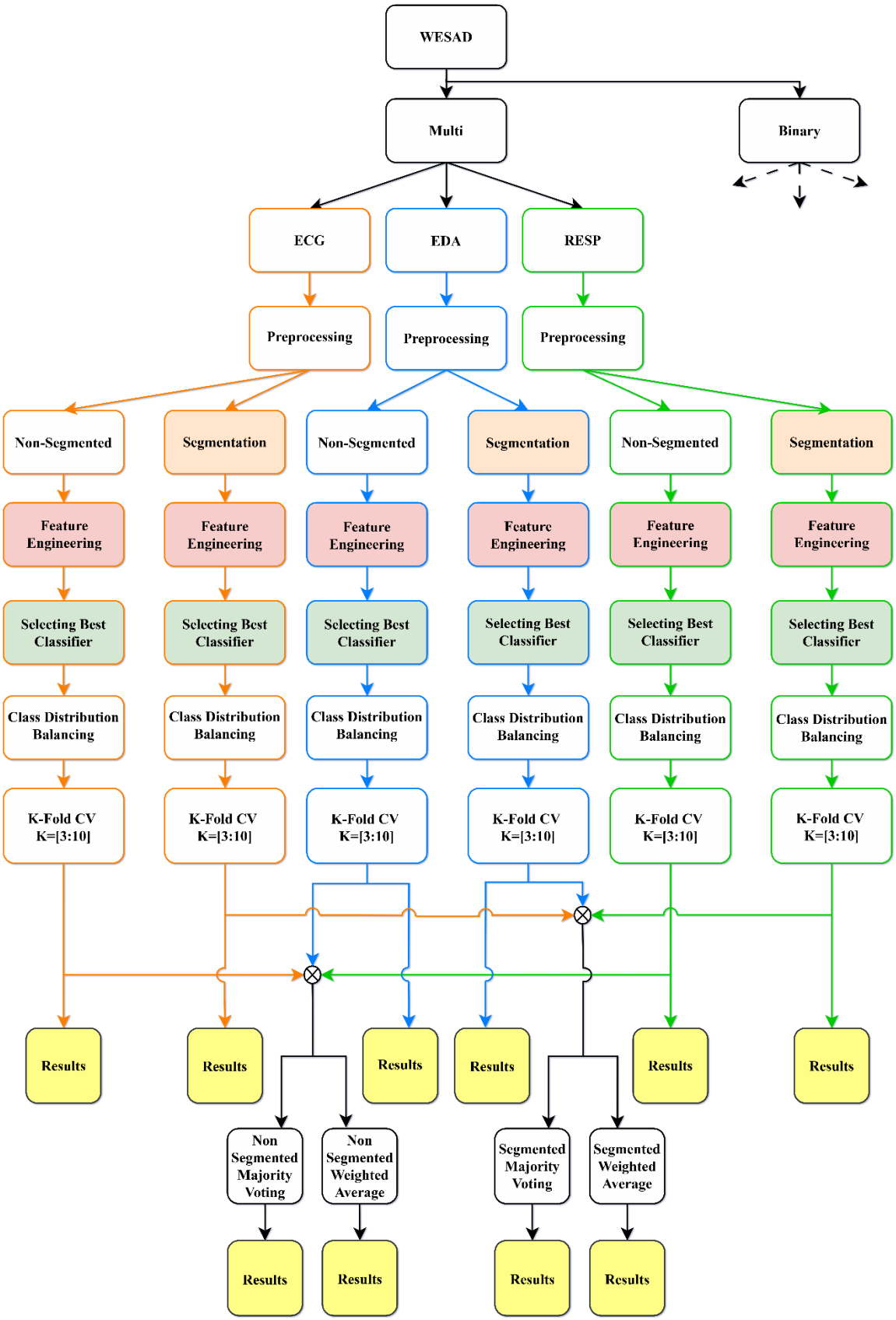

Fig. 1: Comprehensive binary and multi-class stress classification workflow for preprocessing, segmentation, feature engineering, and classifier selection with hyperparameter optimization of ECG, EDA, and RESP modalities from the WESAD dataset.

3.7. Evaluation

The best-performing optimized model for each modality was utilized. To ensure a robust evaluation and mitigate potential performance variance due to data splits, K-fold cross-validation (K-CV) with K values ranging from 3 to 10 was employed [31]. This approach provides a more reliable estimate of performance compared to a single train/test split, as it reduces the risk of overfitting and helps assess the model's generalization to unseen data. Ensemble methods, specifically majority voting (MV) and weighted averaging (WA), were applied to the outputs of ECG, EDA, and RESP. This investigation aimed to determine potential performance gains from a multi-modal approach compared to evaluating each modality independently. Accuracy (ACC) quantifies the proportion of correct predictions, while Precision (P) measures the accuracy of identifying positive labels correctly. Recall (R) indicates the percentage of actual positive cases the model successfully identifies. The F-measure (F1) score, the harmonic mean of precision and recall, provides a single metric balancing these two aspects. Additionally, the Area Under the Receiver Operating Characteristic Curve (AUC) was used to evaluate the model's performance, reflecting its ability to distinguish between positive and negative classes. These five evaluation metrics were calculated for each fold and averaged across K-CV runs, offering a comprehensive assessment of model performance. The equations are defined below.

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$P = \frac{TP}{TP + FP} \tag{2}$$

$$R = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{4}$$

4. Results

This study evaluated the efficacy of time-domain and frequency-domain statistical features including common features such as mean, variance, and others extracted from ECG, EDA, and RESP for the early detection of psychological stress (binary and multi-class classification). Ten different machine learning (ML) models were tested across these modalities to identify the most effective configurations for stress detection. K-fold cross-validation (K-CV) with K ranging from 3 to 10 was employed to enhance the generalizability and robustness of the results by evaluating performance on multiple non-overlapping data splits. To further leverage the information contained in each modality, we explored the performance of two ensemble methods: MV and WA. These methods combine predictions from individual modalities to potentially achieve improved classification.

4.1. Time-domain

Ensemble methods demonstrably outperformed individual modalities across various configurations for both binary and multiclass classification, as visually depicted in Fig. 2 and Fig. 3 respectively. Fig. 4 depicts the ROC curve for non-overlap WA average binary classification. Table 2 provides a consolidated overview of the peak performance achieved for each modality, classification type, and ensemble method.

4.1.1 Performance Evaluation of ECG Modality

As presented in Table 2, the performance of the ECG modality was assessed across different segmentation strategies and classifiers for binary classification. With unsegmented ECG data, the optimized XGB classifier achieved the highest accuracy of 69.31% using 8-CV. For segmentations with over 50% overlap, a window of 300 seconds with a 10-second shift yielded the highest accuracy of 99.63% using an optimized BAG classifier and 8-CV. In cases of segmentation with 50% and under overlap, a window of 390 seconds with a 210-second shift resulted in the highest accuracy of 89.94% with an optimized ET classifier and 8-CV. For segmentation with no overlap, a window of 120 seconds provided the highest accuracy of 80.34% using an optimized XGB classifier and 7-CV. These results are illustrated in Fig. 2, providing a visual summary of the performance across different segmentation strategies.

As detailed in Table 2, the evaluation of the ECG modality for multi-class classification revealed varying performance across different segmentation approaches and classifiers. With unsegmented ECG data, the optimized BAG classifier achieved the highest accuracy of 53.38% using 7-CV. For segmentations with over 50% overlap, a window of 390 seconds with a 10-second shift yielded the highest accuracy of 96.03% using an optimized XGB classifier and 9-CV. In cases of segmentation with 50% and under overlap, a window of 120 seconds with a 60-second shift resulted in the highest accuracy of 76.65% with an optimized XGB classifier and 9-CV. For segmentation with no overlap, a window of 120 seconds provided the highest accuracy of 68.72% using an optimized RF classifier and 9-CV. Fig. 3 visually summarizes these results, highlighting the performance variations across different segmentation strategies.

4.1.2 Performance Evaluation of EDA Modality

As presented in Table 2, the performance of the EDA modality was assessed across different segmentation strategies and classifiers for binary classification. With unsegmented EDA data, the optimized RF classifier achieved the highest accuracy of 68.75% using 10-CV. For segmentations with over 50% overlap, a window of 300 seconds with a 10-second shift yielded the highest accuracy of 99.84% using an optimized ET classifier and 10-CV. In cases of segmentation with 50% and under overlap, a window of 390 seconds with a 210-second shift resulted in the highest accuracy of 92.35% with an optimized kNN classifier and 6-CV. For segmentation with no overlap, a window of 120 seconds provided the highest accuracy of 86.79% using an optimized kNN classifier and 6-CV. These results are illustrated in Fig. 2, providing a visual summary of the performance across different segmentation strategies.

As detailed in Table 2, the evaluation of the EDA modality for multi-class classification revealed varying performance across different segmentation approaches and classifiers. With unsegmented EDA data, the optimized XGB classifier achieved the highest accuracy of 46.78% using 4-CV. For segmentations with over 50% overlap, a window of 300 seconds with a 10-second shift yielded the highest accuracy of 98.16% using an optimized ET classifier and 9-CV. In cases of segmentation with 50% and under overlap, a window of 60 seconds with a 30-second shift resulted in the highest accuracy of 81.12% with an optimized kNN classifier and 6-CV. For segmentation with no overlap, a window of 60 seconds provided the highest accuracy of 76.44% using an optimized ET classifier and 4-CV. Fig. 3 visually summarizes these results, highlighting the performance variations across different segmentation strategies.

4.1.3 Performance Evaluation of RESP Modality

As presented in Table 2, the performance of the RESP modality was assessed across different segmentation strategies and classifiers for binary classification. With unsegmented RESP data, the optimized LDA classifier achieved the highest accuracy of 60.71% using 10-CV. For segmentations with over 50% overlap, a window of 300 seconds with a 10-second shift yielded the highest accuracy of 99.43% using an optimized XGB classifier and 9-CV. In cases of segmentation with 50% and under overlap, a window of 300 seconds with a 210-second shift resulted in the highest accuracy of 84.13% with an optimized BAG classifier and 6-CV. For segmentation with no overlap, a window of 300 seconds provided the highest accuracy of 77.74% using an optimized RF classifier and 6-CV. These results are illustrated in Fig. 2, providing a visual summary of the performance across different segmentation strategies.

As detailed in Table 2, the evaluation of the RESP modality for multi-class classification revealed varying performance across different segmentation approaches and classifiers. With unsegmented RESP data, the optimized kNN classifier achieved the highest accuracy of 44.29% using 10-CV. For segmentations with over 50% overlap, a window of 390 seconds with a 10-second shift yielded the highest accuracy of 94.28% using an optimized XGB classifier and 7-CV. In cases of segmentation with 50% and under overlap, a window of 60 seconds with a 30-second shift resulted in the highest accuracy of 66.76% with an optimized ET classifier and 8-CV. For segmentation with no overlap, a window of 300 seconds provided the highest accuracy of 70.82% using an optimized RF classifier and 6-CV. Fig. 3 visually summarizes these results, highlighting the performance variations across different segmentation strategies.

4.1.4 Performance Evaluation of Modalities MV Ensemble Method

As presented in Table 2, the performance of the modalities MV ensemble was assessed across different segmentation strategies and classifiers for binary classification. With unsegmented MV Ensemble data, the optimized XGB, RF, and LDA classifiers for ECG, EDA, and RESP respectively achieved the highest accuracy

of 74.17% using 8-CV. For segmentations with over 50% overlap, a window of 300 seconds with a 10-second shift yielded the highest accuracy of 99.96% using optimized BAG, ET, and XGB classifiers for ECG, EDA, and RESP respectively, and 4-CV. In cases of segmentation with 50% and under overlap, a window of 210 seconds with a 120-second shift resulted in the highest accuracy of 92.41% with optimized RF, DT, and kNN classifiers for ECG, EDA, and RESP respectively, and 5-CV. For segmentation with no overlap, a window of 210 seconds provided the highest accuracy of 89.88% using optimized BAG, kNN, and QDA classifiers for ECG, EDA, and RESP respectively, and 4-CV. These results are illustrated in Fig. 2, providing a visual summary of the performance across different segmentation strategies.

As detailed in Table 2, the evaluation of the modalities MV ensemble for multi-class classification revealed varying performance across different segmentation approaches and classifiers. With unsegmented MV Ensemble data, the optimized BAG, XGB, and kNN classifiers for ECG, EDA, and RESP respectively achieved the highest accuracy of 57.33% using 5-CV. For segmentations with over 50% overlap, a window of 300 seconds with a 10-second shift yielded the highest accuracy of 99.43% using optimized ET, ET, and XGB classifiers for ECG, EDA, and RESP respectively, and 9-CV. In cases of segmentation with 50% and under overlap, a window of 60 seconds with a 30-second shift resulted in the highest accuracy of 81.82% with optimized XGB, kNN, ET classifiers for ECG, EDA, and RESP respectively, and 10-CV. For segmentation with no overlap, a window of 60 seconds provided the highest accuracy of 73.70% using optimized XGB, ET, and kNN classifiers for ECG, EDA, and RESP respectively, and 7-CV. Fig. 3 visually summarizes these results, highlighting the performance variations across different segmentation strategies.

4.1.5 Performance Evaluation of Modalities WA Ensemble Method

As presented in Table 2, the performance of the modalities WA ensemble was assessed across different segmentation strategies and classifiers for binary classification. With unsegmented WA Ensemble data, the optimized XGB, RF, and LDA classifiers for ECG, EDA, and RESP respectively achieved the highest accuracy of 74.10% using 5-CV. For segmentations with over 50% overlap, a window of 300 seconds with a 10-second shift yielded the highest accuracy of 99.96% using optimized BAG, XGB, and kNN classifiers for ECG, EDA, and RESP respectively, and 4-CV. In cases of segmentation with 50% and under overlap, a window of 390 seconds with a 210-second shift resulted in the highest accuracy of 94.70% with optimized ET, kNN, LR classifiers for ECG, EDA, and RESP respectively, and 6-CV. For segmentation with no overlap, a window of 210 seconds provided the highest accuracy of 91.67% using optimized BAG, kNN, and QDA classifiers for ECG, EDA, and RESP respectively and 8-CV, additionally ROC is depicted in Fig. 4. These results are illustrated in Fig. 2, providing a visual summary of the performance across different segmentation strategies.

As detailed in Table 2, the evaluation of the modalities WA ensemble for multi-class classification revealed varying performance across different segmentation approaches and classifiers. With unsegmented WA Ensemble data, the optimized BAG, XGB, and kNN classifiers for ECG, EDA, and RESP respectively achieved the highest accuracy of 65.28% using 8-CV. For segmentations with over 50% overlap, a window of 300 seconds with a 10-second shift yielded the highest accuracy of 99.59% using optimized ET, ET, and XGB classifiers for ECG, EDA, and RESP respectively, and 9-CV. In cases of segmentation with 50% and under overlap, a window of 120 seconds with a 60-second shift resulted in the highest accuracy of 87.94% with optimized XGB, kNN, kNN classifiers for ECG, EDA, and RESP respectively, and 9-CV. For segmentation with no overlap, a window of 120 seconds provided the highest accuracy of 81.67% using optimized RF, BAG, and kNN classifiers for ECG, EDA, and RESP respectively, and 8-CV. Fig. 3 visually summarizes these results, highlighting the performance variations across different segmentation strategies.

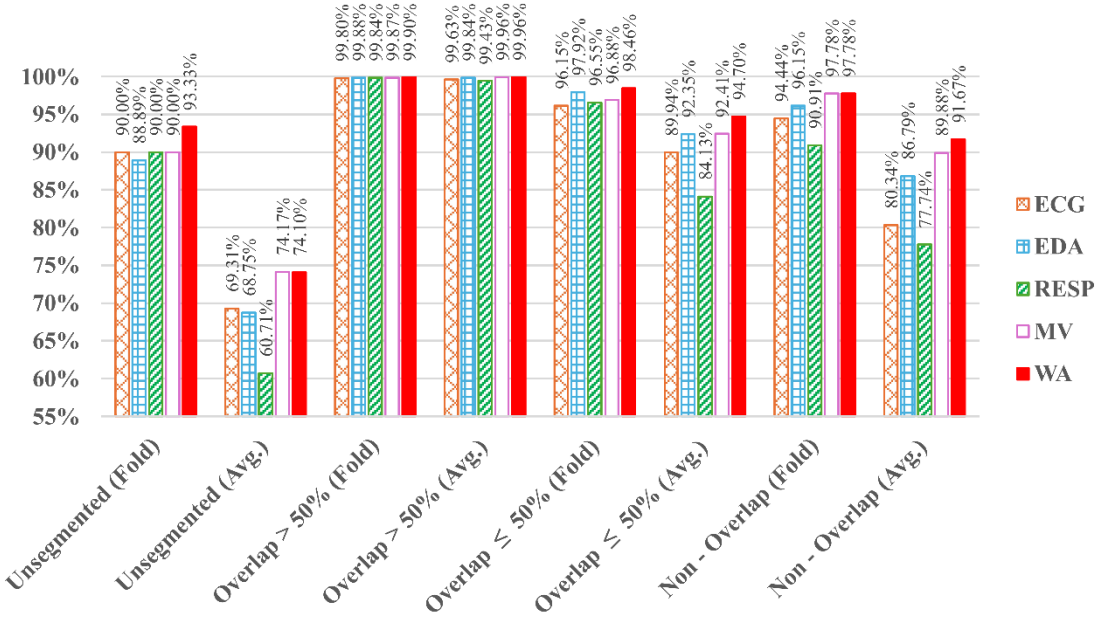

Fig. 2: Best binary classification results based on time domain features

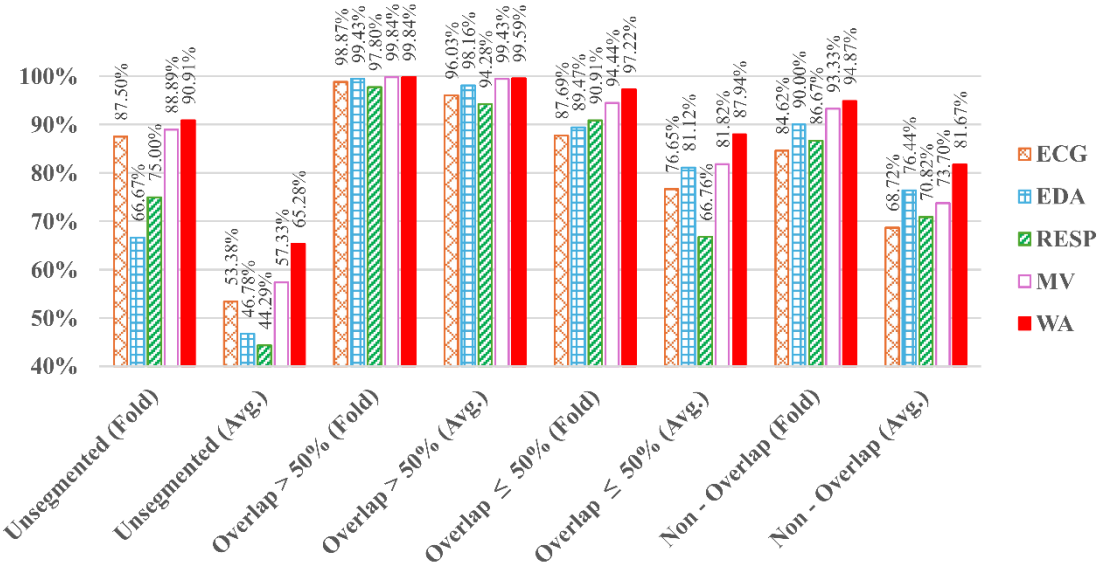

Fig. 3: Best multi-class classification results based on time domain features

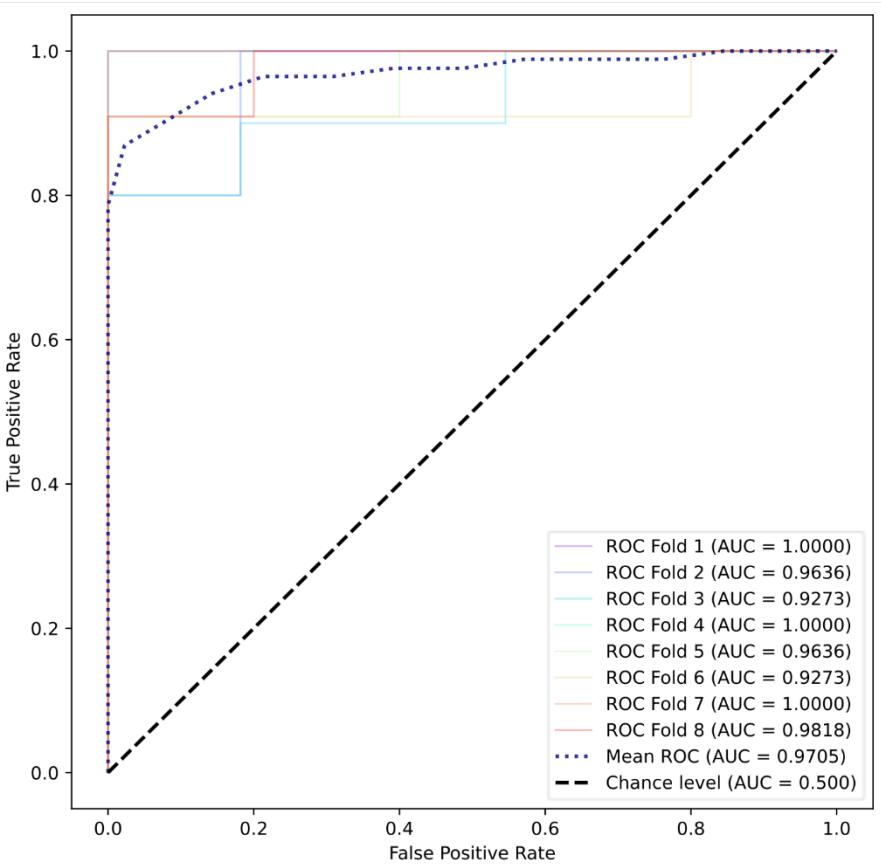

Fig. 4: Binary classification WA 210_210 ROC curve based on time domain features

Table 2: Best results based on time domain features

Modality	Type	Segmentation (Window_Shift)	Model	K-CV	Fold	ACC%	P%	R%	F1%
ECG	Binary	Unsegmented (Fold)	XGB	7	6	90.00	100	80.00	88.89
		Unsegmented (Avg.)	XGB	8	Avg.	69.31	75.83	59.38	63.96
		300_10 (Fold)	BAG	5	2	99.80	99.59	100	99.80
		300_10 (Avg.)	BAG	8	Avg.	99.63	99.59	99.67	99.63
		390_210 (Fold)	ET	5	1	96.15	92.86	100	96.30
		390_210 (Avg.)	ET	8	Avg.	89.94	88.87	92.19	90.29
		120_120 (Fold)	XGB	10	2	94.44	90.00	100	94.74
		120_120 (Avg.)	XGB	7	Avg.	80.34	79.96	81.08	80.31
	Multi	Unsegmented (Fold)	BAG	9	9	87.50	91.67	87.50	86.67
		Unsegmented (Avg.)	BAG	7	Avg.	53.38	58.05	53.38	52.67
		390_10 (Fold)	XGB	10	1	98.87	98.93	98.87	98.87
		390_10 (Avg.)	XGB	9	Avg.	96.03	96.09	96.03	96.03
		120_60 (Fold)	XGB	10	10	87.69	87.89	87.69	87.74
		120_60 (Avg.)	XGB	9	Avg.	76.65	77.42	76.65	76.51
		300_300 (Fold)	kNN	10	5	84.62	88.46	84.62	83.66
		120_120 (Avg.)	RF	9	Avg.	68.72	71.05	68.72	68.41
EDA	Binary	Unsegmented (Fold)	RF	9	2	88.89	80.00	100	88.89
		Unsegmented (Avg.)	RF	10	Avg.	68.75	76.67	61.67	64.95
		300_10 (Fold)	ET	3	1	99.88	100	99.75	99.88
		300_10 (Avg.)	ET	10	Avg.	99.84	99.92	99.75	99.84
		210_120 (Fold)	DT	6	3	97.92	96.00	100	97.96
		390_210 (Avg.)	kNN	6	Avg.	92.35	88.72	96.97	92.61
		300_300 (Fold)	ET	5	1	96.15	100	92.31	96.00
		120_120 (Avg.)	kNN	9	Avg.	86.79	83.25	93.30	87.71
	Multi	Unsegmented (Fold)	XGB	9	1	66.67	61.11	66.67	59.26
		Unsegmented (Avg.)	XGB	4	Avg.	46.78	45.68	46.78	43.86

Modality	Type	Segmentation (Window_Shift)	Model	K-CV	Fold	ACC%	P%	R%	F1%
		300_10 (Fold)	ET	7	2	99.43	99.44	99.43	99.43
		300_10 (Avg.)	ET	9	Avg.	98.16	98.19	98.16	98.15
		300_210 (Fold)	RF	8	5	89.47	91.58	89.47	89.16
		60_30 (Avg.)	kNN	6	Avg.	81.12	81.38	81.12	80.93
		390_390 (Fold)	RF	9	8	90.00	93.33	90.00	89.33
		60_60 (Avg.)	ET	4	Avg.	76.44	78.17	76.44	76.33
RESP	Binary	Unsegmented (Fold)	LDA	8	2	90.00	100	80.00	88.89
		Unsegmented (Avg.)	LDA	10	Avg.	60.71	59.17	51.67	54.83
		300_10 (Fold)	XGB	4	2	99.84	99.67	100	99.84
		300_10 (Avg.)	XGB	9	Avg.	99.43	99.67	99.18	99.42
		210_120 (Fold)	kNN	10	3	96.55	100	92.86	96.30
		300_210 (Avg.)	BAG	6	Avg.	84.13	83.96	86.75	84.89
		390_390 (Fold)	LR	9	6	90.91	83.33	100	90.91
		300_300 (Avg.)	RF	6	Avg.	77.74	79.56	75.45	77.29
	Multi	Unsegmented (Fold)	kNN	9	9	75.00	66.67	75.00	68.75
		Unsegmented (Avg.)	kNN	10	Avg.	44.29	38.45	44.29	38.15
		390_20 (Fold)	ET	10	4	97.80	97.86	97.80	97.80
		390_10 (Avg.)	XGB	7	Avg.	94.28	94.35	94.28	94.27
		390_300 (Fold)	DT	9	9	90.91	93.94	90.91	90.91
		60_30 (Avg.)	ET	8	Avg.	66.76	67.74	66.76	66.76
		300_300 (Fold)	RF	9	4	86.67	88.33	86.67	86.48
		300_300 (Avg.)	RF	6	Avg.	70.82	75.26	70.82	70.48
All (MV)	Binary	Unsegmented (Fold)	XGB+RF+LDA	8	2	90.00	100	80.00	88.89
		Unsegmented (Avg.)	XGB+RF+LDA	8	Avg.	74.17	83.13	64.38	70.82
		120_10 (Fold)	XGB+ET+XGB	5	4	99.87	100	99.74	99.87
		300_10 (Avg.)	BAG+ET+XGB	4	Avg.	99.96	100	99.92	99.96
		390_210 (Fold)	ET+kNN+LR	4	4	96.88	94.12	100	96.97
		210_120 (Avg.)	RF+DT+kNN	5	Avg.	92.41	93.75	91.03	92.22
		120_120 (Fold)	XGB+kNN+DT	8	1	97.78	100	95.45	97.67
		210_210 (Avg.)	BAG+kNN+QDA	4	Avg.	89.88	91.91	88.10	89.65
	Multi	Unsegmented (Fold)	BAG+XGB+kNN	8	8	88.89	92.59	88.89	88.15
		Unsegmented (Avg.)	BAG+XGB+kNN	5	Avg.	57.33	61.93	57.33	55.79
		300_10 (Fold)	ET+ET+XGB	4	4	99.84	99.84	99.84	99.84
		300_10 (Avg.)	ET+ET+XGB	9	Avg.	99.43	99.43	99.43	99.43
		210_120 (Fold)	RF+kNN+kNN	8	4	94.44	95.68	94.44	94.33
		60_30 (Avg.)	XGB+kNN+ET	10	Avg.	81.82	84.04	81.82	82.08
		300_300 (Fold)	kNN+RF+RF	9	4	93.33	95.00	93.33	93.14
		60_60 (Avg.)	DT+ET+kNN	7	Avg.	73.70	78.39	73.70	74.37
All (WA)	Binary	Unsegmented (Fold)	XGB+RF+LDA	5	1	93.33	88.89	100	94.12
		Unsegmented (Avg.)	XGB+RF+LDA	5	Avg.	74.10	72.28	74.64	72.99
		210_10 (Fold)	BAG+kNN+BAG	3	1	99.90	99.81	100	99.90
		300_10 (Avg.)	BAG+XGB+kNN	4	Avg.	99.96	100	99.92	99.96
		120_60 (Fold)	XGB+RF+ET	10	7	98.46	96.97	100	98.46
		390_210 (Avg.)	ET+kNN+LR	6	Avg.	94.70	92.99	96.97	94.78
		120_120 (Fold)	XGB+kNN+DT	8	1	97.78	100	95.45	97.67
		210_210 (Avg.)	BAG+kNN+QDA	8	Avg.	91.67	89.65	95.23	91.99
	Multi	Unsegmented (Fold)	BAG+XGB+kNN	7	2	90.91	93.18	90.91	90.04
		Unsegmented (Avg.)	BAG+XGB+kNN	8	Avg.	65.28	64.81	65.28	61.89
		300_10 (Fold)	ET+ET+XGB	4	4	99.84	99.84	99.84	99.84
		300_10 (Avg.)	ET+ET+XGB	9	Avg.	99.59	99.60	99.59	99.59
		210_120 (Fold)	RF+kNN+kNN	8	4	97.22	97.57	97.22	97.21
		120_60 (Avg.)	XGB+kNN+kNN	9	Avg.	87.94	88.67	87.94	87.91
		120_120 (Fold)	RF+BAG+kNN	9	9	94.87	95.48	94.87	94.86
		120_120 (Avg.)	RF+BAG+kNN	8	Avg.	81.67	83.37	81.67	81.31

4.2. Frequency-domain

Ensemble methods mostly maintained their dominance in the frequency domain for both binary and multiclass classification, as depicted in Fig. 5 and Fig. 6 respectively. Fig. 7 depicts the ROC curve for non-overlap WA average binary classification. Table 3 summarizes the peak performance achieved for each modality, classification type, and ensemble method.

4.2.1 Performance Evaluation of ECG Modality

As presented in Table 3, the performance of the ECG modality was assessed across different segmentation strategies and classifiers for binary classification. With unsegmented ECG data, the optimized LDA classifier achieved the highest accuracy of 72.78% using 3-CV. For segmentations with over 50% overlap, a window of 390 seconds with a 10-second shift yielded the highest accuracy of 97.28% using an optimized kNN classifier and 9-CV. In cases of segmentation with 50% and under overlap, a window of 210 seconds with a 120-second shift resulted in the highest accuracy of 84.48% with an optimized RF classifier and 10-CV. For segmentation with no overlap, a window of 300 seconds provided the highest accuracy of 76.15% using an optimized ET classifier and 8-CV. These results are illustrated in Fig. 5, providing a visual summary of the performance across different segmentation strategies.

As detailed in Table 3, the evaluation of the ECG modality for multi-class classification revealed varying performance across different segmentation approaches and classifiers. With unsegmented ECG data, the optimized BAG classifier achieved the highest accuracy of 56.07% using 4-CV. For segmentations with over 50% overlap, a window of 300 seconds with a 20-second shift yielded the highest accuracy of 90.41% using an optimized ET classifier and 10-CV. In cases of segmentation with 50% and under overlap, a window of 60 seconds with a 30-second shift resulted in the highest accuracy of 73.66% with an optimized ET classifier and 5-CV. For segmentation with no overlap, a window of 210 seconds provided the highest accuracy of 73.50% using an optimized XGB classifier and 7-CV. Fig. 6 visually summarizes these results, highlighting the performance variations across different segmentation strategies.

4.2.2 Performance Evaluation of EDA Modality

As presented in Table 3, the performance of the EDA modality was assessed across different segmentation strategies and classifiers for binary classification. With unsegmented EDA data, the optimized kNN classifier achieved the highest accuracy of 78.29% using 5-CV. For segmentations with over 50% overlap, a window of 390 seconds with a 10-second shift yielded the highest accuracy of 95.19% using an optimized XGB classifier and 7-CV. In cases of segmentation with 50% and under overlap, a window of 210 seconds with a 120-second shift resulted in the highest accuracy of 86.89% with an optimized DT classifier and 9-CV. For segmentation with no overlap, a window of 60 seconds provided the highest accuracy of 83.88% using an optimized BAG classifier and 6-CV. These results are illustrated in Fig. 5, providing a visual summary of the performance across different segmentation strategies.

As detailed in Table 3, the evaluation of the EDA modality for multi-class classification revealed varying performance across different segmentation approaches and classifiers. With unsegmented EDA data, the optimized LR classifier achieved the highest accuracy of 53.47% using 8-CV. For segmentations with over 50% overlap, a window of 300 seconds with a 10-second shift yielded the highest accuracy of 80.37% using an optimized ET classifier and 9-CV. In cases of segmentation with 50% and under overlap, a window of 300 seconds with a 210-second shift resulted in the highest accuracy of 71.24% with an optimized XGB classifier and 9-CV. For segmentation with no overlap, a window of 120 seconds provided the highest accuracy of 70.99% using an optimized BAG classifier and 10-CV. Fig. 6 visually summarizes these results, highlighting the performance variations across different segmentation strategies.

4.2.3 Performance Evaluation of RESP Modality

As presented in Table 3, the performance of the RESP modality was assessed across different segmentation strategies and classifiers for binary classification. With unsegmented RESP data, the optimized LR classifier achieved the highest accuracy of 73.93% using 10-CV. For segmentations with over 50% overlap, a window of 390 seconds with a 10-second shift yielded the highest accuracy of 92.59% using an optimized ET classifier and 9-CV. In cases of segmentation with 50% and under overlap, a window of 300 seconds with a 210-second shift resulted in the highest accuracy of 78.95% with an optimized BAG classifier and 8-CV. For segmentation with no

overlap, a window of 300 seconds provided the highest accuracy of 79.95% using an optimized XGB classifier and 9-CV. These results are illustrated in Fig. 5, providing a visual summary of the performance across different segmentation strategies.

As detailed in Table 3, the evaluation of the RESP modality for multi-class classification revealed varying performance across different segmentation approaches and classifiers. With unsegmented RESP data, the optimized LDA classifier achieved the highest accuracy of 37.82% using 6-CV. For segmentations with over 50% overlap, a window of 390 seconds with a 20-second shift yielded the highest accuracy of 79.78% using an optimized ET classifier and 9-CV. In cases of segmentation with 50% and under overlap, a window of 120 seconds with a 60-second shift resulted in the highest accuracy of 64.73% with an optimized RF classifier and 8-CV. For segmentation with no overlap, a window of 390 seconds provided the highest accuracy of 69.44% using an optimized ET classifier and 6-CV. Fig. 6 visually summarizes these results, highlighting the performance variations across different segmentation strategies.

4.2.4 Performance Evaluation of Modalities MV Ensemble Method

As presented in Table 3, the performance of the modalities MV ensemble was assessed across different segmentation strategies and classifiers for binary classification. With unsegmented MV Ensemble data, the optimized LDA, kNN, and LR classifiers for ECG, EDA, and RESP respectively achieved the highest accuracy of 78.14% using 4-CV. For segmentations with over 50% overlap, a window of 390 seconds with a 10-second shift yielded the highest accuracy of 98.42% using optimized kNN, XGB, and ET classifiers for ECG, EDA, and RESP respectively, and 6-CV. In cases of segmentation with 50% and under overlap, a window of 210 seconds with a 120-second shift resulted in the highest accuracy of 90.70% with optimized RF, DT, and LR classifiers for ECG, EDA, and RESP respectively, and 9-CV. For segmentation with no overlap, a window of 300 seconds provided the highest accuracy of 83.78% using optimized ET, RF, and XGB classifiers for ECG, EDA, and RESP respectively, and 4-CV. These results are illustrated in Fig. 5, providing a visual summary of the performance across different segmentation strategies.

As detailed in Table 3, the evaluation of the modalities MV ensemble for multi-class classification revealed varying performance across different segmentation approaches and classifiers. With unsegmented MV Ensemble data, the optimized BAG, LR, and LDA classifiers for ECG, EDA, and RESP respectively achieved the highest accuracy of 52.14% using 6-CV. For segmentations with over 50% overlap, a window of 390 seconds with a 20-second shift yielded the highest accuracy of 90.94% using optimized ET, RF, and ET classifiers for ECG, EDA, and RESP respectively, and 8-CV. In cases of segmentation with 50% and under overlap, a window of 60 seconds with a 30-second shift resulted in the highest accuracy of 74.36% with optimized ET, BAG, and kNN classifiers for ECG, EDA, and RESP respectively, and 6-CV. For segmentation with no overlap, a window of 120 seconds provided the highest accuracy of 77.46% using optimized ET, BAG, and ET classifiers for ECG, EDA, and RESP respectively, and 5-CV. Fig. 6 visually summarizes these results, highlighting the performance variations across different segmentation strategies.

4.2.5 Performance Evaluation of Modalities WA Ensemble Method

As presented in Table 3, the performance of the modalities WA ensemble was assessed across different segmentation strategies and classifiers for binary classification. With unsegmented WA Ensemble data, the optimized LDA, kNN, and LR classifiers for ECG, EDA, and RESP respectively achieved the highest accuracy of 78.22% using 3-CV. For segmentations with over 50% overlap, a window of 390 seconds with a 10-second shift yielded the highest accuracy of 98.36% using optimized kNN, XGB, and ET classifiers for ECG, EDA, and RESP respectively, and 10-CV. In cases of segmentation with 50% and under overlap, a window of 210 seconds with a 120-second shift resulted in the highest accuracy of 90.35% with optimized RF, DT, and LR classifiers for ECG, EDA, and RESP respectively, and 6-CV. For segmentation with no overlap, a window of 120 seconds provided the highest accuracy of 86.53% using optimized BAG, RF, and RF classifiers for ECG, EDA, and RESP respectively and 10-CV, additionally ROC is depicted in Fig. 7. These results are illustrated in Fig. 5, providing a visual summary of the performance across different segmentation strategies.

As detailed in Table 3, the evaluation of the modalities WA ensemble for multi-class classification revealed varying performance across different segmentation approaches and classifiers. With unsegmented WA Ensemble data, the optimized BAG, LR, and LDA classifiers for ECG, EDA, and RESP respectively achieved the highest accuracy of 57.50% using 8-CV. For segmentations with over 50% overlap, a window of 300 seconds with a 20-second shift yielded the highest accuracy of 96.26% using optimized ET, XGB, and ET classifiers for ECG, EDA,

and RESP respectively, and 10-CV. In cases of segmentation with 50% and under overlap, a window of 60 seconds with a 30-second shift resulted in the highest accuracy of 83.34% with optimized ET, BAG, and kNN classifiers for ECG, EDA, and RESP respectively, and 5-CV. For segmentation with no overlap, a window of 120 seconds provided the highest accuracy of 83.38% using optimized ET, BAG, and ET classifiers for ECG, EDA, and RESP respectively, and 5-CV. Fig. 6 visually summarizes these results, highlighting the performance variations across different segmentation strategies.

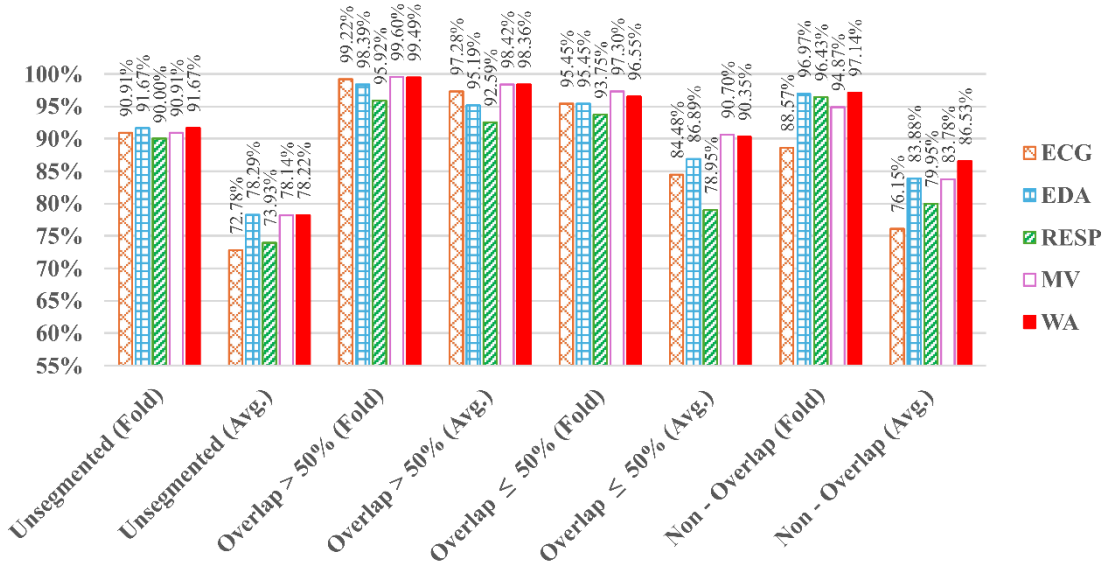

Fig. 5: Best binary classification results based on frequency domain features

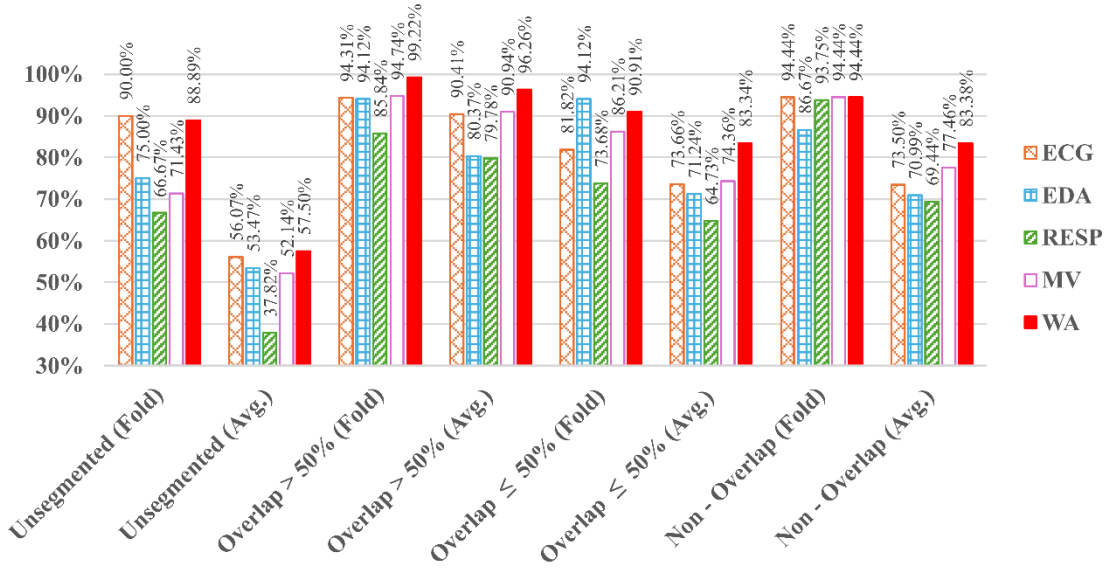

Fig. 6: Best multi-class classification results based on frequency domain features

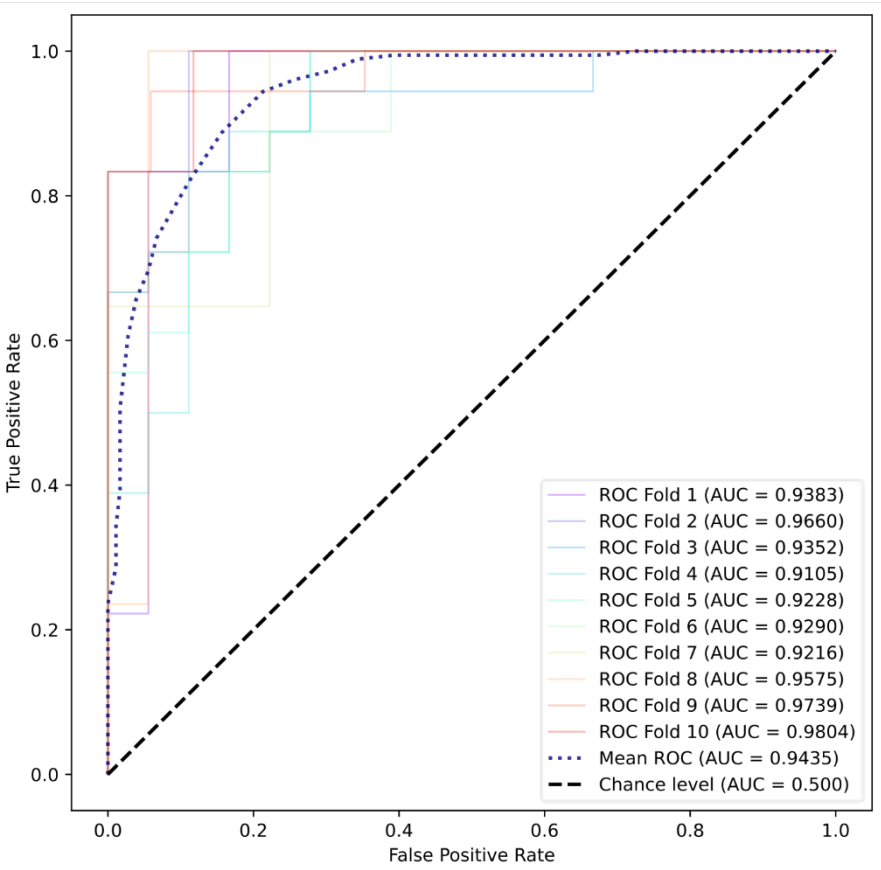

Fig. 7: Binary classification WA 120_120 ROC curve based on frequency domain features

Table 3: Best results based on frequency domain features

Modality	Type	Segmentation (Window_Shift)	Model	K-CV	Fold	ACC%	P%	R%	F1%
ECG	Binary	Unsegmented (Fold)	LDA	7	3	90.91	100	80.00	88.89
		Unsegmented (Avg.)	LDA	3	Avg.	72.78	80.30	58.97	67.78
		390_20 (Fold)	ET	7	4	99.22	98.46	100	99.22
		390_10 (Avg.)	kNN	9	Avg.	97.28	96.35	98.30	97.31
		390_210 (Fold)	XGB	6	1	95.45	91.67	100	95.65
		210_120 (Avg.)	RF	10	Avg.	84.48	82.44	88.95	85.26
		120_120 (Fold)	BAG	10	10	88.57	93.75	83.33	88.24
		300_300 (Avg.)	ET	8	Avg.	76.15	75.49	78.47	76.51
	Multi	Unsegmented (Fold)	BAG	8	2	90.00	93.33	90.00	89.33
		Unsegmented (Avg.)	BAG	4	Avg.	56.07	68.49	56.07	55.98
		300_20 (Fold)	ET	10	2	94.31	94.52	94.31	94.33
		300_20 (Avg.)	ET	10	Avg.	90.41	90.82	90.41	90.39
		390_300 (Fold)	ET	10	3	81.82	87.12	81.82	80.35
		60_30 (Avg.)	ET	5	Avg.	73.66	74.51	73.66	73.79
		210_210 (Fold)	XGB	9	9	94.44	95.56	94.44	94.36
		210_210 (Avg.)	XGB	7	Avg.	73.50	76.09	73.50	73.08
EDA	Binary	Unsegmented (Fold)	kNN	6	3	91.67	100	83.33	90.91
		Unsegmented (Avg.)	kNN	5	Avg.	78.29	84.29	69.29	74.74
		390_30 (Fold)	XGB	10	2	98.39	96.88	100	98.41
		390_10 (Avg.)	XGB	7	Avg.	95.19	93.12	97.62	95.31
		390_210 (Fold)	RF	6	1	95.45	91.67	100	95.65
		210_120 (Avg.)	DT	9	Avg.	86.89	87.45	86.81	86.79
		390_390 (Fold)	kNN	3	2	96.97	100	93.75	96.77
		60_60 (Avg.)	BAG	6	Avg.	83.88	84.52	83.33	83.83
	Multi	Unsegmented (Fold)	LR	9	9	75.00	87.50	75.00	75.00
		Unsegmented (Avg.)	LR	8	Avg.	53.47	50.70	53.47	48.47

Modality	Type	Segmentation (Window_Shift)	Model	K-CV	Fold	ACC%	P%	R%	F1%
		300_210 (Fold)	XGB	9	5	94.12	95.59	94.12	93.95
		300_10 (Avg.)	ET	9	Avg.	80.37	80.65	80.37	80.23
		300_210 (Fold)	XGB	9	5	94.12	95.59	94.12	93.95
		300_210 (Avg.)	XGB	9	Avg.	71.24	75.95	71.24	69.95
		390_390 (Fold)	DT	6	6	86.67	90.00	86.67	86.29
		120_120 (Avg.)	BAG	10	Avg.	70.99	74.28	70.99	70.36
RESP	Binary	Unsegmented (Fold)	LR	7	6	90.00	100	80.00	88.89
		Unsegmented (Avg.)	LR	10	Avg.	73.93	80.17	66.67	70.90
		390_10 (Fold)	ET	9	5	95.92	95.92	95.92	95.92
		390_10 (Avg.)	ET	9	Avg.	92.59	92.13	93.21	92.65
		300_210 (Fold)	BAG	10	2	93.75	88.89	100	94.12
		300_210 (Avg.)	BAG	8	Avg.	78.95	78.17	84.31	80.05
		210_210 (Fold)	XGB	6	5	96.43	93.33	100	96.55
		300_300 (Avg.)	XGB	9	Avg.	79.95	77.52	86.31	81.40
	Multi	Unsegmented (Fold)	LDA	6	6	66.67	88.89	66.67	70.56
		Unsegmented (Avg.)	LDA	6	Avg.	37.82	46.68	37.82	38.48
		390_20 (Fold)	ET	8	3	85.84	85.81	85.84	85.74
		390_20 (Avg.)	ET	9	Avg.	79.78	80.72	79.78	79.88
		390_210 (Fold)	kNN	7	4	73.68	74.04	73.68	73.43
		120_60 (Avg.)	RF	8	Avg.	64.73	66.65	64.73	64.34
		390_390 (Fold)	ET	6	5	93.75	95.31	93.75	93.75
		390_390 (Avg.)	ET	6	Avg.	69.44	75.02	69.44	69.06
All (MV)	Binary	Unsegmented (Fold)	LDA+kNN+LR	7	1	90.91	100	83.33	90.91
		Unsegmented (Avg.)	LDA+kNN+LR	4	Avg.	78.14	84.15	69.72	75.60
		390_10 (Fold)	kNN+XGB+ET	7	5	99.60	100	99.21	99.60
		390_10 (Avg.)	kNN+XGB+ET	6	Avg.	98.42	97.78	99.10	98.43
		210_120 (Fold)	RF+DT+LR	8	2	97.30	95.00	100	97.44
		210_120 (Avg.)	RF+DT+LR	9	Avg.	90.70	89.94	92.40	90.87
		120_120 (Fold)	BAG+RF+RF	9	9	94.87	100	90.00	94.74
		300_300 (Avg.)	ET+RF+XGB	4	Avg.	83.78	84.55	83.09	83.75
	Multi	Unsegmented (Fold)	BAG+LR+LDA	10	7	71.43	71.43	71.43	66.67
		Unsegmented (Avg.)	BAG+LR+LDA	6	Avg.	52.14	62.97	52.14	51.88
		390_120 (Fold)	RF+RF+DT	10	10	94.74	95.79	94.74	94.50
		390_20 (Avg.)	ET+RF+ET	8	Avg.	90.94	91.73	90.94	90.91
		210_120 (Fold)	kNN+XGB+kNN	10	5	86.21	86.50	86.21	86.13
		60_30 (Avg.)	ET+BAG+kNN	6	Avg.	74.36	78.07	74.36	74.73
		300_300 (Fold)	RF+RF+RF	7	5	94.44	95.56	94.44	94.36
		120_120 (Avg.)	ET+BAG+ET	5	Avg.	77.46	80.82	77.46	77.38
All (WA)	Binary	Unsegmented (Fold)	LDA+kNN+LR	6	3	91.67	100	83.33	90.91
		Unsegmented (Avg.)	LDA+kNN+LR	3	Avg.	78.22	80.98	72.65	75.98
		390_10 (Fold)	kNN+XGB+ET	9	3	99.49	99.00	100	99.50
		390_10 (Avg.)	kNN+XGB+ET	10	Avg.	98.36	97.37	99.44	98.38
		210_120 (Fold)	RF+DT+LR	10	2	96.55	93.33	100	96.55
		210_120 (Avg.)	RF+DT+LR	6	Avg.	90.35	88.19	93.83	90.70
		120_120 (Fold)	BAG+RF+RF	10	8	97.14	94.44	100	97.14
		120_120 (Avg.)	BAG+RF+RF	10	Avg.	86.53	83.44	91.60	87.16
	Multi	Unsegmented (Fold)	BAG+LR+LDA	8	8	88.89	94.44	88.89	88.89
		Unsegmented (Avg.)	BAG+LR+LDA	8	Avg.	57.50	58.75	57.50	54.98
		390_20 (Fold)	ET+RF+ET	7	4	99.22	99.25	99.22	99.22
		300_20 (Avg.)	ET+XGB+ET	10	Avg.	96.26	96.45	96.26	96.25
		390_300 (Fold)	ET+RF+LR	9	7	90.91	93.94	90.91	90.30
		60_30 (Avg.)	ET+BAG+kNN	5	Avg.	83.34	84.50	83.34	83.48
		300_300 (Fold)	RF+RF+RF	7	5	94.44	95.83	94.44	94.44
		120_120 (Avg.)	ET+BAG+ET	5	Avg.	83.38	85.34	83.38	83.44

5. Discussion

In this study, we investigated the impact of wearable devices on the early detection of psychological stress, employing both binary and five-class classifications. Our findings reveal significant correlations between the modalities, ECG, EDA, and RESP, and stress levels, suggesting the efficacy of these biomarkers for stress detection. Additionally, we employed two ensemble methods that simultaneously integrated these modalities. These results are consistent with previous research (e.g., [14], [16], [17], [19], [20], [21], [23], [25]), which also highlighted the utility of physiological signals in stress monitoring. Our investigation into five-class stress classification and the use of ensemble methods provides a novel contribution to the field. These results suggest that wearable devices could significantly enhance stress monitoring and management, improving overall quality of life. Future research should aim to address current limitations and refine these models further.

5.1. Comparative Evaluation of Methodologies

Table 4 juxtaposes our research findings with relevant studies, delineating comparisons based on the utilization of individual modalities, namely, ECG, EDA, and RESP. Each of our employed modalities is juxtaposed against corresponding studies employing singular modalities. Additionally, our ensemble methodology is compared against studies integrating all three modalities simultaneously.

Zhu et al (2023) [32] conducted research on binary stress classification utilizing exclusively the EDA modality from four distinct datasets: CLAS, UTD, VerBIO, and WESAD. The primary focus is on the WESAD dataset, which our study employed. In their investigation, an accuracy of 86.5% was achieved by utilizing segmentation without overlap, employing a 30-second window size, and employing the RF classifier with LOSO cross-validation. Conversely, our study attained a slightly higher accuracy of 86.79% under similar settings, exclusively utilizing the EDA modality, employing segmentation without overlap, employing a 120-second window size, and utilizing the kNN classifier with 9-CV.

Adarsh et al (2024) [33] investigated binary stress classification by leveraging the ECG modality from two distinct datasets: SWELL and WESAD. The primary focus is on the WESAD dataset, which our study employed. In the research conducted by the authors, an accuracy of 97.75% was achieved through segmentation with more than 50% overlap, utilizing a window size of 5 seconds with a 0.25-second shift, and employing graph convolutional Networks (GCN) with 5-CV. Conversely, the investigation conducted in our study yielded a higher accuracy of 99.63% under analogous conditions, exclusively employing the ECG modality, employing segmentation with more than 50% overlap, utilizing a window size of 300 seconds with a 10-second shift, and employing the BAG classifier with 8-CV.

Schmidt et al (2018) [16] explored binary stress classification by utilizing the Respiration (RESP) modality from the WESAD dataset, which was also employed in our study. In the investigation conducted by Schmidt et al, an accuracy of 88.09% was achieved through segmentation with more than 50% overlap, employing a window size of 60 seconds with a 0.25-second shift, and utilizing LDA with LOSO cross-validation. Conversely, our study attained a higher accuracy of 99.43% under similar conditions, exclusively utilizing the RESP modality, employing segmentation with more than 50% overlap, utilizing a window size of 300 seconds with a 10-second shift, and employing the XGB classifier with 9-CV.

Rashid et al (2023) [34] investigated binary stress classification through a multimodal approach incorporating ECG, EDA, and RESP modalities from the WESAD dataset, which aligns with the dataset utilized in our study. In their investigation, Rashid et al. achieved an accuracy of 81.62% by employing segmentation with more than 50% overlap, utilizing a window size of 60 seconds with a 5-second shift, and employing AB with LOSO cross-validation. Conversely, our study achieved a notably higher accuracy of 99.96% under analogous conditions, employing a multimodal WA ensemble of ECG, EDA, and RESP modalities. Our methodology involved segmentation with more than 50% overlap, utilizing a window size of 300 seconds with a 10-second shift, and employing BAG, XGB, and kNN classifiers for ECG, EDA, and RESP, respectively, with 4-CV.

Table 4: Comparison with related studies

Paper References	Dataset	Modality	Validation	Model	ACC%	Our Proposal
Zhu et al (2023) [32]	CLAS	EDA Non-overlap	LOSO	SVM	68.5	EDA 120_120 9-CV kNN **86.79**
	UTD			RF	73.1	
	VerBIO			SVM	92.9	
	WESAD			RF	**86.5**	
Adarsh et al (2024) [33]	**WESAD**	ECG Overlap>50%	5-CV	GCN	**97.75**	ECG 300_10 8-CV BAG **99.63**
	SWELL				94.48	
Schmidt et al (2018) [16]	WESAD	RESP Overlap>50%	LOSO	LDA	88.09	RESP 300_10 9-CV XGB **99.43**
Rashid et al (2023) [34]	WESAD	ECG+EDA+RESP Overlap>50%	LOSO	AB	**81.62**	WA 300_10 4-CV BAG+XGB+ kNN **99.96**
		ACCE+ECG+EDA Overlap>50%			86.37	

5.2. Principle Findings

We utilized three different modalities: ECG, EDA, and RESP. Additionally, we employed two ensemble methods that simultaneously integrated these modalities. Our investigation into commercially available wearable devices that offer these modalities is summarized in Table 5, derived from Taskasaplidis et al (2024) [35].

Table 5: Commercial wearable devices that provide ECG, EDA, and RESP modalities.

Device Name	Body Location	ECG	EDA	RESP	Additional Features
Fitbit Sense 2	Wrist	Yes	Yes	Yes	SpO2, SKT
Flowtime	Head	Yes	No	No	2-channel brainwave
Movesense	Chest	Yes	No	No	Motion measurement
Prana	Weist	No	No	Yes	Posture
Sentio Solutions Feel Therapeutics	Wrist	Yes	Yes	No	Physical activity, SKT

Based on our findings, we recommend the configurations presented in Table 6 for optimal performance in real-world applications tailored to the system type and available modalities.

Table 6: Recommended Configurations for Optimal Performance.

System	Modality	Type	Features Domain	Segmentation (Window_Shift)	Model
Offline	ECG	Binary	Time	120_120	XGB
		Multi	Frequency	210_210	XGB
	EDA	Binary	Time	120_120	kNN
		Multi	Time	60_60	ET
	RESP	Binary	Frequency	300_300	XGB
		Multi	Time	300_300	RF
	All (WA)	Binary	Time	210_210	BAG+kNN+QDA
		Multi	Frequency	120_120	ET+BAG+ET
Online	ECG	Binary	Time	300_10	BAG
		Multi	Time	390_10	XGB
	EDA	Binary	Time	300_10	ET
		Multi	Time	300_10	ET
	RESP	Binary	Time	300_10	XGB
		Multi	Time	390_10	XGB
	All (WA)	Binary	Time	300_10	BAG+XGB+kNN
		Multi	Time	300_10	ET+ET+XGB

6. Conclusion

This study explored the effectiveness of wearable devices in the early detection of psychological stress, utilizing both binary and five-class classification models. Our findings demonstrated significant correlations between stress levels and physiological signals from ECG, EDA, and RESP, confirming these modalities as reliable biomarkers for stress detection. We tested ten different classifiers: Random Forest (RF), Extreme Gradient Boosting (XGB), k-nearest Neighbors (kNN), Logistic Regression (LR), Decision Tree (DT), AdaBoost (AB), Extra Trees (ET), Bagging (BAG), Quadratic Discriminant Analysis (QDA), and Linear Discriminant Analysis (LDA). We also applied hyperparameter optimization using grid search, incorporating time and frequency domain features separately in our analyses. We employed two ensemble methods, Majority Voting (MV) and Weighted Averaging (WA), to integrate these modalities, enhancing the accuracy and robustness of the stress detection system. Additionally, we reviewed commercially available wearable devices capable of providing these physiological measurements. Based on our findings, we recommend the configurations detailed in Table 6 for optimal performance in real-world applications. These recommendations are tailored to the specific system types and available modalities, ensuring maximum effectiveness and utility. This research underscores the potential of multimodal wearable devices in the early detection and monitoring of psychological stress, offering a foundation for future research and practical applications in wearable health technology. While our study provides valuable insights, future research could benefit from exploring the integration of both time and frequency domain features, as well as investigating the potential of deep learning models to enhance detection capabilities further.

Data availability and materials

The data used in the publication is publicly available.

Code availability

The developed code can be provided by the Corresponding Author upon reasonable request.

Ethical approval

No ethical approval is required for this study.

Competing interests

The authors declare no competing interests.

Author contributions

Basil A. Darwish: Writing – original draft, Methodology, Formal analysis, Data curation. **Nancy M. Salem:** Writing – review & editing, Validation, Supervision, Methodology, Formal analysis, Conceptualization. **Ghada Kareem:** Writing – review & editing, Validation, Methodology, Conceptualization. **Lamees N. Mahmoud:** Writing – review & editing, Validation, Methodology, Conceptualization. **Ibrahim Sadek:** Writing – review & editing, Validation, Supervision, Methodology, Formal analysis, Conceptualization.

Appendix: List of Abbreviations

AB: AdaBoost

ACC: Accuracy

ACCE: Three-Axis Acceleration

AI: Artificial Intelligence

Avg: Average

BAG: Bagging

BVP: Blood Volume Pulse

cGAN: Conditional Generative Adversarial Network

CNN: Convolutional Neural Network

DL: Deep Learning

DNN: Deep Neural Network

DT: Decision Tree

ECG: Electrocardiogram

EDA: Electrodermal Activity

ET: Extra Trees

F1: F1 Score

FCN: Fully Convolutional Network

FFT: Fast Fourier Transform

GAN: Generative Adversarial Network

GCN: Graph Convolutional Networks

GSR: Galvanic Skin Response

HR: Heart Rate

HRV: Heart Rate Variability

IMU: Inertial Measurement Unit

kNN: k-Nearest Neighbors

K-CV: K-fold Cross-Validation

LDA: Linear Discriminant Analysis

LOSO: Leave-One-Subject-Out

LR: Logistic Regression

LSTM: Long Short-Term Memory

ML: Machine Learning

MLP: Multi-Layer Perceptron

MMTM: Multimodal Transfer Module

MV: Majority Voting

P: Precision

PANAS: Positive and Negative Affect Schedule Questionnaire

PPG: Photoplethysmography

PRV: Pulse Rate Variability

PSD: Power Spectral Density

QDA: Quadratic Discriminant Analysis

R: Recall

RESP: Respiration

RF: Random Forest

ROC: Receiver Operating Characteristic

SFM: Select From Model

SKT: Skin Temperature

SVM: Support Vector Machine

TEMP: Body Temperature

TSST: Trier Social Stress Test

WA: Weighted Average

WESAD: Wearable Stress and Affect Detection Dataset

XGB: Extreme Gradient Boosting

References

[1] A. F. A. Mentis, D. Lee, and P. Roussos, "Applications of artificial intelligence−machine learning for detection of stress: a critical overview," *Molecular Psychiatry 2023*, pp. 1–13, Apr. 2023, doi: 10.1038/s41380-023-02047-6.

[2] S. Sharma, G. Singh, and M. Sharma, "A comprehensive review and analysis of supervised-learning and soft computing techniques for stress diagnosis in humans," *Comput Biol Med*, vol. 134, p. 104450, Jul. 2021, doi: 10.1016/J.COMPBIOMED.2021.104450.

[3] R. Li and Z. Liu, "Stress detection using deep neural networks," *BMC Med Inform Decis Mak*, vol. 20, no. 11, pp. 1–10, Dec. 2020, doi: 10.1186/S12911-020-01299-4/TABLES/5.

[4] A. Arsalan and M. Majid, "Human stress classification during public speaking using physiological signals," *Comput Biol Med*, vol. 133, p. 104377, Jun. 2021, doi: 10.1016/J.COMPBIOMED.2021.104377.

[5] S. Cohen and D. Janicki-Deverts, "Who's Stressed? Distributions of Psychological Stress in the United States in Probability Samples from 1983, 2006, and 20091," *J Appl Soc Psychol*, vol. 42, no. 6, pp. 1320–1334, Jun. 2012, doi: 10.1111/J.1559-1816.2012.00900.X.

[6] M. Kivimäki and A. Steptoe, "Effects of stress on the development and progression of cardiovascular disease," *Nat Rev Cardiol*, vol. 15, no. 4, pp. 215–229, Apr. 2018, doi: 10.1038/NRCARDIO.2017.189.

[7] J. Wang *et al.*, "The application of machine learning techniques in posttraumatic stress disorder: a systematic review and meta-analysis," *npj Digital Medicine 2024 7:1*, vol. 7, no. 1, pp. 1–13, May 2024, doi: 10.1038/s41746-024-01117-5.

[8] N. K. Iyortsuun, S. H. Kim, M. Jhon, H. J. Yang, and S. Pant, "A Review of Machine Learning and Deep Learning Approaches on Mental Health Diagnosis," *Healthcare*, vol. 11, no. 3, Feb. 2023, doi: 10.3390/HEALTHCARE11030285.

[9] W. N. Price and I. G. Cohen, "Privacy in the age of medical big data," *Nature Medicine 2019 25:1*, vol. 25, no. 1, pp. 37–43, Jan. 2019, doi: 10.1038/s41591-018-0272-7.

[10] E. Smets *et al.*, "Large-scale wearable data reveal digital phenotypes for daily-life stress detection," *npj Digital Medicine 2018 1:1*, vol. 1, no. 1, pp. 1–10, Dec. 2018, doi: 10.1038/s41746-018-0074-9.

[11] A. B. R. Shatte, D. M. Hutchinson, and S. J. Teague, "Machine learning in mental health: a scoping review of methods and applications," *Psychol Med*, vol. 49, no. 9, pp. 1426–1448, Jul. 2019, doi: 10.1017/S0033291719000151.

[12] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine 2019 25:1*, vol. 25, no. 1, pp. 44–56, Jan. 2019, doi: 10.1038/s41591-018-0300-7.

[13] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, and R. Ranganath, "A Review of Challenges and Opportunities in Machine Learning for Health.," *AMIA Jt Summits Transl Sci Proc*, vol. 2020, pp. 191–200, 2020, Accessed: May 21, 2024. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/32477638

[14] E. Smets *et al.*, "Comparison of machine learning techniques for psychophysiological stress detection," *Communications in Computer and Information Science*, vol. 604, pp. 13–22, 2016, doi: 10.1007/978-3-319-32270-4_2/FIGURES/3.

[15] M. Gjoreski, M. Luštrek, M. Gams, and H. Gjoreski, "Monitoring stress with a wrist device using context," *J Biomed Inform*, vol. 73, pp. 159–170, Sep. 2017, doi: 10.1016/J.JBI.2017.08.006.

[16] P. Schmidt, A. Reiss, R. Duerichen, and K. Van Laerhoven, "Introducing WeSAD, a multimodal dataset for wearable stress and affect detection," in *ICMI 2018 - Proceedings of the 2018 International Conference on Multimodal Interaction*, Association for Computing Machinery, Inc, Oct. 2018, pp. 400–408. doi: 10.1145/3242969.3242985.

[17] Y. S. Can, N. Chalabianloo, D. Ekiz, and C. Ersoy, "Continuous Stress Detection Using Wearable Sensors in Real Life: Algorithmic Programming Contest Case Study," *Sensors 2019, Vol. 19, Page 1849*, vol. 19, no. 8, p. 1849, Apr. 2019, doi: 10.3390/S19081849.

[18] P. Siirtola and J. Röning, "Comparison of Regression and Classification Models for User-Independent and Personal Stress Detection," *Sensors 2020, Vol. 20, Page 4402*, vol. 20, no. 16, p. 4402, Aug. 2020, doi: 10.3390/S20164402.

[19] E. E. Kaczor, B. Chapman, S. Carreiro, P. Indic, and J. Stapp, "Objective Measurement of Physician Stress in the Emergency Department Using a Wearable Sensor," *Proc Annu Hawaii Int Conf Syst Sci*, vol. 2020, p. 3729, 2020, doi: 10.24251/hicss.2020.456.

[20] T. Iqbal *et al.*, "A Sensitivity Analysis of Biophysiological Responses of Stress for Wearable Sensors in Connected Health," *IEEE Access*, vol. 9, pp. 93567–93579, 2021, doi: 10.1109/ACCESS.2021.3082423.

[21] T. Iqbal, A. Elahi, W. Wijns, and A. Shahzad, "Exploring Unsupervised Machine Learning Classification Methods for Physiological Stress Detection," *Front Med Technol*, vol. 4, p. 782756, Mar. 2022, doi: 10.3389/FMEDT.2022.782756/BIBTEX.

[22] M. Ehrhart, B. Resch, C. Havas, and D. Niederseer, "A Conditional GAN for Generating Time Series Data for Stress Detection in Wearable Physiological Sensor Data," *Sensors 2022, Vol. 22, Page 5969*, vol. 22, no. 16, p. 5969, Aug. 2022, doi: 10.3390/S22165969.

[23] R. Kuttala, R. Subramanian, and V. R. M. Oruganti, "Multimodal Hierarchical CNN Feature Fusion for Stress Detection," *IEEE Access*, vol. 11, pp. 6867–6878, 2023, doi: 10.1109/ACCESS.2023.3237545.

[24] P. Kalra and V. Sharma, "Mental Stress Assessment Using PPG Signal a Deep Neural Network Approach," *IETE J Res*, vol. 69, no. 2, pp. 879–885, Feb. 2023, doi: 10.1080/03772063.2020.1844068.

[25] A. Greco *et al.*, "Acute Stress State Classification Based on Electrodermal Activity Modeling," *IEEE Trans Affect Comput*, vol. 14, no. 1, pp. 788–799, Jan. 2023, doi: 10.1109/TAFFC.2021.3055294.

[26] R. Richer *et al.*, "Machine learning-based detection of acute psychosocial stress from body posture and movements," *Scientific Reports 2024 14:1*, vol. 14, no. 1, pp. 1–19, Apr. 2024, doi: 10.1038/s41598-024-59043-1.

[27] M. Albaladejo-González, J. A. Ruipérez-Valiente, and F. Gómez Mármol, "Evaluating different configurations of machine learning models and their transfer learning capabilities for stress detection using heart rate," *J Ambient Intell Humaniz Comput*, vol. 14, no. 8, pp. 11011–11021, Aug. 2023, doi: 10.1007/S12652-022-04365-Z/TABLES/6.

[28] M. Huljanah, Z. Rustam, S. Utama, and T. Siswantining, "Feature Selection using Random Forest Classifier for Predicting Prostate Cancer," *IOP Conf Ser Mater Sci Eng*, vol. 546, no. 5, p. 052031, Jun. 2019, doi: 10.1088/1757-899X/546/5/052031.

[29] M. Frid-Adar, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Synthetic data augmentation using GAN for improved liver lesion classification," *Proceedings - International Symposium on Biomedical Imaging*, vol. 2018-April, pp. 289–293, May 2018, doi: 10.1109/ISBI.2018.8363576.

[30] I. J. Goodfellow *et al.*, "Generative Adversarial Nets," *Adv Neural Inf Process Syst*, vol. 27, 2014, Accessed: Apr. 29, 2024. [Online]. Available: http://www.github.com/goodfeli/adversarial

[31] M. A. Aboamer, A. T. Azar, A. S. A. Mohamed, K. J. Bär, S. Berger, and K. Wahba, "Nonlinear features of heart rate variability in paranoid schizophrenic," *Neural Comput Appl*, vol. 25, no. 7–8, pp. 1535–1555, Dec. 2014, doi: 10.1007/S00521-014-1621-1/TABLES/8.

[32] L. Zhu *et al.*, "Stress Detection Through Wrist-Based Electrodermal Activity Monitoring and Machine Learning," *IEEE J Biomed Health Inform*, vol. 27, no. 5, pp. 2155–2165, May 2023, doi: 10.1109/JBHI.2023.3239305.

[33] V. Adarsh and G. R. Gangadharan, "Mental stress detection from ultra-short heart rate variability using explainable graph convolutional network with network pruning and quantisation," *Mach Learn*, pp. 1–28, Jan. 2024, doi: 10.1007/S10994-023-06504-9/TABLES/6.

[34] N. Rashid, T. Mortlock, and M. A. Al Faruque, "Stress Detection Using Context-Aware Sensor Fusion From Wearable Devices," *IEEE Internet Things J*, vol. 10, no. 16, pp. 14114–14127, Aug. 2023, doi: 10.1109/JIOT.2023.3265768.

[35] G. Taskasaplidis, D. A. Fotiadis, and P. D. Bamidis, "Review of Stress Detection Methods Using Wearable Sensors," *IEEE Access*, vol. 12, pp. 38219–38246, 2024, doi: 10.1109/ACCESS.2024.3373010.