BIOSIGNAL ANALYSIS METHODS FOR THE ASSESSMENT OF STRESS

by

FROUDAS G. MICHAIL
BSc. Applied Mathematics, University of Crete 2021

MASTER OF SCIENCE

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING,
MECHANICAL ENGINEERING
SCHOOL OF ENGINEERING
HELLENIC MEDITERRANEAN UNIVERSITY



MSC IN INFORMATICS ENGINEERING

Heraklion
2025

Approved by:
Major Professor
Professor Manolis Tsiknakis

COPYRIGHT

copyright© 2025

Froudas G. Michail. All rights reserved. This thesis/document is the intellectual property of the author and may not be reproduced, distributed, or used in any form without explicit permission. Any reference to this work should be properly cited.

Abstract

In recent years, stress has emerged as one of the most significant factors affecting both mental and physical health. Prolonged exposure to stressful situations can lead to serious conditions such as anxiety, depression and cardiovascular disease, negatively affecting quality of life. Early detection and accurate classification of stress is crucial for the development of effective management and intervention strategies.

In this study, we investigate the use of machine learning techniques to classify stress through physiological signals using the Wearable Stress and Affect Detection (WESAD) public dataset. WESAD includes recordings from wearable sensor devices, such as the Empatica E4 wristband and a chest belt (Respiban), and contains information from several physiological signals, such as heart rate (BVP - Blood Volume Pulse), electrodermal activity (EDA), respiratory patterns (RESP), as well as acceleration (ACC) and body temperature. Participants in the experiment were exposed to different emotional states, including baseline, stress, amusement and meditation, allowing the study of the physiological response of the body to each of them.

In our approach we exploit BVP and EDA signals from the Empatica E4 and develop state-of-the-art techniques for the pre-processing of signals, the extraction of features in both the time and frequency domains, and the development of machine learning algorithms for the detection and classification of emotional states, specifically focusing on stress. Specifically, we applied Random Forest, Support Vector Machines (SVM), Logistic regression, k-nearest neighbors (KNN), Gradient boosting and Naive Bayes algorithms, and performed a comparative analysis of their performance based on accuracy. To ensure that features have similar scales standardization and normalization techniques were deployed to optimize the accuracy of the classifiers.

The results of the study show that machine learning algorithms can classify stress with results achieving an accuracy of 94% using Random Forest, making stress detection systems based on wearables a reliable method for monitoring psychosomatic health. This study contributes to the development of intelligent stress recognition systems, which could be integrated into applications with the use of wearable devices, enabling early detection and tailoring personalized management strategies. Such technological developments have the potential to improve mental health care, increase productivity in the workplace and support individuals in managing their stress in a more effective way.

Περίληψη

Στη σύγχρονη εποχή, το στρες είναι ένας από τους σημαντικότερους παράγοντες που επηρεάζουν τόσο την ψυχική όσο και τη σωματική υγεία. Η παρατεταμένη έκθεση σε στρεσογόνες καταστάσεις μπορεί να οδηγήσει σε σοβαρές καταστάσεις όπως άγχος, κατάθλιψη και καρδιαγγειακές παθήσεις, επηρεάζοντας αρνητικά την ποιότητα ζωής. Η έγκαιρη ανίχνευση και η ακριβής ταξινόμηση του στρες είναι ζωτικής σημασίας για την ανάπτυξη αποτελεσματικών στρατηγικών διαχείρισης και παρέμβασης.

Σε αυτή τη μελέτη, διερευνούμε τη χρήση τεχνικών μηχανικής μάθησης για την ταξινόμηση του στρες μέσω φυσιολογικών σημάτων χρησιμοποιώντας το δημόσιο σύνολο δεδομένων Wearable Stress and Affect Detection (WESAD). Το WESAD περιλαμβάνει εγγραφές από φορητές συσκευές αισθητήρων, όπως το περικάρπιο Empatica E4 και μια ζώνη στήθους, και περιέχει πληροφορίες από διάφορα φυσιολογικά σήματα, όπως καρδιακό ρυθμό (BVP - Blood Volume Pulse), ηλεκτροδερμική δραστηριότητα (EDA), αναπνευστικά μοτίβα (RESP), επιτάχυνση (ACC) και θερμοκρασία σώματος. Οι συμμετέχοντες στο πείραμα εκτέθηκαν σε διαφορετικές καταστάσεις, συμπεριλαμβανομένης της ουδέτερης κατάστασης, του στρες, της διασκέδασης και του διαλογισμού, επιτρέποντας τη μελέτη της φυσιολογικής απόκρισης του σώματος σε καθεμία από αυτές.

Η προσέγγισή μας περιλαμβάνει προ επεξεργασία σημάτων, εξαγωγή χαρακτηριστικών από το πεδίο του χρόνου και των συχνοτήτων και εκπαίδευση αλγορίθμων μηχανικής μάθησης για την ταξινόμηση των συναισθηματικών καταστάσεων. Συγκεκριμένα, εφαρμόζουμε Random Forest, Support Vector Machines (SVM), Logistic regression, KNN, Gradient boosting και Naive Bayes συγκρίνοντας την απόδοσή τους μετρώντας την ακρίβεια, την ανάκληση και τη βαθμολογία F1. Η τυποποίηση μεθόδων χαρακτηριστικών, όπως η Τυποποίηση και η Κανονικοποίηση, συμβάλλει στη βελτιστοποίηση της ακρίβειας των ταξινομητών.

Τα αποτελέσματα της μελέτης δείχνουν ότι οι αλγόριθμοι μηχανικής μάθησης μπορούν να ταξινομήσουν το στρες με υψηλή ακρίβεια φτάνοντας το 94%, καθιστώντας τα φορετά συστήματα ανίχνευσης στρες μια αξιόπιστη μέθοδο για την παρακολούθηση της ψυχοσωματικής υγείας. Αυτή η μελέτη συμβάλλει στην αναγνώριση ότι η ανάπτυξη έξυπνων συστημάτων αναγνώρισης στρες που θα μπορούσαν να ενσωματωθούν σε φορητές συσκευές, επιτρέποντας έτσι τον έγκαιρο εντοπισμό και την προσαρμογή εξατομικευμένων στρατηγικών διαχείρισης. Τέτοιες τεχνολογικές εξελίξεις έχουν τη δυνατότητα να βελτιώσουν την φροντίδα της ψυχικής υγείας, να αυξήσουν την παραγωγικότητα στο χώρο εργασίας και να υποστηρίξουν τα άτομα στη διαχείριση του στρες τους με πιο αποτελεσματικό τρόπο.

Table of Contents

1.INTRODUCTION 1

1.1 DEFINITION OF STRESS 2
1.2 OBJECTIVE OF THIS THESIS 4

2. LITERATURE REVIEW 6

2.1 STRESS DETECTION THROUGH PHYSIOLOGICAL SIGNALS 6
2.1.1 ELECTRODERMAL ACTIVITY (EDA) - ANALYSIS AND USE 6
2.1.2. BLOOD VOLUME PULSE (BVP) - HEART RATE AND HRV 7
2.1.3 OTHER BIOSIGNALS IN STRESS DETECTION 8
2.1.4 STUDIES USING PHYSIOLOGICAL SIGNALS FOR STRESS DETECTION 9
2.2. COMPARISON OF EXISTING DATASETS FOR STRESS DETECTION 15
2.2.1. WESAD – ANALYSIS AND APPLICATIONS 15
2.2.2. PUBLIC DATASETS – CHARACTERISTICS AND LIMITATIONS 15

3. METHODOLOGY 18

3.1 DATA PREPROCESSING 18
3.1.1 SIGNAL FILTERING 19
3.1.2. DATA SEGMENTATION 20
3.2. FEATURE EXTRACTION 24
3.2.1. TIME DOMAIN FEATURES 24
3.2.2. FREQUENCY DOMAIN FEATURES 28
3.2.3. DISCRETE DOMAIN FEATURES 31
3.3. FEATURE SELECTION 31
3.3.1. FILTER METHODS 31
3.3.2. WRAPPER METHODS 32
3.3.3. EMBEDDED METHODS 32
3.4. CLASSIFICATION 34
3.4.2. MODEL EVALUATION METHODS 43
3.4.3. EVALUATION METRICS 44

4. IMPLEMENTATION 47

4.1. DESCRIPTION OF THE WESAD DATASET AND DATA 47
4.1.1 DESCRIPTION OF THE TOOLS AND SOFTWARE USED 49
4.2. DATA PROCESSING PROCESS 50

4.2.1. Signal filtering and normalization 61
4.2.2. Feature extraction and selection from EDA and BVP 65
4.3. Implementation of machine learning algorithms 74

5.RESULTS 75

5.1 EDA classification results (with feature selection) 76
5.2 BVP classification results (with feature selection) 82
5.3 EDA and BVP classification results (with all features) 91

6. CONCLUSIONS AND FUTURE WORK 95

6.1 Future work 96

REFERENCES 98

List of tables

Table 1. Overview of sensor modalities, devices, recorded physiological parameters, and sampling frequencies used in the WESAD dataset. 4
Table 2 Comparison of stress detection studies 14
Table 3 Overview of publicly available stress dataset..... 17
Table 4 Comparison of event-based and sliding window segmentation methods, including their use cases, advantages, and restrictions..... 21
Table 5 Confusion matrix representation showing the relationship between actual and predicted classifications 45
Table 6 Label definitions for different physiological states in the dataset 49
Table 7 Sensor specifications for physiological signal acquisition, including device type and sampling frequency. 49
Table 8 Time Domain signal features 66
Table 9 Time domain signal Dynamics 66
Table 10 Frequency domain Features 67
Table 11 Time Domain Statistical Features extracted for BVP signal 69
Table 12 Time Domain Statistical Features extracted for BVP signal. 69
Table 13 Frequency - Domain Features extracted for BVP signal. 70
Table 14 Classification performance of different machine learning models in EDA signal using the top 10 features with a 60-second. (Label 1: Baseline, Label 2 : Stress, Label 3 : Amusement , Label 4: Meditation)..... 76
Table 15 Classification performance of different machine learning models in EDA signal using the top 5 features with a 60-second. (Label 1: Baseline, Label 2 : Stress, Label 3 : Amusement , Label 4: Meditation)..... 77
Table 16 Classification performance of different machine learning models in EDA signal using the top 10 features with a 60-second with 25% overlap. (Label 1: Baseline, Label 2 : Stress, Label 3 : Amusement , Label 4: Meditation) 77
Table 17 Classification performance of different machine learning models in EDA signal using the top 5 features with a 60-second with 25% overlap. (Label 1: Baseline, Label 2 : Stress, Label 3 : Amusement , Label 4: Meditation) 78
Table 18 Classification performance of different machine learning models in EDA signal using the top 10 features with a 60-second with 50% overlap. (Label 1: Baseline, Label 2 : Stress, Label 3 : Amusement , Label 4: Meditation) 79
Table 19 Classification performance of different machine learning models in EDA signal using the top 5 features with a 60-second with 50% overlap. (Label 1: Baseline, Label 2 : Stress, Label 3 : Amusement , Label 4: Meditation) 79

Table 20 Classification performance of different machine learning models in BVP signal using the top 5 features with a 60-second window. (Label 1: Baseline, Label 2 : Stress, Label 3 : Amusement, Label 4: Meditation) 83
Table 21 Classification performance of different machine learning models in BVP signal using the top 10 features with a 60-second. (Label 1: Baseline, Label 2 : Stress, Label 3 : Amusement , Label 4: Meditation) 83
Table 22 Classification performance of different machine learning models using the top 5 features with a 60-second with 25% overlap. (Label 1: Baseline, Label 2 : Stress, Label 3 : Amusement , Label 4: Meditation) 84
Table 23 Classification performance of different machine learning models using the top 10 features with a 60-second with 25% overlap. (Label 1: Baseline, Label 2 : Stress, Label 3 : Amusement , Label 4: Meditation) 85
Table 24 Classification performance of different machine learning models using the top 5 features with a 60-second with 50% overlap. (Label 1: Baseline, Label 2 : Stress, Label 3 : Amusement , Label 4: Meditation) 86
Table 25 Performance of different machine learning models using the top 10 features with a 60-second with 50% overlap. (Label 1: Baseline, Label 2 : Stress, Label 3 : Amusement , Label 4: Meditation) 87
Table 26 BVP classification results without feature selection methods 93
Table 27 EDA classification results without feature selection. 94

List of Figures

Figure 1 Poincare plot for healthy subject and patients with arrhythmia, heart failure, and syncope. (A) Healthy subject. (B) Patient with arrhythmia. (C) Patient with heart failure. (D) Patient with syncope. (figure taken from [19]) 8
Figure 2 Overview of stress recognition process 19
Figure 3 Overview of time and event-based segmentation. Figure taken from [47] 20
Figure 4 Overview of non- overlapping(a) and overlapping windows (b). Figure taken from [52] 22
Figure 5 Overview of overlapped window segmentation for EEG signal. Figure taken from [53] 23
Figure 6 Overview of non-overlapping windows. Figure taken from [54] 23
Figure 7 Overview of SVM algorithms. Figure taken from [93] 35
Figure 8 Overview of Decision Trees. Figure taken from [94] 36
Figure 9 Overview of Random Forest algorithm. Figure taken from [95] 36
Figure 10 Overview of Neural networks algorithms. Figure taken from[96] 37
Figure 11 Overview of Gradient Boosting algorithms. Figure taken from [97] 37
Figure 12 Overview of Naïve Bayes Classifier. Figure taken from [98] 39
Figure 13 Overview of KNN classifier. Figure taken from [99] 39
Figure 14 Overview of Logistic Regression Classifier. Figure taken from [100] 40
Figure 15. Overview of hierarchical clustering algorithm. Figure taken from [105] 41
Figure 16 Overview of PCA Component Analysis. Figure taken from [109] 42
Figure 17 Visualization of the embedding space learned on SearchSnippets using Sentence Transformer. Figure taken from [112] 43
Figure 18 Illustration of k-fold cross-validation. Figure taken from [116] 44
Figure 19 Time intervals table representing the times that the label is changing. 50
Figure 20 ECG signal over time (top) and corresponding labels (bottom) 51
Figure 21 Electrodermal activity (EDA) signals from wrist (top) and chest (bottom) with labeled intervals 52
Figure 22 Electrodermal activity (EDA) signals from the wrist with labeled time intervals. . 53
Figure 23 Time-series plots of physiological signals, including Blood Volume Pulse (BVP), Accelerometer (ACC), Electrodermal Activity (EDA), and Temperature (TEMP), illustrating variations over time. 54
Figure 24 Eda baseline condition original and zoomed in 56
Figure 25 Eda stress condition and zoomed in. 56
Figure 26 Eda amusement condition original and zoomed in 57
Figure 27 Eda meditation condition original and zoomed in 58
Figure 28 BVP baseline condition original and zoomed in 59

Figure 29 BVP stress condition original and zoomed in 59
Figure 30 BVP meditation condition original and zoomed in 60
Figure 31 BVP amusement condition original and zoomed in 60
Figure 32 Comparison of the original EDA signal and the filtered EDA signal using a low-pass Butterworth filter (bottom) 61
Figure 33 Decomposition of the Electrodermal Activity (EDA) signal into tonic and phasic components, highlighting slow baseline changes (tonic) and rapid responses (phasic). ... 62
Figure 34 Electrodermal Activity (EDA) signal with detected peaks 63
Figure 35 Processing of the Blood Volume Pulse (BVP) signal, showing the original signal), filtered signal, and smoothed..... 64
Figure 36 Filtered Blood Volume Pulse (BVP) signal with detected peaks 65
Figure 37 Feature importance comparison across different machine learning models for EDA signal..... 71
Figure 38 Overall feature importance across all models for EDA signal 72
Figure 39 Feature importance comparison across different machine learning models for BVP signal..... 73
Figure 40 Overall feature importance across all models for BVP signal comment. 73
Figure 41 Confusion matrix for binary classification using labels 1 & 2 stress non-stress for Eda signal. 80
Figure 42 Confusion matrix for binary classification using labels 1,3,4 as non-stress and 2 as stress for EDA signal. 81
Figure 43 Confusion matrix for multiclass classification for EDA signal..... 82
Figure 44 Confusion matrix for binary classification using labels 1 & 2 for BVP signal. 88
Figure 45 Confusion matrix for binary classification using labels 1,3,4 as non-stress and 2 as stress for BVP signal. 89
Figure 46 Confusion matrix for multiclass classification for BVP signal. 90

1.Introduction

In recent decades, the prevalence of digital overload, heightened occupational demands, and socio-economic instability has contributed to the emergence of stress as a major public health concern. Numerous studies show that an increasing percentage of the population regardless of age or professional status is exposed to high levels of stressful stimuli, resulting in a burden on both physical and mental health. The continuous activation of physiological stress mechanisms, especially when not accompanied by an adequate recovery period, is associated with serious consequences such as cardiovascular diseases, depression, anxiety, cognitive dysfunction and immune system weakness. The need for early recognition and management of stress, especially with modern means such as sensor technology and machine learning techniques, is now more urgent than ever, both at a personal and social level.

Stress constitutes a physiological and psychological response to external or internal stimuli that are perceived as demanding, unpredictable, or threatening. It is an adaptive mechanism that has evolved to help humans respond to challenges by activating the Fight or Flight response. During this response, the Sympathetic Nervous System (SNS) stimulates the production of hormones such as adrenaline and cortisol, causing physiological changes such as an increase in heart rate, blood pressure and skin conductivity. Several factors influence the emerge of stress, including biological (genetic predisposition, hormone levels), psychological (perception and management of stress) and environmental (work and social conditions). Although acute stress can be beneficial, improving performance and preparedness, chronic stress has been linked to serious health effects such as cardiovascular disease and psychological disorders.
Two of the most important physiological signals associated with stress are:

- Electrodermal Activity (EDA): The measurement of electrodermal activity captures changes in skin conductance, which are directly related to the activity of the Sympathetic Nervous System. EDA is one of the most reliable indicators of the autonomic nervous system and is widely used to assess emotional and cognitive load.
- Blood Volume Pulse (BVP): The BVP signal captures peripheral blood flow variations through photoplethysmography (PPG), a non-invasive optical technique and is used to extract critical cardiovascular characteristics, such as heart rate (HR) and heart rate variability (HRV). HRV is an important indicator of the balance between the Sympathetic and Parasympathetic Nervous Systems, which makes it a useful tool for studying physiological responses to stress.

This work focuses on investigating how physiological signals can be used to detect stress, with the goal of developing methods that could be applied in real-world settings, such as health and well-being. To achieve this goal, a methodological approach is followed that initially includes a literature review, which aims to clarify existing approaches and identify the most important research trends in the field of stress detection and management through physiological signals. At the same time, an evaluation of the available methods and techniques that have been proposed in the literature is carried out, with the aim of identifying the most efficient approaches for stress diagnosis and management.

An open-source experimental dataset, which includes recordings of physiological EDA and BVP signals, is used to test and evaluate the various analysis algorithms. These data are processed and analyzed to investigate the effectiveness of the methods applied in the classification of stress states.

The methodological process followed initially includes an exploratory data analysis (EDA), during which visualization and initial analysis of the signals are performed, aiming to identify basic patterns and variations This is followed by signal analysis, where the structure of the physiological signals is examined and possible patterns related to the body's response to stress are identified.

A critical stage of the process is Feature Extraction, which involves the calculation of key temporal and frequency features from the EDA and BVP signals, which can be used in further analysis and classification. This is followed by Feature Selection, where the most important features that contribute to the discrimination of different states and improve the performance of classification algorithms are determined.

The last stage involves Classification, where machine learning models are developed and evaluated to categorize physiological signals into four states: Baseline, Stress, Amusement and Meditation. Through the overall analysis, the work seeks to contribute to the understanding of how physiological signals can be used to detect stress, with the aim of creating practical applications in the fields of health and wellness.

1.1 Definition of stress

Stress is a complex biological and psychological response of the body to factors that it perceives as threatening or demanding. It is a natural adaptive mechanism that has evolved to increase the body's alertness, attention and readiness in the face of challenges. However, while stress can be useful in short periods, its prolonged presence has serious implications for health and mental wellbeing. There are two main types of stress:

- **Acute stress**: is the body's instantaneous response to a sudden stressor. For example, when someone is in danger or has to cope with a difficult situation, the body activates mechanisms to increase its performance. Acute stress is not necessarily harmful - it often helps to increase concentration and efficiency.

- **Chronic stress**: When a person is continuously exposed to stressful situations without adequate recovery time, stress becomes a chronic condition. Chronic stress is associated with serious health effects such as increased risk of cardiovascular disease, weakened immune system, mental disorders and impaired cognitive function.[1]

1.1.1 The role of the Autonomic Nervous System (ANS)

The Autonomic Nervous System (ANS) is part of the peripheral nervous system and is responsible for the control and regulation of involuntary bodily functions, such as heart rate, blood pressure, body temperature, respiration, and digestive activity. Its function is crucial for maintaining homeostasis of the body, as it adapts physiological activity to external and internal demands. The ANS is divided into two main functional branches: the Sympathetic Nervous System (SNS) and the Parasympathetic Nervous System (PNS), which act competitively and cooperatively to achieve dynamic balance of the body's systems.

The Sympathetic branch is mainly activated during stressful or dangerous situations, leading to the initiation of the so-called "fight or flight" response. Its activation causes a multitude of physiological responses, such as acceleration of the heart rate (tachycardia), vasoconstriction, increase in blood pressure, dilation of the bronchi, suppression of digestive function, as well as the release of catecholamines (such as adrenaline and noradrenaline) from the adrenal glands, which prepare the body for immediate action against a threat. These responses are the result of activation of the hypothalamic-pituitary-adrenal axis (HPA axis) and the sympathoadrenal system, which enhance the mobilization of energy reserves and increase alertness.

In contrast, the Parasympathetic Nervous System is activated in states of rest and recovery, promoting the conservation of energy and the functional restoration of the body. Its action is characterized by a decrease in heart rate (bradycardia), a decrease in blood pressure, an increase in digestive and metabolic activity, and generally supporting processes such as digestion and tissue regeneration. The balance between the activities of these two subsystems is fundamental for the regulation of the physiological response to stress, and their dysfunction has been associated with a number of psychosomatic disorders, such as anxiety, depression and cardiovascular diseases.

It is also important to note that changes in the balance of the ANS can be observed and recorded through physiological signals, such as heart rate variability (HRV), electrodermal

activity (EDA) and heartbeat characteristics (BVP), making the ANS particularly important in the study of stress and psychophysiology in general.[2]

1.2 Objective of this thesis

The present study aims to contribute meaningfully to the field of stress detection through the analysis of physiological signals, with a particular focus on electrodermal activity (EDA) and blood volume pulse (BVP). More specifically, the efficiency of various feature extraction and selection methods, as well as the performance of multiple machine learning algorithms in the classification of emotional states are investigated. The classification problem is approached both at the binary level (stress vs. non-stress) and at the multi-class level, allowing a comprehensive evaluation of the ability of the models to distinguish between different emotional states in realistic experimental contexts.

The WESAD (Wearable Stress and Affect Detection) dataset [3], [4], [5], [6], [7], [8], an openly available and widely used experimental dataset for the analysis of emotional states through wearable sensors, was chosen for the implementation of the study. WESAD includes recordings of physiological signals such as EDA, BVP, respiration, skin temperature, and accelerometer, as presented in Table 1. The data were collected in a controlled laboratory environment during four emotional states: calm (baseline), stress, amusement, and meditation. The selection of this dataset was based on the completeness of the signals, the high quality of the recordings, as well as its wide acceptance in the literature [4][5][6][7] which allows for comparison of the results with other studies.

Table 1. Overview of sensor modalities, devices, recorded physiological parameters, and sampling frequencies used in the WESAD dataset.

Sensor	Device	Parameters	Frequency (Hz)
EDA	*Empatica E4*	*Skin conductivity*	*4 Hz*
BVP	*Empatica E4*	*Blood Volume*	*64 Hz*
ECG	*RespiBAN*	*Cardiac activity*	*700 Hz*
Respiration	*RespiBAN*	*Respiration*	*700 Hz*
ACC	*Empatica E4*	*Movement (x, y, z)*	*32 Hz*
ACC	*RespiBAN*	*Movement (x, y, z)*	*64 Hz*

By comparing various models and parameter configurations—such as the number of features and temporal overlap rates—this study seeks to identify the most effective

techniques for reliable stress detection, with the goal of enabling their application in domains such as healthcare, occupational environments, and overall well-being.

2. Literature Review

2.1 Stress Detection Through Physiological Signals

One of the main factors affecting human health and well-being is stress, combining psychological and physiological reactions. Chronic stress is a major contributor to the development of serious diseases such as cardiovascular disease, hypertension and various mental disorders such as anxiety and depression [9].The need for stress detection is vital as early prediction and intervention is required.

With the advancement of technology and the development of biomedical technologies and wearable devices, the scientific community has turned to the use of physiological signals for real-time stress detection. The assessment of physiological responses is a key element in understanding and detecting stress, as the Autonomic Nervous System (ANS) exhibits measurable changes during different emotional and cognitive states. Various physiological signals, such as Electrodermal Activity (EDA), Blood Volume Pulse (BVP), Heart Rate Variability (HRV), and respiration, provide valuable information about the body's neurophysiological response to stress [10] .Among the various biomarkers, EDA and BVP have been extensively studied [11][12][13][14] and are considered reliable indicators of sympathetic nervous system activity. The analysis of these signals through advanced machine learning and deep learning techniques allows the development of accurate and reliable stress recognition systems. These signals reflect basic physiological mechanisms, such as sweat gland activation and cardiovascular dynamics, which are directly affected by the presence of stressors. In the present study, we focus on EDA and BVP signals for in-depth analysis of the stress state.

This literature review to record and analyze the available datasets, the methods of analyzing physiological signals for stress detection, as well as the machine learning and deep learning techniques that have been applied in the relevant literature, is conducted. At the same time, the main challenges and open issues that remain to be investigated in the field of stress detection are highlighted. In this context, the physiological signals used for stress detection, the available datasets, the methodologies that have been applied for data analysis, the machine learning and deep learning techniques that are considered most effective, as well as the limitations that still exist and the possible future directions that could contribute to enhancing the accuracy and reliability of stress recognition systems are examined.

2.1.1 Electrodermal Activity (EDA) - Analysis and use

Electrodermal activity is one of the most widespread and reliable biomarkers for assessing the activity of the Sympathetic Nervous System The increase in skin conductance is directly

related to the stimulation of the sweat glands, which is caused by stress. The Electrodermal Activity (EDA) consists of two primary components: the Tonic Component, known as Skin Conductance Level (SCL), and the Phasic Component, referred to as Skin Conductance Response (SCR) [15]. SCL represents the baseline level of skin conductance and changes gradually over time, making it particularly useful for detecting prolonged stress and general emotional arousal. In contrast, SCR corresponds to short-term fluctuations in skin conductance, typically associated with transient emotional or cognitive stimuli, reflecting immediate physiological responses to external or internal events.

2.1.2. Blood Volume Pulse (BVP) - Heart Rate and HRV

Pulse Blood Volume (BVP) is recorded via photoplethysmography (PPG), which is based on the measurement of changes in blood circulation. Photoplethysmography (PPG) is an optical technique for measuring blood flow and heart rate through changes in light absorption by tissue. It uses a light source (usually LED) and a light sensor (photodiode), which records changes in the intensity of reflected light due to fluctuations in blood volume in the blood vessels.

The main features that can be extracted from BVP are heart rate (HR) and heart rate variability (HRV), which is considered particularly important for and critical for the diagnosis of stress.[16]

Heart Rate Variability (HRV) refers to the fluctuations in the time interval between successive heartbeats and serves as a crucial indicator of the balance between the Sympathetic and Parasympathetic Nervous Systems (PNS). Higher HRV values are generally associated with a healthy autonomic balance, reflecting greater adaptability and resilience to physiological and psychological stressors. Conversely, lower HRV values or a lack of significant fluctuations may indicate heightened sympathetic dominance, often linked to stress, anxiety, or cardiovascular dysfunction, suggesting a reduced ability of the autonomic nervous system to effectively regulate physiological responses.

There are three main methods for analyzing HRV. Namely: i) Time domain analysis, which focuses on extracting features such as SDNN (Standard Deviation of NN Intervals) and RMSSD (Root Mean Square of Successive Differences) [17] ii) Frequency domain analysis, which focuses on and analyzes the high and low-frequency components of the heart rate, determining the dominance of the sympathetic or parasympathetic system [18]nonlinear analysis, which uses methods such as the Poincare plot and entropy to estimate fluctuations.

In the work of ishaque et al. [19] , the authors explore the effectiveness of classifying stress levels by analyzing physiological signals collected during virtual reality (VR)

simulations. The study focuses on the application of decision tree algorithms for stress classification, demonstrating that these models provide reliable and interpretable results. Among the visual analysis tools employed, the Poincaré plot is highlighted as an effective nonlinear method for assessing heart rate variability and detecting potential abnormalities in autonomic regulation. Their graphical representation allows the visualization of short-term and long-term variations in interbit intervals, providing valuable insight into physiological responses under different conditions.

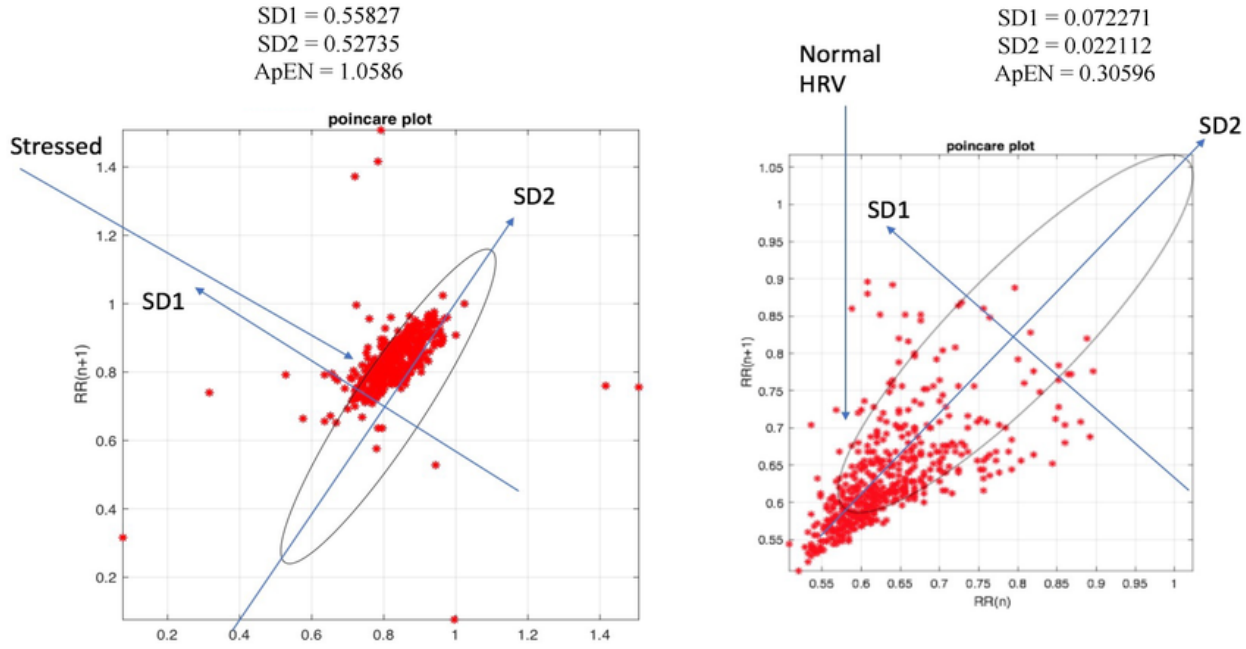


Figure 1 Poincare plot for healthy subject and patients with arrhythmia, heart failure, and syncope. (A) Healthy subject. (B) Patient with arrhythmia. (C) Patient with heart failure. (D) Patient with syncope. (figure taken from [19])

The Poincaré plot is a nonlinear time series analysis method widely used to study heart rate variability (HRV). Essentially, it is a scatter plot in which each time interval between two consecutive heartbeats (RR interval) is plotted as a function of the next. The resulting geometric shape provides information about short-term and long-term heart rate variability, with two key indices, SD1 and SD2, describing the range of dispersion vertically and diagonally, respectively. The Poincare plot is often used in research on the psychophysiological state of an individual, such as stress, fatigue, or general well-being, due to its ability to reveal patterns that are not apparent with linear techniques [20].

2.1.3 Other biosignals in stress detection

In addition to EDA and BVP, various physiological signals contribute to the detection of stress. The most important of these are the respiratory rate, which focuses on the detection

of respiratory frequency and the change in breathing pattern under stress conditions. Body temperature (Skin Temperature). Stress often causes vasoconstriction, resulting in a decrease in temperature, which is measured with thermosensors. And finally, the electroencephalogram (EEG), which refers to the recording of brain waves.[21] The analysis of these waves allows the analysis of the cognitive and emotional response to stressful situations. Using frequencies such as beta waves (13-30 hz) and theta waves (4-8 hz) to detect mental arousal.

2.1.4 Studies using physiological signals for stress detection

A significant body of research has been dedicated to the development of methodologies for stress detection, leveraging physiological signals, machine learning techniques, and wearable sensor technologies. Understanding and accurately identifying stress is crucial due to its profound impact on physical and mental health. In this section, we provide an overview of some of the most influential studies in the field, highlighting their methodologies, key findings, and contributions to the advancement of stress detection systems.

The creators of WESAD [3] used signals from Empatica E4 and Respiban. The E4[1] is a medical-grade wearable device that *offers real-time physiological data acquisition and streaming*. They recorded ECG, EDA, EMG, respiration, body temperature and acceleration signals from the RespiBAN device (700 Hz), as well as BVP, EDA, body temperature and acceleration signals from the Empatica E4 device (64 Hz, 4 Hz, 4 Hz and 32 Hz respectively). Regarding preprocessing, they used methods such as resampling, to harmonize the different frequencies of the sensors. In addition, they used noise filtering methods by applying Butterworth filters to remove high and low-frequency noise. as well as outlier removal through statistical data to eliminate extreme values in physiological signals. 60-second windows with 50% overlap were used to extract the features. For classification, they used SVM and random forest, which managed to achieve stress detection levels of up to 93% in binary classification. Additionally, they performed multiclass classification in which no specific accuracy is clearly stated. This percentage confirms the suitability of the dataset for use in relevant applications.[3]

A study that focuses exclusively on data derived from EDA, BVP and skin temperature from the empatica E4 device developed by Ninh et. al in [21]. The researchers extracted statistical features from the raw signals, such as the mean, variance, and skewness and skewness coefficients. The preprocessing of the data includes a Butterworth filter to eliminate high-frequency and low frequency noise, and outlier removal also performed to mitigate the influence of extreme values. The data were used as inputs to a simple neural network (MLP)

[1] https://www.empatica.com/en-eu/store/e4-wristband/

known for its capacity to model complex relationships in data. For feature extraction, 60-second windows with 50% overlap were used to maintain the continuity of the features. The accuracy of the model reached 86% in binary stress classification based on the features given to it, reinforcing the idea that wearable biosensors can be effectively utilized in mental health monitoring applications.

In another study, Rashid et al. [11] used only the BVP from the Empatica E4 device with a frequency of 64hz. They had a different approach to how effective deep learning methods can be. So, they developed a CNN model, which uses automatically extracted features through convolutional layers. Also, they manually selected features. Some of these are heart rate and temporal fluctuations of the signal. Noise reduction techniques were applied to the raw BVP signal to eliminate motion artifacts and high-frequency noise. Feature extraction was carried out using 30-second windows with 50% overlap to enhance the performance of the classifier. They performed binary classification for stress detection, the results of which showed an accuracy of 85% in stress detection, demonstrating that CNNs can be effectively utilized in similar applications.

In a related study [12], the electrodermal activity (EDA) signal—collected via the Empatica wearable device operating at a sampling frequency of 4 Hz—was highlighted as a key physiological indicator for stress detection, due to its high sensitivity to sympathetic nervous system activation. They implemented self-supervised learning with CNN, utilizing data transfer methods to address the problem of small data amounts. This method involved pretraining the model on unlabeled data to intrinsic patterns. The preprocessing of the signals included the application of 60-second windows, without reference to overlap. The final model achieved an accuracy of 80% in binary stress classification, which indicates the possibility of applying such methods to real-world scenarios, such as continuous monitoring of stress levels.

Authors in [22] used data from Respiban with a recording frequency of 700hz . They used ECG and EDA as signals, using 30-second windows with 25% overlap ensuring a smooth transition between successive samples. The analysis of the features was done in both the time and frequency domains, incorporating features such as RMSSD and LF/HR ratio for HRV estimation. Regarding classification, they applied algorithms such as SVM, LightGBM, Random Forest, and XGBoost. They also performed binary stress-non-stress classification. The highest accuracy achieved was 88% with XGBoost. A result that demonstrates how effective the gradient boosting method is in stress detection.

A different approach was taken by Greco et al. in [13] , who proposed cvxEDA. A convex optimization method for EDA signal analysis. This approach allowed the distinction between tonic and phasic features, improving the modeling of sympathetic nervous system

responses. The method was applied to data with 60-second non-overlapping windows, allowing the analysis of dynamic electrodermal activity over long time intervals. The study evaluated the effectiveness of the algorithm in 3 different experimental sessions by examining their robustness to noise, their ability to distinguish and recognize stimuli, and their ability to correctly describe the activity of the autonomic nervous system in response to intense emotional arousal. However, this study did not focus on classification procedures (binary or multiclass) nor did it report relative accuracy rates.

A report that focused only on the ECG signal examining HRV features is [23]. More specifically, they collected data from 16 participants during real driving conditions. The HRV analysis included features such as maximum and minimum R-R interval, median and mean RRI, mean heart rate HR, standard deviation. Additionally, they dealt with features related to frequencies such as VLF, LF, HF, and LF/HF ratio. Finally, they dealt with nonlinear characteristics SD1 and SD2 (axes of the Poincare diagram). Using an SVM classifier with an RBF kernel, they achieved 83% accuracy in predicting driver stress levels. This study highlights the importance of HRV analysis for assessing stress while driving.

In the study [24] .The researchers proposed a simple, practical and cost-effective calibration technique for creating accurate and personalized stress prediction models. Data were collected from a large population, which was aimed at the subjectivity of stress detection and the diversity of reactions between individuals. The data were collected from 2 different data sets. Included physiological signals such as HRV and EDA. 7 Various features were extracted from these signals in both the time and frequency domains. For example, features such as the mean, median, standard deviation of RR intervals, RMSSD (root mean square difference of consecutive RR intervals) and pNNx (percentage of consecutive RR intervals that differ by more than x ms) were extracted from HRV. From the EDA, features such as mean, standard deviation, and statistics of galvanic skin response responses were extracted. The frozen data were processed to remove noise and balance classes. Finally, they trained general models using data from the entire population without individualization. These models achieved an accuracy of only 42.5% ± 19.9%. To improve performance, the researchers incorporated a small number (e.g., 100) of calibration samples from each individual into the general dataset. This hybrid approach resulted in a significant increase in accuracy, reaching 95.2% ± 0.5%.

Szakonyi et al. in [25] focused on the characteristics of HRV from the ECG. The study involved 7 people who underwent two stress tests. The Trier Social Stress Test and the Stroop Color-Word Test. ECG data were recorded with wearable devices and feature extraction was performed in 5-minute overlapping windows allowing detailed analysis of heart rate changes. The researchers extracted different fields for feature extraction. Such as in the time domain of frequency and some nonlinear features. Regarding classification, they tested

various machine learning algorithms including Random Forest, XGBoost Tree and Random Trees. The best results were achieved with the Random Forest algorithm, using HRV features in the time domain and nonlinear features, achieving an accuracy of 96.31% and an F1 score of 96.26%.

One of the few studies that dealt with the study of temperature is presented in [14].The researchers analyzed 420 different features from temperature as well as from ACC, EDA and BVP of RespiBAN and Empatica E4 devices. A total of 420 features were extracted from these signals using 30-s windows with 50% overlap. . To classify the data, the Logistic Regression algorithm was applied, achieving 75% accuracy in stress detection.

Accordingly, the research conducted in [26] evaluated a range of algorithms such as (Logistic Regression, Decision Tree, Random Forest, Stacking Ensemble Learning) using EDA, BVP and body temperature data from the Empatica E4 device. Features were extracted using 60-second windows with 50% overlap, optimizing the use of the data. For the classification they used algorithms such as logistic regression, decision tree which was applied to count the decisions that lead to the stress classification, Random Forest and stacking ensemble learning. In the personalized models, they achieved accuracy of up to 99%, while in the generalized ones, 91%, highlighting the importance of personalization in stress classification models.

As part of the research in [27], EDA, BVP, acceleration and respiration signals from the Empatica E4 and RespiBAN devices were analyzed. The researchers applied preprocessing and feature extraction techniques to detect stress, highlighting the importance of multimodal analysis. Feature extraction was performed on 60-second windows with 50% overlap, allowing for detailed recording of physiological changes associated with stress. Further experiments using deep learning models like GoogLeNet and SqueezeNet, combined with data overlapping techniques, boosted the accuracy of Random Forest up to 99.5%, demonstrating the strong potential of advanced signal processing and learning techniques for reliable stress detection.

Researchers in [28] developed a real-time system using data from wearable devices. Twenty-one subjects participated, from whom data were collected during daily activities, recording physiological signals such as electrocardiogram (ECG) and electrodermal activity (EDA). Preprocessing of the signals included noise filtering and feature extraction in 60-second windows with 50% overlap. For classification, the Random Forest algorithm was used, achieving 92% accuracy in stress detection.

A different study was conducted by Sano and Picard[29], using wearable devices to predict stress and sleep levels. Data was collected from 18 individuals, recording data such as EDA, body temperature, and movement. In the pre-processing part, the authors used 5-minute

windows and various signal smoothing techniques. Using linear regression, 75% accuracy was achieved in predicting stress.

Another study that concerned the detection of stress through physiological signals while driving was that of Healey and Picard.[30]. The data collected were ECG, EDA and respiration. Preprocessing included noise filtering and feature extraction in 60-second windows. For classification, the k-Nearest Neighbors algorithm was used, achieving 97% accuracy in stress detection.

In the study of Hernandez et al. [31], a stress detection system using wearable devices and behavioral analysis, was developed. Data were collected from 30 participants, recording signals such as heart rate, EDA, and movement. Preprocessing involved feature extraction in 2-minute windows with 50% overlap. Using the Support Vector Machine algorithm, 80% accuracy in stress detection was achieved.

Plarre et al.[32] examined stress detection via wearable devices in a work environment. Data were collected from 33 participants, recording signals such as ECG and EDA. Preprocessing included noise filtering and feature extraction in 5-minute windows. For classification, the Logistic Regression algorithm was used, achieving 85% accuracy in stress detection.

Table 2 Comparison of stress detection studies

Study	Signals Used	Window Size	Overlap	Feature Extraction	Classification Methods	Accuracy (%)
Schmidt et al. (2018) [3]	ECG, EDA, EMG, Respiration, Temp, ACC, BVP	60s	50%	Statistical, HRV	SVM, RF	93%
Ninh et al. (2022)[21]	EDA, BVP, Skin Temp	60s	50%	Statistical (Mean, Variance, Skewness)	MLP	86%
Rashid et al. (2021)[11]	BVP	30s	50%	Auto & Manual Feature Selection	CNN	85%
Islam et al. (2023) [12]	EDA	60s	Not Stated	Not Stated	Self-Supervised Learning (CNN)	80%
Zhou et al. (2024)[22]	ECG, EDA	30s	25%	Time & Frequency Domain Features (RMSSD, LF/HF Ratio)	SVM, LightGBM, RF, XGBoost	88%
Greco et al. (2016) [13]	EDA	60s	0%	Tonic & Phasic Features	Not Stated	Not Stated
ICABME (2015)[23]	ECG	5min	50%	HRV Analysis (RRI, Mean HR, SD)	SVM (RBF Kernel)	83%
Nkurikiyeyezu et al. (2020) [24]	HRV, EDA	Varied	Varied	HRV & EDA Statistical Features	Random Forest, ExtraTrees	95.2% ± 0.5%
Szakonyi et al. (2021) [25]	HRV	5min	50%	HRV Features	Random Forest	96.31%
Fauzi et al. (2022) [14]	Temp, ACC, EDA, BVP	30s	50%	Statistical Features	Logistic Regression	75%
Garg et al. (2021)[26]	EDA, BVP, Body Temp	60s	50%	Statistical Features	Stacking Ensemble Learning	99% (Personalized) / 91% (Generalized)
Campanella et al. (2023)[27]	EDA, BVP, ACC, Respiration	60s	50%	Multimodal Analysis	Not Stated	Not Stated
Gjoreski et al. (2016)[28]	ECG, EDA	60s	50%	HRV & EDA	Random Forest	92%
Sano and Picard (2013)[29]	EDA, Body Temp, Movement	5min	0%	Signal Smoothing	Linear Regression	75%
Healey and Picard (2005)[30]	ECG, EDA, Respiration	60s	50%	Feature Extraction from ECG & EDA	k-NN	97%
Hernandez et al.(2011) [31]	Heart Rate, EDA, Movement	2min	50%	Time-domain & statistical features	SVM	80%
Plarre et al.(2011)[32]	ECG, EDA	5 min	0%	Time & frequency-domain features	Logistic Regression	85%

2.2. Comparison of existing datasets for stress detection

Different datasets have been involved in the study and detection of stress through physiological signals in addition to WESAD[3], SWELL-KW[33], Affective ROAD[34], DREAMER[35], DEAP[36] and AMIGOS[37] are 5 of the most widespread datasets that have been used in research to identify various emotional states. Each of them has different characteristics and applies different methodologies for data collection.

2.2.1. WESAD – Analysis and applications

WESAD (wearable Stress and Affect Detection) is considered one of the most recognized public datasets related to the classification of stress. The purpose of its creation is to develop and evaluate methods for stress detection using different physiological signals.

The experimental protocol used includes 15 participants, who were invited to a controlled data collection process that included different emotional states. The sensors used during the experiment were two: the Empatica E4 on the wrist and the RespiBan on the chest. The measurements were collected in a controlled environment where the participants were asked to face different emotional and physiological states.

The experiment included four different conditions.

Baseline: The participants were at rest without external stimuli.

Stress: Stress induction through the trier social stress test (TSST).

Amusement: Participants are exposed to funny videos to induce a feeling of amusement.

Meditation: Guided meditation session in order to relax the participants.

The role of the Trier Social Stress Test

The Trier Social Stress Test is a widely used experimental method that aims to induce stress in a controlled environment. In the WESAD dataset, the TSST was used to create a reliable and objective stress state. This process subjects' participants to solving numerical problems under time pressure as well as public speaking. These activities are considered to be a normal stress response [3]

2.2.2. Public datasets – Characteristics and limitations

1.SWELL-KW: dataset focuses on detecting stress in a work environment. It involves 25 different participants, who performed different activities in their work environment, such as writing reports and reading emails. During the experiments, they introduced email interruptions and time pressure as stressors. Stress levels were measured through EDA,

ECG, HRV and ACC. A sampling frequency of 128 hz was used to record the data, and 30s windows with 50% overlap were used. The main problems that exist are the limited number of participants and the artificial induction of stress that may not be fully reflected in real conditions of work stress.[33]

2. Affective ROAD: Focuses on detecting emotional stress in real driving conditions. There were 10 participants, and they were subjected to different driving scenarios, such as urban and highway driving. Data were collected from Zephyr Bio harness 3.0 and Empatica E4 devices. They include EDA, BVP, HRV, ECG and facial recognition video. The sampling frequency is 64hz and 50s windows with 50% overlap were used. The only problem we face is the alteration of physiological signals by noises coming from external factors making the analysis of the signals difficult.[34]

3. DREAMER: is a multimodal dataset that focuses on the analysis of stress through EDA, BVP and ECG. The 23 participants watched videos and listened to music while their physiological signals were recorded. The sampling frequency is 128hz and the data were processed with 30s windows and 50% overlap. The stimulation of emotions was carried out in a controlled laboratory environment, which may not be fully reflected in real conditions.[35]

4.DEAP: is one of the most widely used datasets for emotion recognition that concerns the neurophysiological response of humans to visual and auditory stimuli. The 32 participants were exposed to 40 emotionally charged videos. The recordings include EDA, BVP, HRV and EEG, with a sampling frequency of 512hz for ECG and 128hz for the other signals. Its main drawback is the lack of data from real stress scenarios.[36]

5. AMIGOS: focuses on emotion recognition at both the individual and group levels. It includes 40 participants who watched short and long videos. EDA, BVP, HRV, EEG and facial expression data were recorded with a sampling frequency of 128hz. The ability to collect data from interactions in a group environment makes it unique but presents challenges in terms of the consistency of subjective emotional responses.[37]

Although each dataset has its own advantages, the main limiting factors to consider include the method of stress induction, the number of participants, and whether the data were collected in laboratory or real-world conditions. The WESAD dataset is considered more comprehensive in terms of multimodality and accuracy of recording physiological signals, making it ideal for stress detection applications via EDA and BVP.

Table 3 Overview of publicly available stress dataset

Dataset	Participants	Signals	Sampling frequency	Recording Environment	Stress Induction
WESAD	*15*	*EDA, BVP, ECG, Respiration, ACC*	*64 Hz (BVP, EDA), 700 Hz (ECG)*	*Laboratory*	*TSST (Trier Social Stress Test)*
SWELL-KW	*25*	*EDA, ECG, HRV, ACC*	*128 Hz*	*Work environment*	*Cognitive load & pressure conditions*
Affective ROAD	*40*	*EDA, BVP, HRV, EEG, Face recognition video*	*64 Hz*	*Driving (real-world conditions)*	*Traffic pressure scenarios*
DREAMER	*23*	*EDA, BVP, EEG*	*128 Hz*	*Laboratory*	*Audiovisual stimuli (video, music)*
DEAP	*32*	*EDA, BVP, HRV, EEG*	*512 Hz (EEG), 128 Hz (*	*Laboratory*	*Emotional videos*
AMIGOS	*40*	*EDA, BVP, HRV, EEG, facial expression*	*128 Hz*	*Laboratory & Social interactions*	*Interaction of people*

3. Methodology

In this study, stress recognition was approached through the analysis of physiological signals using machine learning techniques, with data sourced from the WESAD dataset. The primary focus was on the Electrodermal Activity (EDA) and Blood Volume Pulse (BVP) signals, which have been extensively studied for their correlation with autonomic nervous system responses to stress. [11], [12], [13], [21], [26]. The methodology followed a structured pipeline, including data preprocessing, feature extraction, and classification, to effectively distinguish between different physiological states such as Baseline, Stress, Amusement, and Meditation. Various feature selection techniques were applied to optimize the most relevant physiological characteristics, while multiple machine learning algorithms, including Random Forest, Gradient Boosting, SVM, KNN, Logistic Regression, and Naïve Bayes, were evaluated for their classification performance. We examined the impact of different time windows, feature subsets, and overlapping segmentation strategies on classification accuracy. The results demonstrated that Random Forest and Gradient Boosting achieved the highest accuracy in binary classification tasks (95.97% and 94.31%) respectively, while multiclass classification remained a more complex challenge, with performance varying across different algorithms. The findings highlight the significance of signal preprocessing and feature selection in enhancing stress recognition accuracy, paving the way for the development of more efficient, real-time stress monitoring systems based on wearable sensor data.

3.1 Data preprocessing

Perhaps one of the most critical stages in the analysis of physiological signals is data preprocessing as it is directly related to the quality of the features to be extracted and therefore the performance of the classification numbers. The need for preprocessing arises from the fact that data collected through sensors often contain noise, data loss and various types of distortions due to user movement. This has the consequence that the use of appropriate techniques is considered necessary as it improves the overall quality of the signals making them more reliable for further analysis [38], [39], [40], [41]

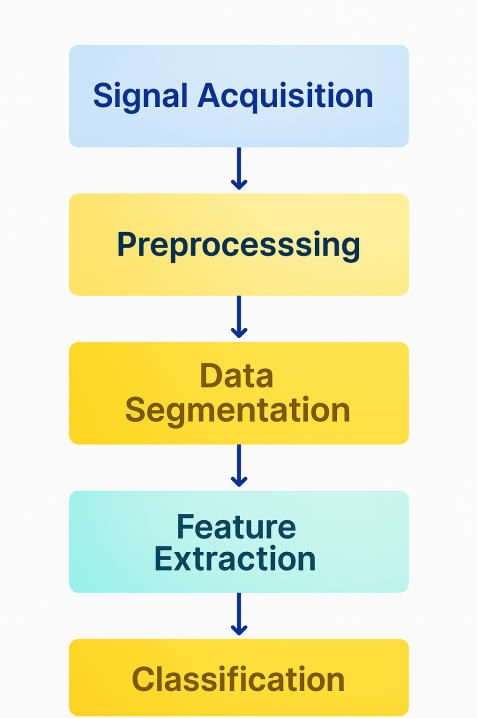


Figure 2 Overview of stress recognition process

3.1.1 Signal filtering

The data collected by portable sensor devices most often contain interference and unwanted components that may affect the subsequent processing of the signal. Signal filtering often aims to remove noise from the analysis and isolate useful information. There are various types of noise and different methods for dealing with them which will be analyzed.

Noise therefore comes from different causes such as environmental noise which may come from electromagnetic interference (e.g. electronic devices)[42] . From the user's movement as the sensors placed on it may record abnormal changes due to it[43] . Finally, the appearance of low-intensity distortions due to changes in body temperature or skin conductivity.

Various filtering methods are used to remove these effects.

- **Low-pass filters:** eliminate high frequencies and abrupt changes caused by interference
- **High-pass filters:** remove low frequencies allowing more significant changes in the signal to be highlighted

• **Band-pass filters:** are a combination of the low-pass and High-pass filters that isolate a specific frequency range

• **Wavelet transform** used for more advanced denoising especially in living signals

• **Moving average filter**: a basic technique that helps to smooth small changes in the signal

the choice of the appropriate method depends on the type of signal and the desired level of smoothing as mentioned in [34].

3.1.2. Data Segmentation

Data segmentation is a crucial step in the preprocessing of physiological signals as it allows us to divide a continuous signal into smaller ones which are more manageable and allow us to further analyze the signal. We have 2 approaches regarding segmentation which are event-based segmentation and sliding window segmentation.[44], [45], [46]

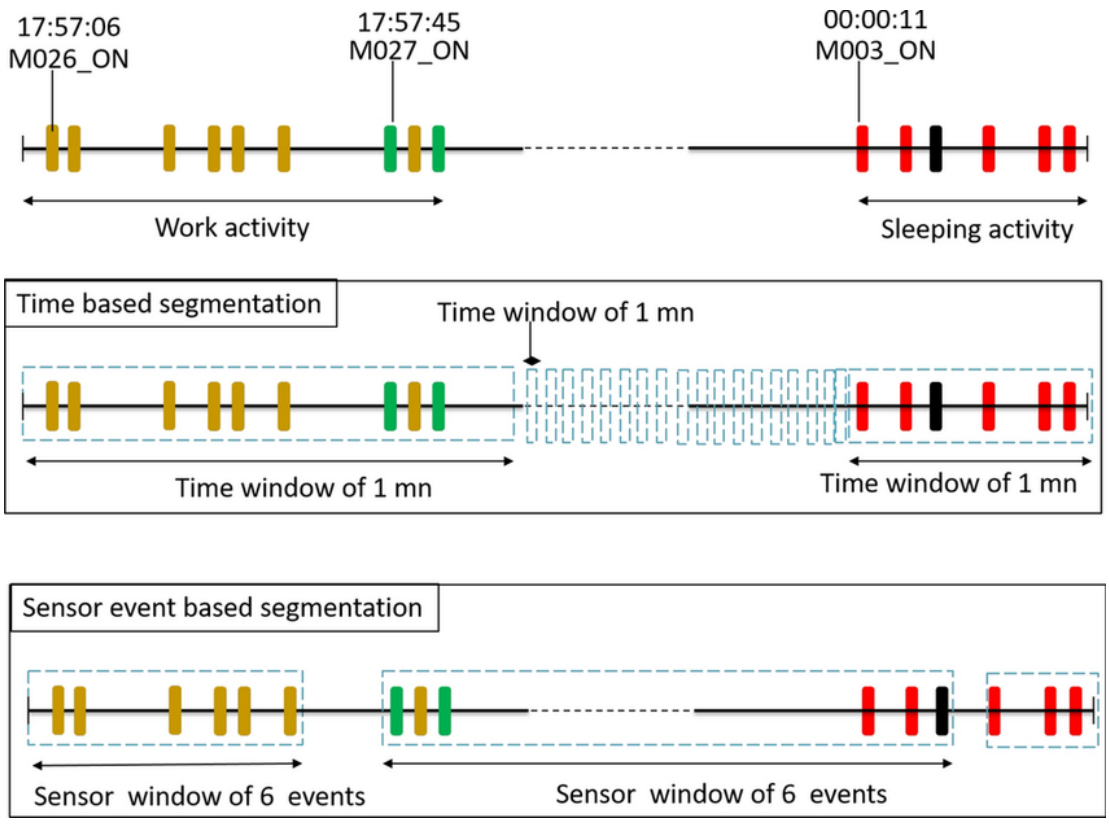


Figure 3 Overview of time and event-based segmentation. Figure taken from [47]

The choice of method depends on different factors and the nature of the problem. An analytical table (Table 4) with the advantages and limitations of the mentioned methods.

Table 4 Comparison of event-based and sliding window segmentation methods, including their use cases, advantages, and restrictions

Method	When is it used	Advantages	Restrictions
Event-Based Segmentation	When the data has clear events (e.g. heartbeats, neural spikes)	Focuses on critical signal changes	It needs accurate event detection algorithms
Sliding Window Segmentation	When signals do not have clear changes, but analysis is required at specific time intervals	Simpler implementation, less processing power	Can cut off critical signal changes

3.1.2.1. Event-based segmentation

In event-based segmentation, the segmentation of a signal is done according to how it is detected when specific events or changes in the signal are detected. This method is particularly useful when the interesting features of the signal appear at irregular intervals. A typical example is in the analysis of the EEG, where signal segmentation can be based on abrupt changes in or specific patterns that indicate the onset of an event [48].

Event-based segmentation is mainly applied to systems with non-stationary abrupt changes where the occurrence of an event is what defines the boundaries of the signal segments. Such cases are often encountered in electrocardiograms, electroencephalograms, thermal signals, and respiratory signals. For example, in the cardiogram, signal segmentation can be done based on the onset of the heartbeat after the detection of R-peaks. In the electroencephalogram, it can be separated when epileptic seizures are detected or increased brain activity [49], [50].

The signal segmentation process using this method involves a series of structured steps. Initially, the input signal is continuously compared with a reference channel in real time to detect distinct features indicative of abrupt fluctuations, discontinuities, or recurring patterns. Upon identifying such an event, it is marked as a potential boundary—either a start or end point—for a new segment. These boundaries are then used to divide the signal into distinct portions, enabling a more targeted and effective analysis of each segment individually [51].

3.1.2.2. Sliding window segmentation

In window segmentation, a continuous signal is divided into overlapping or non-overlapping segments of fixed length, also known as windows. This method is widely used due to its simplicity and its ability to capture local features of the signal. The choice of length and the percentage of overlap can vary and depends on the physiological signal we are studying. Research has shown that overlapping windows can improve classification accuracy in applications [52].

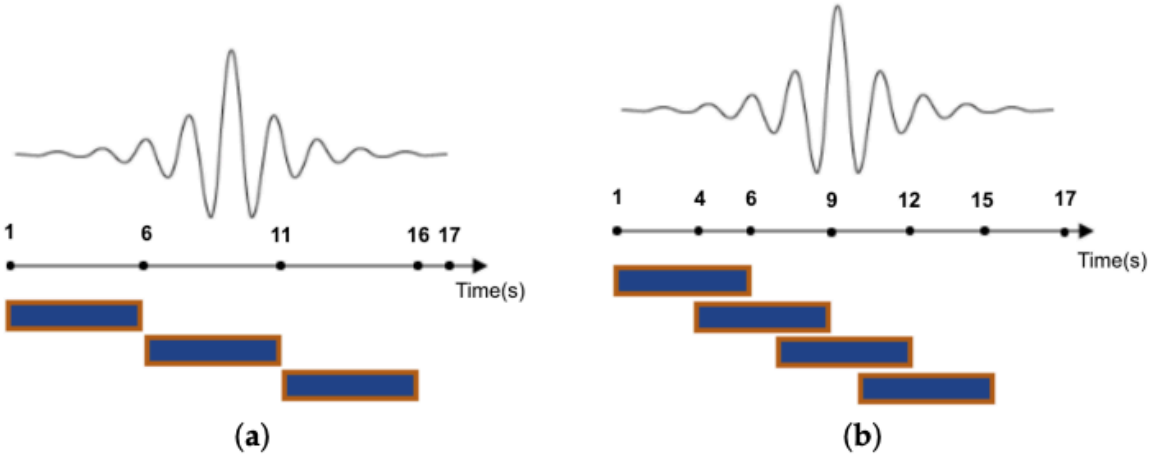


Figure 4 Overview of non- overlapping(a) and overlapping windows (b). Figure taken from [52]

According to studies [49], [50], [52] overlapping windows increase the number of data available for analysis, which can lead to better performance of classifiers. However, excessive overlap can increase the cultural load without significant performance improvement. Therefore, choosing the appropriate overlap rate is critical to optimizing system performance. We may conclude that the choice of overlap rate is critical for the performance of both physiological signal analysis systems. Batch overlap requires careful tuning to avoid unnecessary increases in computational load. The optimal choice depends on the specific signal and the application.

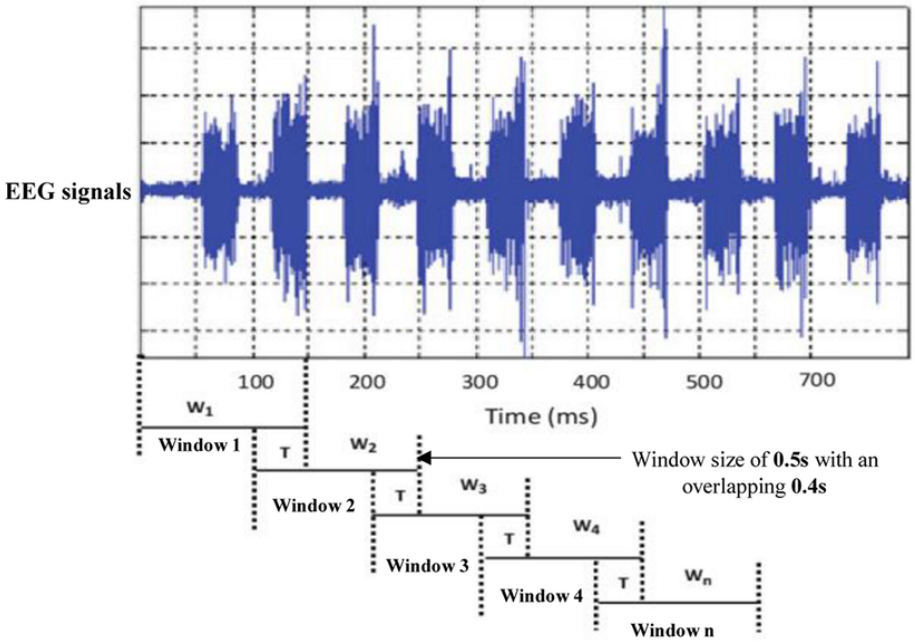


Figure 5 Overview of overlapped window segmentation for EEG signal. Figure taken from [53]

Non-overlapping windows

In this case the continuous signal is divided into successive fixed-length segments without any overlap between them. Each segment is analyzed independently which reduces the computational burden for the complexity of the processing however this approach may not fully capture the transient to dynamic changes that occur at the window boundaries leading to possible loss of important information.

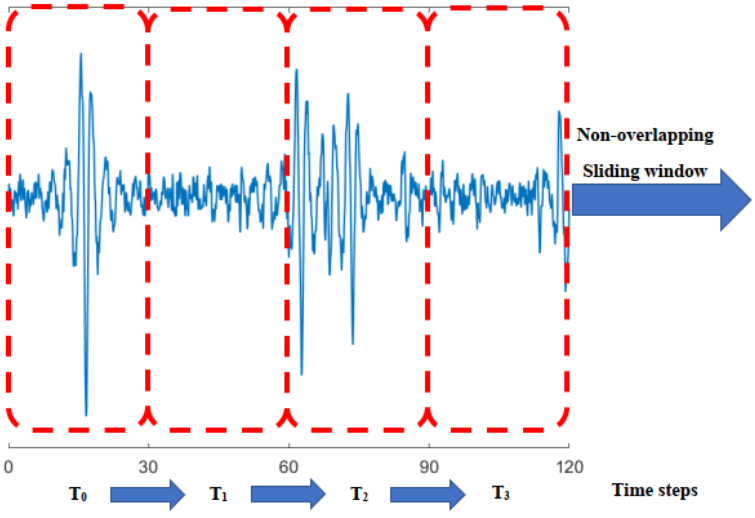


Figure 6 Overview of non-overlapping windows. Figure taken from [54]

3.2. Feature Extraction

One of the most critical steps in physiological signal analysis is feature extraction as it allows us to extract important information from the data collected through sensors. This process helps to reduce the complexity of the data while preserving the essential elements necessary for further analysis and development of machine learning models. Feature extraction can be done based on time, frequency, or in some cases, spatial analysis where the data is multidimensional [55], [56], [57]. Effective feature extraction plays a vital role in the success of applications such as emotion recognition, medical diagnosis, and real-time monitoring of an individual's physiological state. When recognizing emotions from physiological signals, the careful selection of relevant features can substantially enhance the performance and accuracy of machine learning models used for classifying emotional states.[58]

3.2.1. Time Domain Features

Time domain features are among the most important and widely used in the analysis of physiological signals. Features are extracted directly from the signal in time without requiring transformation to another frequency domain.

Their extraction enables us to analyze various statistical and dynamic properties of the signal which can be used in applications such as classifying emotional states or detecting pathological states[58], [59],[60]

3.2.1.1. Statistical and Mathematical Functions

Some of the most important features will be analyzed as well as some mathematical functions will be given.

• Mean: Is the arithmetic average value of the signal over a period.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

• Median: Is the middle value that is not affected by extreme values (outliers).

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}}, & \text{if n is odd} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & \text{if n is even} \end{cases}$$

• Standard deviation: Shows how much the samples of the signal differ from the average value.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2} \ [61]$$

• Variance: The square of the standard deviation which shows the overall variability of the signal.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \ [61]$$

• Min/Max: The extreme values that the signal takes in a specific time window.

$$\max(x_1, x_2, \dots, x_n) \ [61]$$

• Range: The difference between the maximum and the change in the value.

$$\text{Range} = \max(x_1, x_2, \dots, x_n) - \min(x_1, x_2, \dots, x_n) \ [61]$$

• Root mean square: The square of the signal which shows the overall strength of the signal.

$$\text{RMS} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2} \ [62]$$

Correlation and Signal Analysis Features

• **Correlation**: shows the degree of similarity between 2 signals

• **Cross-correlation**: measures the similarity between 2 signals the map of your time shift to the other

• **Autocorrelation**: the cross-correlations of a signal with itself therefore the relationship always shows a peak at time shift zero and their size represents the signal

• **Linear correlation** coefficient: determines the legal relationship between 2 signals. A value close to +1 or -1 means a high linear correlation while a value close to 0 means that the signals have no linear correlation.

Statistical Distribution Characteristics

• **Kurtosis**: measures the sharpness of the distribution of the signal values; if the value is high, the signal has more pronounced peaks.[63]

$$K = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} [61]$$

• **Skewness**: measures the degree of symmetry of the signal distribution. If the value is close to zero, the distribution is symmetric, while if it is positive or negative, it indicates whether the signal has larger values in one direction or the other [63].

$$S = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}} [61]$$

• **Mode**: the value that appears most often in the signal.

• **Trimean**: the average calculated by ignoring a percentage of the extreme values of the signal.

$$\text{Trimean} = \frac{Q_1 + 2Q_2 + Q_3}{4} [61]$$

• **Asymmetry coefficient**: is calculated as the first human pile of the data divided by the standard deviation.

$$\gamma_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}} [61]$$

• **Integration**: measures the area under the signal curve, which can be used to estimate the total energy of the signal.

$$\int_a^b f(x)\ dx\ [61]$$

3.2.1.2. Additional Computational Methods

Signal Variation Analysis Methods

In the analysis of physiological signals, beyond the basic statistical features, computational methods are used and added to extract more complex information. These methods allow for a better analysis of signal variations and dynamics, improving the accuracy of classification and classification of physiological data [37 - 40].

• **Zero crossing rate:** The total number of signal transitions from and to negative values and vice versa, normalized to the length of the window

$$\text{ZCR} = \frac{1}{N-1}\sum_{i=1}^{N-1} 1[(x_i \cdot x_{i+1}) < 0] \text{ [64]}$$

• **Mean crossing rate**:

The number of times the signal changes from below its mean value normalized to the length of the window.

$$\text{MCR} = \frac{1}{N-1}\sum_{i=1}^{N-1} 1[(x_i - \bar{x}) \cdot (x_{i+1} - \bar{x}) < 0] \text{ [64]}$$

• **Differences**

Difference between consecutive samples which can be used to estimate the variability and dynamic changes of the signal.

$$\text{Mean Difference} = \frac{1}{N-1}\sum_{i=1}^{N-1} |x_{i+1} - x_i| \text{ [64]}$$

Motion & Tilt Analysis Methods

• **Angular velocity**

The rate of change of angular displacement, which determines the rotation speed of an object and its axis of rotation

$$\omega_i = \frac{\theta_{i+1} - \theta_i}{t_{i+1} - t_i} \text{ [64]}$$

• **Tilt angle**

The angle of inclination of the Gottsche plane and the angle of inclination with respect to the vector axis of gravity, calculated through the acceleration in the 3 axes.

$$\theta_y = \tan^{-1}\left(\frac{a_y}{\sqrt{a_x^2 + a_z^2}}\right) \text{ [64]}$$

Signal Magnitude Analysis Methods

• **Signal magnitude vector**: The sum of the Euclidean measure of the signal in the 3 axes normalized to the length of the window.

$$\text{SMM} = \sum_{i=1}^{N} \sqrt{x_i^2 + y_i^2 + z_i^2} \text{ [64]}$$

• **Normalized signal magnitude area**(**sma**): The sum of the magnitude reached in the 3 axes normalized to the length of the window.

$$\text{NSMA} = \frac{1}{N} \sum_{i=1}^{N} \frac{|x_i| + |y_i| + |z_i|}{3} \text{ [64]}$$

3.2.2. Frequency Domain Features

The analysis of physiological signals in the frequency domain provides important information that is not easily identified in the time domain. The features obtained from the frequency domain and from the spectral analysis of the signal are widely used in applications such as pattern recognition, pathology detection and signal classification [60], [65], [66].

Some of them are

- Shannon Energy: Measures the entropy of the signal's power distribution.

$$SE = -\sum_{i=1}^{N} p_i \log_2(p_i), \quad \text{where } p_i = \frac{x_i^2}{\sum_{i=1}^{N} x_i^2} \text{ [67]}$$

- Log Energy: Measures the total logarithmic energy of the signal.

$$LE = \sum_{i=1}^{N} \log(x_i^2) \text{ [67]}$$

- Skewness: Measures the asymmetry of the signal's distribution.

$$S = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^3}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2\right)^{3/2}}, \quad \text{with } \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \text{ [67]}$$

3.2.2.1. Fourier Transformations

The Fourier Transform (FT) is a fundamental technique that transforms a signal from the time domain to the frequency domain, allowing the analysis of its synchronous components. The most widespread implementation is the fast Fourier transform, which is used for efficient processing of large signals. Different types of Fourier transforms are distinguished, along with their mathematical equations.[68], [69], [70]

Continuous Fourier Transform (CFT) Defined by the equation:

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi ft}\, dt \text{ [67]}$$

where:

- x(t): is the signal in the time domain,
- X(f): is the signal in the frequency domain,
- F: is the frequency, and
- J: is the complex unit.
- Discrete Fourier Transform (DFT)

Discrete Fourier Transform (DFT)

The numerical version of FT, which is used for digital signals is given by

where N is the number of samples.

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N}, \quad k = 0,1, \dots, N-1 \text{ [67]}$$

Power Spectral Density – PSD:

Expresses the distribution of the signal power in the frequency domain, with the following mathematical equation

$$S_X(k) = \frac{1}{N}|X_k|^2, \quad k = 0,1, \dots, N-1 \text{ [67]}$$

Dominant Frequency – DF:

f is the frequency with the maximum power in the signal spectrum.

$$f_{\text{dom}} = \frac{\arg\max_k |X_k|}{N} f_s \text{ [67]}$$

Mean Frequency – MF:

It is the weighted average of the frequencies. With mathematical equation:

$$f_{\text{mean}} = \frac{\sum_{k=0}^{N-1} f_k S_X(f_k)}{\sum_{k=0}^{N-1} S_X(f_k)} \text{ [67]}$$

These methods find applications in EEG analysis for the detection of epileptic seizures, andin ECG analysis for detecting normal or pathological cardiac activity. They also have wide use in HRV (Heart Rate Variability) analysis for *for evaluating cardiac autonomic nervous system* (ANS) activity.

3.2.2.2. Wavelet Transforms

Wavelet transform is a signal analysis method that allows analysis in both the temporal and spatial domains, in contrast to the Fourier transform, which provides only frequency information. Is used in the detection of transients and abrupt changes in signals. Next, we will analyze 3 different types of wavelets transforms [71], [72], [73].

Continuous Wavelet Transform – CWT

is used to analyze a signal in both time and frequency domains by applying scaled and shifted wavelet functions**.**

$$W(a,b) = \int_{-\infty}^{\infty} x(t)\psi_{a,b}^{*}(t)\ dt [74]$$

Discrete Wavelet Transform – DWT

DWT is used in the analysis of digital signals through a hierarchical multi-resolution method

$$W_{j,k} = \sum_{n=0}^{N-1} x_n \psi_{j,k}(n)\ [74]$$

Wavelet Spectral Entropy – WSE

WSE measures the disorder in the wavelet spectrum

$$\mathrm{WE} = -\sum_{j=1}^{J} P_j \log P_j\ \ [74]$$

These analyses find applications in various fields of biomedical engineering. In EEG analysis, features are used to detect abnormalities such as epileptic seizures, allowing for better diagnosis and monitoring of electoral disorders. Regarding cardiac activity and EEG processing, particular emphasis is given to the previous methods to detect abnormalities in cardiac activity, contributing to the early diagnosis of heart diseases and arrhythmias. Finally, regarding EDA, changes in skin conductance are related to the activation of the sympathetic nervous system, which allows us to detect emotional reactions.

3.2.3. Discrete Domain Features

Discrete domain features refer to properties extracted from signals that have been sampled and quantized, converting them from continuous to discrete form. The process allows the application of digital processing techniques for the analysis and interpretation of signals. One of the most basic features is the autocorrelation function which measures the similarity of a signal to its shifted versions at different time delays. The function allows us to detect periodicities and estimate noise in the signal [75].

Additionally, the Discrete Cross-Correlation Function is used to measure the similarity between 2 different signals, which allows us to detect delays or advances between them [76].

Regarding the discrete power spectral density, it provides information about the distribution of the signal power in the frequency domain, helping to identify dominant frequencies and analyze the energy content of the signal [76].

Finally, the discrete wavelet transform allows the analysis of the signal at different scales and frequencies, offering a multidimensional approach to detect features that may not be apparent in the time-frequency domain [77].

3.3. Feature Selection

Feature selection is perhaps the most critical process in machine learning for processing natural signals. The main goal is to reduce the number of model inputs while retaining only the most important features that contribute to the distinction of categories. The benefits of the feature selection process are multiple. First, we should mention the reduction of the dimension, which minimizes the complexity of the data, reducing the cost of training the algorithms. In addition, the use of fewer but more important features improves the accuracy of the model, resulting in higher accuracy. It also reduces the chances of the model being overfitted [78],[79],[80].

3.3.1. Filter Methods

Each feature is evaluated independently of the learning model based on statistical criteria and classification techniques. These methods are suitable for large amounts of data as they do not require repeated training of algorithms and as a result, have lower computational costs. The most common methods include:[78], [81], [82]

Pearson Correlation Coefficient [81]**:**

values close to plus or minus one indicate a strong linear relationship.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} [81]$$

• **Mutual Information** (MI) [81]

measures the amount of information shared between 2 variables even in nonlinear relationships.

$$I(X;Y) = \sum_{x \in X}\sum_{y \in Y} p(x,y) \log \quad \frac{p(x,y)}{p(x)p(y)} [81]$$

• **Chi-Square Test** [81]

is used for categorical data to measure the relationships between a feature and the target variable.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} [81]$$

• **ANOVA F-score** [81]

higher value of f means more important features

$$F = \frac{\frac{S_{\text{between}}}{k-1}}{\frac{S_{\text{within}}}{N-k}} [81]$$

3.3.2. Wrapper Methods

Wrapper Methods use machine learning algorithms to evaluate various combinations of features by selecting the subset that improves the performance of the model. Some of the most popular techniques are: [82], [83].

• Recursive Feature Elimination (RFE)[84], and

• Additive and Subtractive Selection (Forward/Backward Selection) [85].

3.3.3. Embedded Methods

Embedded Methods combine feature selection with the training process of the learning model. This means that the model itself determines which features are most important during its training, selecting the dynamics. Embedded methods are superior to filter and wrapped methods because they offer a balanced model, considering the relationship between features while minimizing computational complexity. These methods are particularly useful when the dataset is large and contains redundant or irrelevant features

that can degrade the performance of the model. Some of the main advantages of Embedded Methods are time savings as they do not require additional steps for feature selection. Improved accuracy as they focus on the most critical features, reducing the possibility of overfitting, and finally reduced complexity as they automatically select the features that have the greatest contribution to classification [80], [84], [85].

Examples of Embedded Methods

1. Lasso Regression (L1 Regularization)

Lasso Regression (Least Absolute Shrinkage and Selection Operator) uses L1 regularization, a penalty term that shrinks the coefficients of the least important features towards zero. This results in automatic feature selection, as features that do not contribute to the model are assigned a coefficient of zero and are discarded.

$$\hat{\beta} = \arg\min_{\beta}\left(\sum_{i=1}^{n}(y_i - X_i\beta)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right)[86]$$

When λ is large, more features are discarded, while when it is small, the model retains more features.

2. Ridge Regression (L2 Regularization)

Ridge Regression applies L2 regularization, that is, it adds a penalty to the size of the model coefficients, preventing overfitting.

$$\hat{}\{\beta\} = \arg\min_\{\beta\} \vdash(\sum_\{i=1\}^{n}(y_i - X_i\,\beta)^2 + \lambda\sum_\{j=1\ [86]$$

Key difference between Ridge and Lasso:

• Lasso (L1): Zeroes coefficients and removes features.

• Ridge (L2): Reduces coefficients but does not zero them.

3. Elastic Net Regression

Elastic Net Regression is a combination of Lasso and Ridge Regression, using both regularization terms. It is especially useful when we have many correlated features.

$$\hat{\beta} = \arg\min_{\beta}\left(\sum_{i=1}^{n}(y_i - X_i\beta)^2 + \lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{j=1}^{p}\beta_j^2\right)[86]$$

4. Decision Trees and Random Forests

Decision Trees and Random Forests incorporate feature selection, as they use an information criterion (e.g., Gini impurity or entropy) to select the most important features at each level of the tree.

$$G = \sum_{i=1}^{c} p_i(1 - p_i) = 1 - \sum_{i=1}^{c} p_i^2 \text{ [87]}$$

5. Gradient Boosting Machines (GBM, XGBoost, LightGBM)

Gradient Boosting Models, such as XGBoost and LightGBM, incorporate feature selection, scoring features based on their contribution to the model.

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \text{ [87]}$$

In conclusion, embedded methods are the most efficient approach for feature selection as they do not consider the structure of the data and select the most important features during model training

3.4. Classification

Classification is a fundamental process in machine learning where models are trained to categorize data into predefined categories based on specific features. This technique is very useful in areas such as biomedical signal analysis, emotional state detection, activity recognition and many others. The model is trained to assign observations to predefined categories based on a set of features.

According to the study by Kotsiantis et al.[88], [89], classification is one of the most studied methods in machine learning, with many applications in various fields. Additionally, classification methods have been applied to text analysis such as the automatic categorization of research articles where different learning techniques are evaluated for their performance in content classification. Overall, classification is a critical tool in machine learning with applications ranging from medical diagnosis to natural language processing.[90]

3.4.1. Categories of Machine Learning Algorithms

Machine learning includes different categories of algorithms, the two main ones being supervised learning and unsupervised learning.[89]

3.4.1.1. Supervised Learning Algorithms

In supervised learning algorithms, the model is trained using data that includes both the inputs and the corresponding outputs (labels). The goal is to learn a function that can predict

the label for new unknown inputs. The main types of problems in supervised learning are classification and regression. In terms of classification, labels are distinct categories such as spam detection, image recognition or in our case, state of being or not. In contrast, in regression, labels are continuous values, such as in real estate price forecasting or energy demand modeling.[91], [92]

Some of the most common supervised learning algorithms are as follows.

Support vector machines (SVMs): are used to create hyper-layers that separate categories in data. They are particularly effective in high dimensions and use techniques such as kernel functions to classify non-linearly separated data.

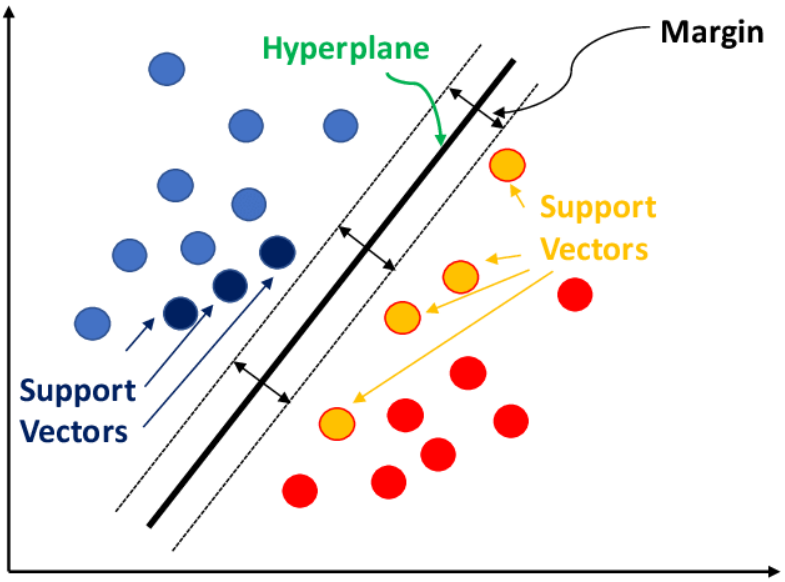


Figure 7 Overview of SVM algorithms. Figure taken from [93]

Decision trees: create a tree-like model to make decisions based on data features. Decision trees select successive features to separate categories, maximizing the information gained.

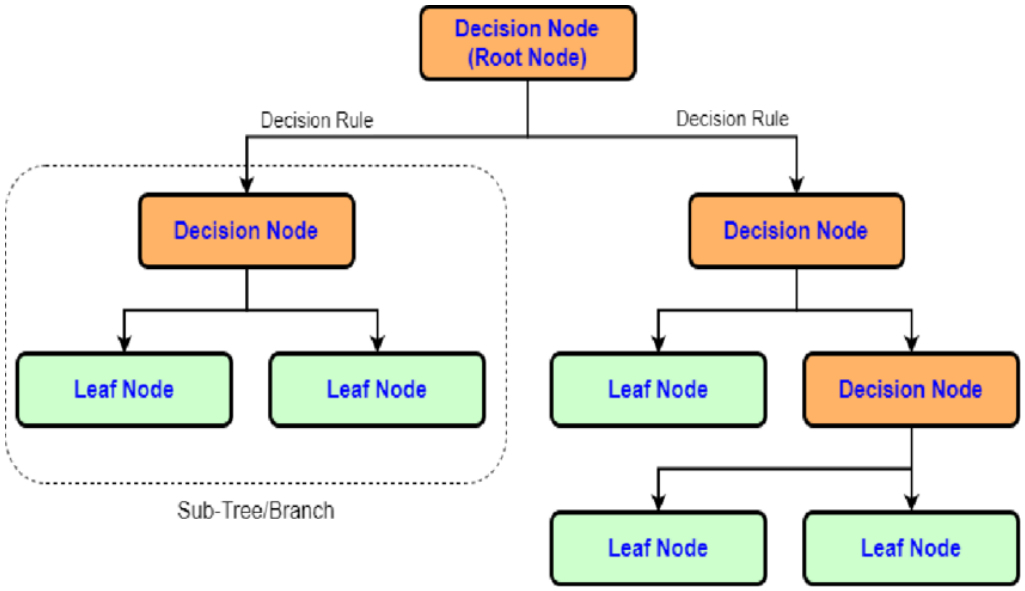


Figure 8 Overview of Decision Trees. Figure taken from [94]

Random Forests (RF). RF combine multiple decision trees to reduce accuracy and avoid overfitting. They use the bagging technique to create a variety of models from different parts of the training data.

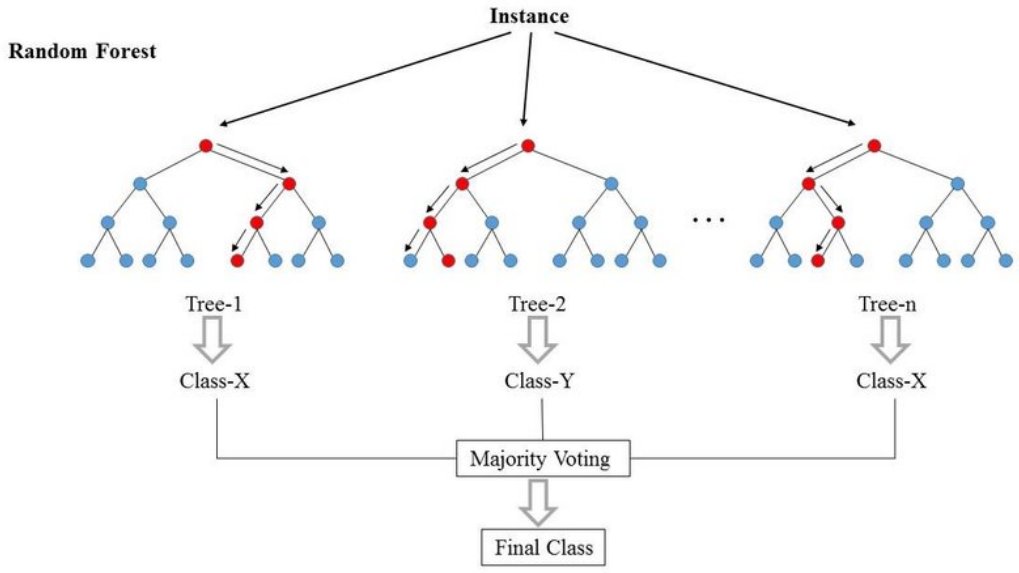


Figure 9 Overview of Random Forest algorithm. Figure taken from [95]

Neural Networks (NN). NN are models inspired by the structure of the human brain. NN could learn complex patterns and are divided into 3 main categories: Simple artificial neural

networks, convolutional neural networks, and recurrent neural networks for image processing and natural language processing applications.

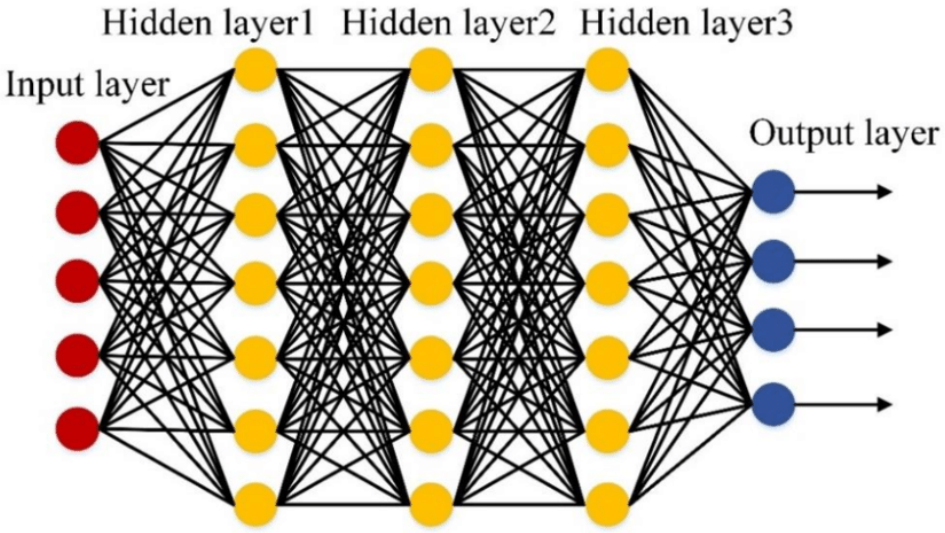


Figure 10 Overview of Neural networks algorithms. Figure taken from[96]

Gradient Boosting Machines. They represent a technique that builds models sequentially, with each new model correcting the errors of the previous one. Popular implementations include XGBOOST, LIGHTGBM, and CATBOOST. They are highly efficient on large datasets and are used extensively in data analysis.

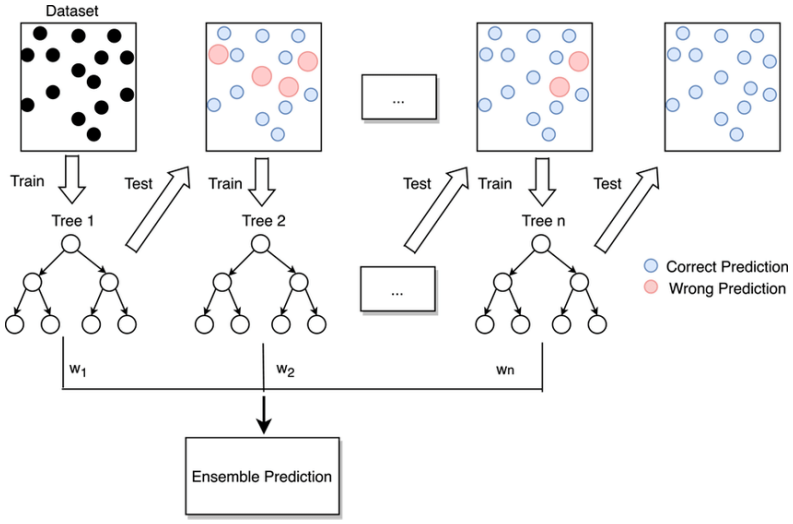
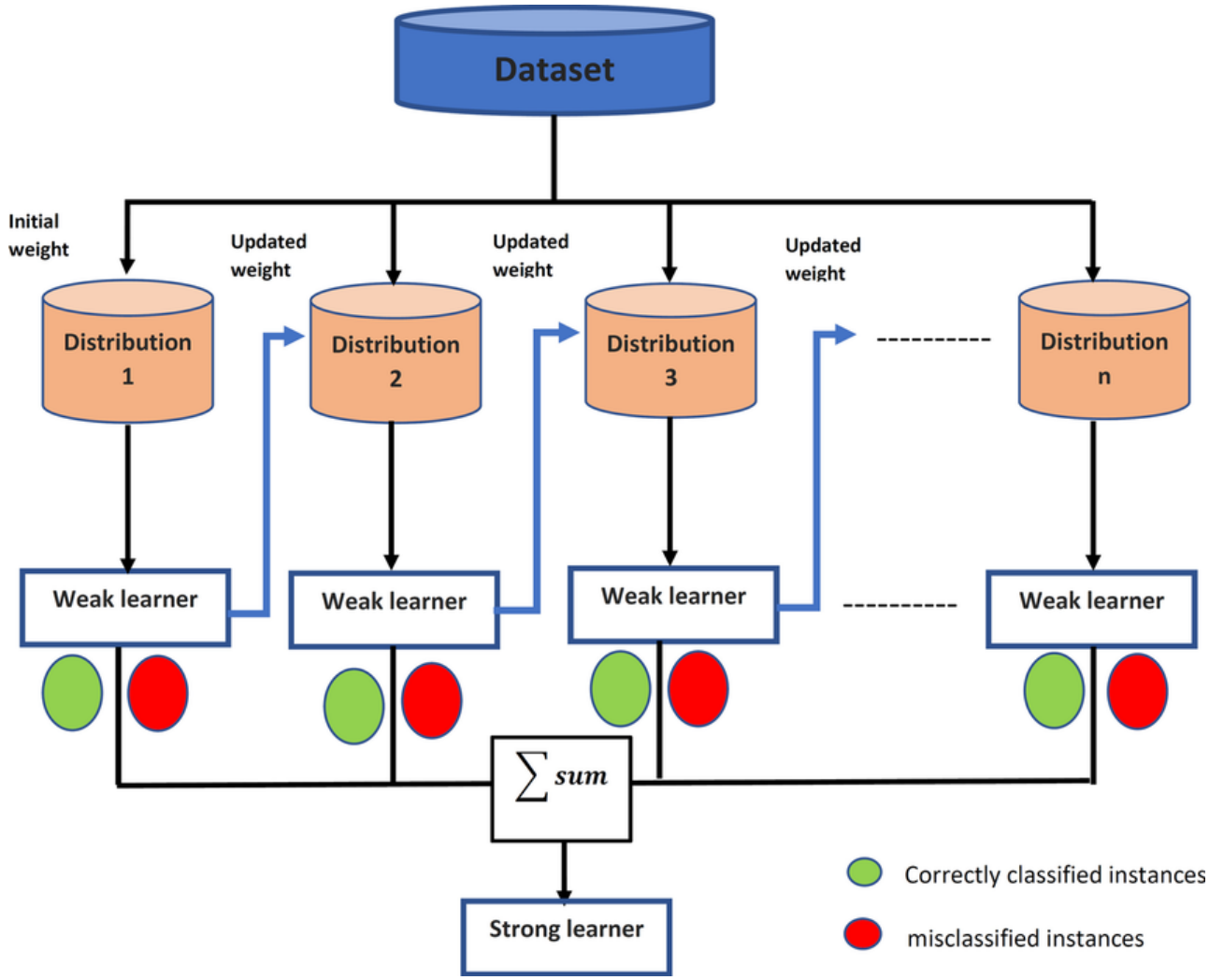


Figure 11 Overview of Gradient Boosting algorithms. Figure taken from [97]

Adaptive boost. Adaptive boost is a boost method that trains successively weak classifiers by giving more weight to samples that were incorrectly classified by previous models.



Naïve bayes classifier. Based on Bayes' theorem and used primarily for text classification and natural language processing, it operates under the assumption that features are independent of each other, which greatly simplifies the calculations.

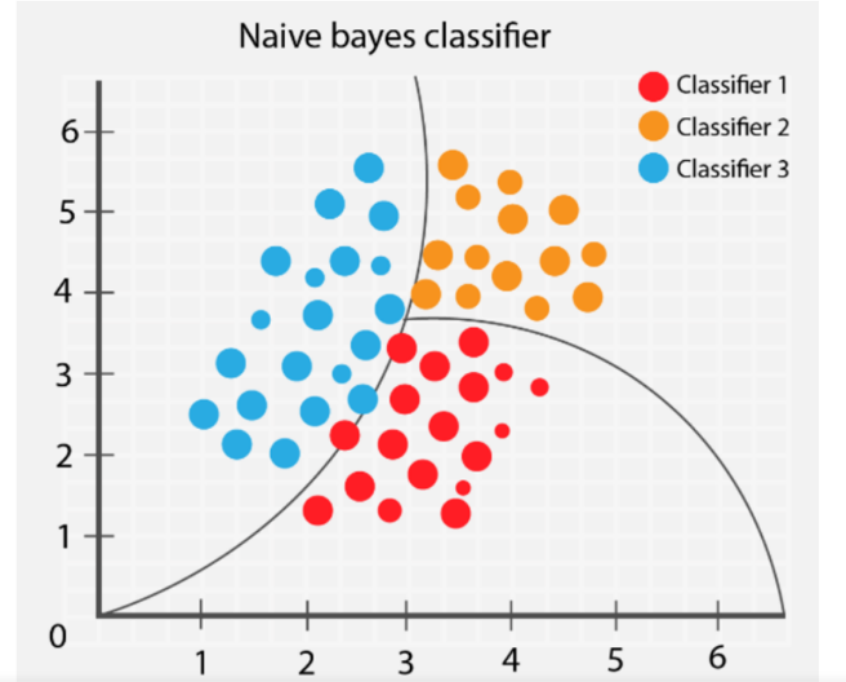


Figure 12 Overview of Naïve Bayes Classifier. Figure taken from [98]

K nearest neighbors (K-NN). K-NN is a non-parametric algorithm that classifies new samples based on their proximity to existing data. It can be quite slow for large sets due to the need to calculate the distances between all points.

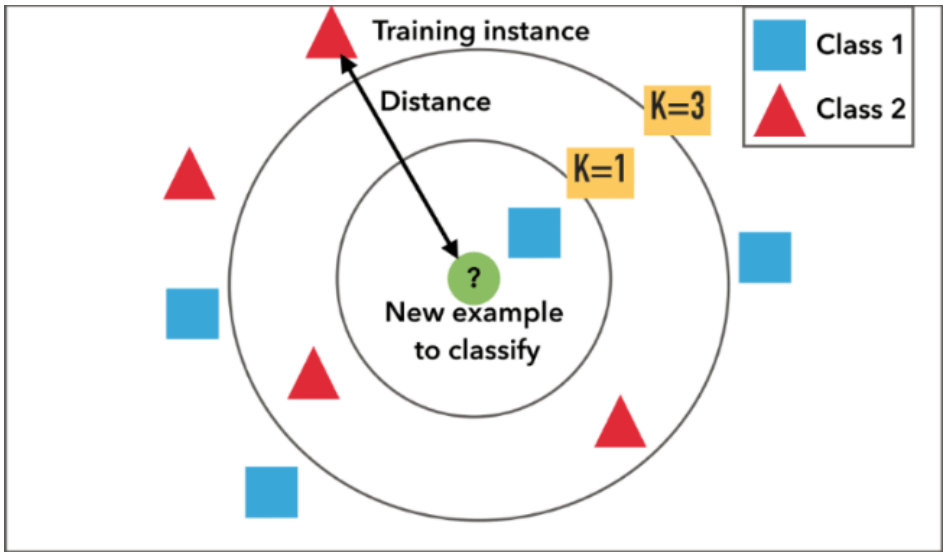


Figure 13 Overview of KNN classifier. Figure taken from [99]

Logistic Regression. The main use of Logistic Regression is binary classification by predicting the probability that a sample belongs to a class using the sigmoid function.

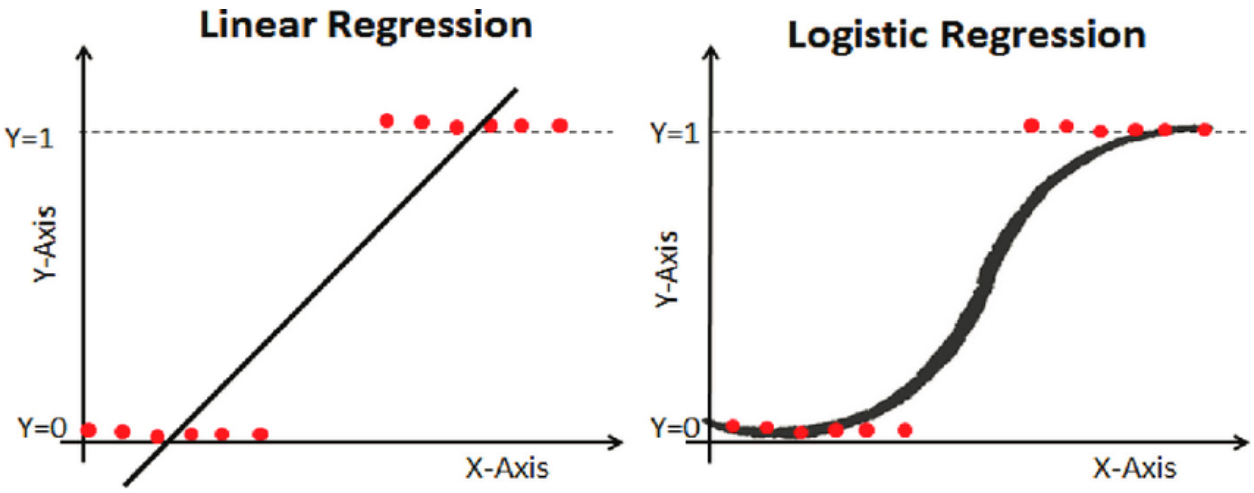


Figure 14 Overview of Logistic Regression Classifier. Figure taken from [100]

3.4.1.2. Unsupervised Learning Algorithms

Regarding unsupervised learning algorithms, the model is trained using unlabeled data. This results in the model discovering hidden patterns and associations in the data instead of learning from predefined outputs. This technique is particularly useful when the process of introducing labels into the data is expensive or impractical. To achieve that process, the main applications are data analysis to find associations, data segmentation and anomaly detection. As mentioned in [101], it is preferable to use large amounts of data. The two main categories are clustering and dimensionality reduction [102], [103].

Clustering is one of the most important techniques as it groups data into clusters, that is, into groups with similar characteristics, without knowing in advance the categories to which they belong. Next, we will analyze some of the most important clustering algorithms [104].

• **K-means clustering**: This algorithm groups data without data into groups based on their similarity. Some of its main advantages are its speed and efficiency for large data sets as it is easy to understand and implement. However, the number of clusters (K)must be determined in advance, and it is sensitive to the initial values of the centroids. Essentially, it calculates the distance between points and the cluster center. The goal of K-Means is to minimize the total distance of all points from the center of the groups.

$$J = \sum_{i=1}^{K} \sum_{x_j \in C_i} \left|\left| x_j - \mu_i \right|\right|^2$$

• **Hierarchical clustering**: in this method a tree-like structure is created, in which the data is gradually notified either by merging or by splitting. The main advantages are that it does

not require the determination of the number of clusters in advance. As well as that a hierarchy is given that can be analyzed at different levels. As for the disadvantages, its increased computational cost in large data sets as well as its sensitivity to noise and anomalies are some of them.

$$C = C_{\text{left}} \cup C_{\text{right}}, \quad \text{where } \min_{\text{split}} \sum_{i} D\left(x_i, C_{\text{split}}\right)$$

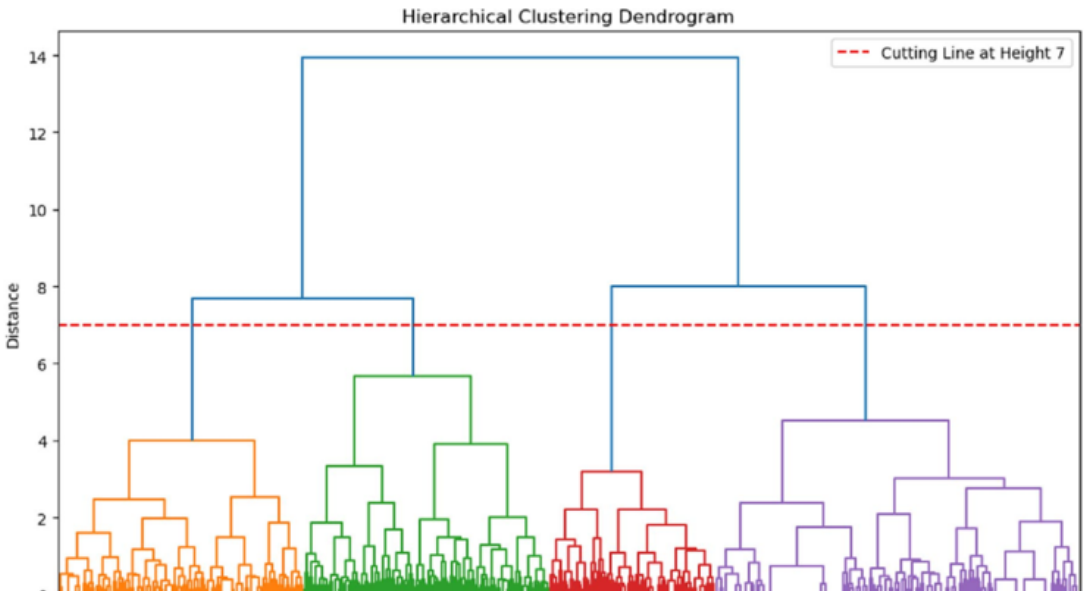


Figure 15. Overview of hierarchical clustering algorithm. Figure taken from [105]

The distance between 2 clusters can be calculated using the following 2 methods:

Nearest neighbor method (single linkage)

$$\hat{y} = \arg\max_{c} \sum_{i \in N_k(x)} \mathbb{1}(y_i = c)$$

Average linkage method (average linkage)

$$D(A, B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a, b)$$

Dimension reduction is a method that reduces the number of features in a dataset, retaining only the important information [106][107].

Principal Component Analysis (PCA)

PCA is one of the most popular algorithms for dimensionality reduction. It was then sent to transform data into a new set of variables that explain the largest possible variation in the data. The main advantages are the reduction of the dimensionality of the data while preserving the most critical information, as well as the simplification of the visualization process. The main disadvantage is the loss of information during dimensionality reduction as well as the difficulty of interpreting the results. to solve the mathematical procedure, we proceed to find eigenvectors and eigenvalues of the covariance matrix [108].

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$$

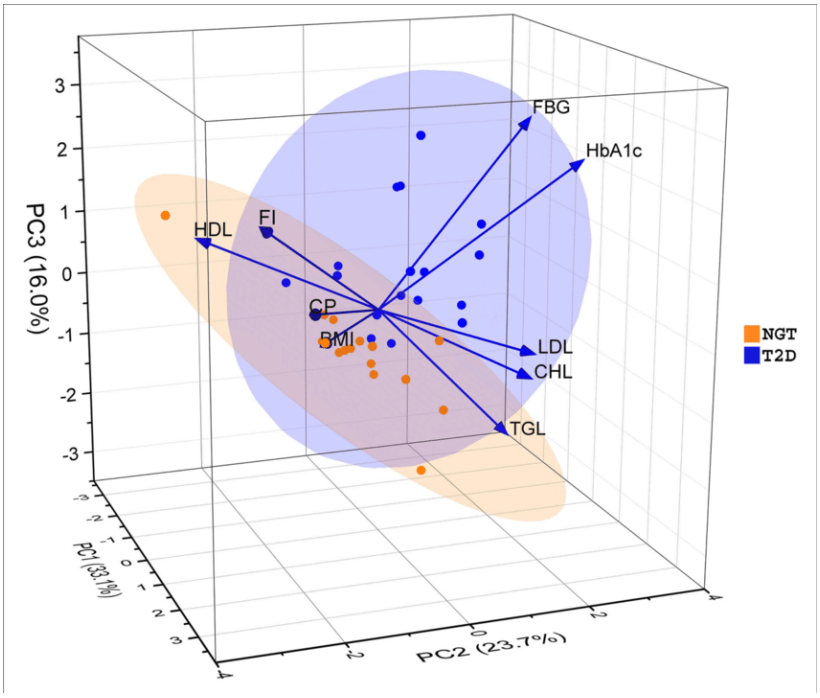


Figure 16 Overview of PCA Component Analysis. Figure taken from [109]

t-SNE (t-Distributed Stochastic Neighbor Embedding

is a nonlinear dimensionality reduction technique primarily intended for high-dimensional data visualization. It allows us to visualize high-dimensional data while preserving local relationships between data. Its main disadvantage is its computational cost and the difficulty in generalizing the results [110][111].

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

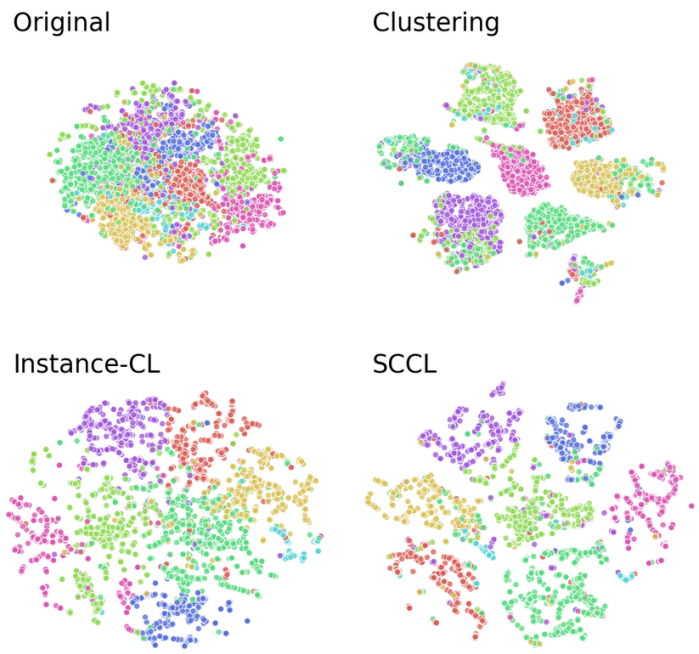


Figure 17 Visualization of the embedding space learned on SearchSnippets using Sentence Transformer. Figure taken from [112]

3.4.2. Model Evaluation Methods

Evaluating a machine learning model is critical to ensuring generalization to new data. The most common cross-validation methods include. A) n-fold cross-validation B) Leave-one-out cross-validation C) Leave-One-Group out validation [113][114]

3.4.2.1. n-Fold Cross Validation

It is a validation technique used to evaluate the performance of a model, maintaining a balance between the number of training data and the number of test data. The procedure is as follows. first, we divide the data set into an equal-sized parts (n). in each iteration the model is trained on n-1 and tested on the remaining one. The process is repeated as many times as the number of data that we have divided. The average of the results is calculated to estimate the overall performance of the model [115].

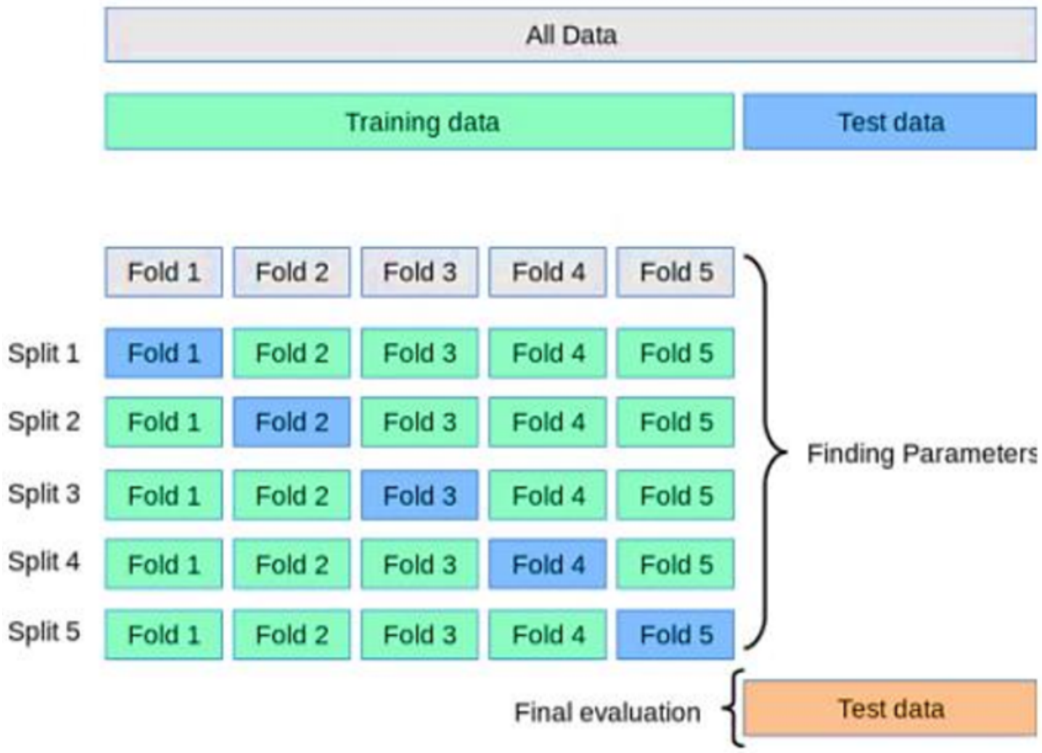


Figure 18 Illustration of k-fold cross-validation. Figure taken from [116]

3.4.2.2. Leave-One-Out Cross Validation (LOOCV)

It is an extreme form of n-Fold Cross Validation, where only one observation is used as input to the test data and all the rest are considered training data, and the process is repeated for all observations. Finally, the average of the individual validation results is used to evaluate the model. Its main disadvantage is the high computational cost, especially for large data sets.[117]

3.4.2.3. Leave-One-Subject-Out Cross Validation (LOSO)

It is a special case of LOOCV, mainly applied to biomedical and physiological data, such as the analysis of physiological signals. Instead of a single observation, each iteration includes an entire participant who is removed from the training and used for testing. The process is repeated for all participants. It is considered extremely useful, especially in cases where the model needs to be generalized to new individuals. Such as emotion detection through physiological signals [118].

3.4.3. Evaluation Metrics

To be able to quantify the performance of a model, we use various metrics. Our main tool is the Confusion Matrix, which allows us to evaluate the performance of a classifier. It is a table

that depicts the actual categories of the data in relation to the predicted categories, allowing us to understand the correct and incorrect predictions of the model [119], [120][114].

Table 5 Confusion matrix representation showing the relationship between actual and predicted classifications

		Predicted class	
		0	**1**
Actual Class	**0**	**True Positive (TP)**	**False Negative (FN)**
	1	**False Negative (FN)**	**True Negative (TN)**

- **True Positive (TP):** The model predicts a positive category, and the actual category is also positive.
- **False Positive (FP):** The model predicts a positive category, but the actual category is negative.
- **True Negative (TN):** The model predicts a negative category, and the actual category is also negative.
- **False Negative (FN):** The model predicts a negative category, but the actual category is positive.

With the help of Table 5, we can calculate important metrics such as accuracy, sensitivity, and specificity.

Accuracy: the percentage of correct predictions is calculated in relation to the total data.

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

- TP is **True Positives**.
- TN is **True Negatives**.
- FP is **False Positives**.
- FN is **False Negatives**.

Precision: expresses the proportion of positive predictions that were actually correct

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

Recall: measures the models ability to correctly detect positive cases

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

F1-Score: is the harmonic mean of precision and recall. F1gives a balanced picture of performance

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

Area Under the Curve (AUC) – ROC

Illustrates the performance of a binary classifier at various decision thresholds. Essentially it measures the overall performance of the model

$$AUC = \int_0^1 \text{TPR}(t)\ d\big(\text{FPR}(t)\big)$$

4. Implementation

4.1. Description of the WESAD dataset and data

WESAD (Wearable Stress and Affect Detection) is a multimodal dataset created to study stress and affect detection using wearable sensors. The WESAD dataset was designed by the authors of [3] and is one of the most important and widely used publicly available datasets for the analysis of physiological signals in the context of emotion recognition and stress detection. The experiment involved 15 participants (12 men and 3 women, with a mean age of 27.5 ± 2.4 years), who underwent a laboratory protocol designed to elicit and record different physiological responses. The protocol included four basic conditions: a neutral condition (Baseline), a stress condition (Stress), an amusement condition (Amusement) and a period of guided meditation (Meditation) as shown in **Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε.**5. The initial neutral condition involved reading neutral content for 20 minutes, while the stress condition was based on the Trier Social Stress Test (TSST), which involved public speaking and numerical calculations under pressure, with a total duration of 10 minutes. The entertainment condition involved viewing 11 humorous videos with a total duration of 392 seconds, designed to induce positive emotions. The meditation phase included a 7-minute guided breathing exercise, which helped to return participants to a neutral emotional state. The total duration of the experiment was approximately two hours, while the order of the conditions was alternated for different participants to avoid biases due to the order of exposure to stimuli.

Data collection was carried out via two different wearable devices: the Empatica E4 bracelet and the RespiBAN chest sensor. The Empatica E4 recorded blood pulse volume (BVP) at 64 Hz, electrodermal activity (EDA) at 4 Hz, skin temperature (TEMP) at 4 Hz, and 3D acceleration (ACC) at 32 Hz. RespiBAN, a medical-grade sensor, recorded electrocardiogram (ECG) at 700 Hz, electrodermal activity (EDA), electromyogram (EMG), respiratory rate (RESP), body temperature (TEMP), and acceleration (ACC). The two devices were manually synchronized with a double tap at the start and end of the recording, while the data was stored locally to avoid losses due to wireless transmission.

To validate the accuracy of the recorded data and confirm that the experiment induced the desired emotional states, five self-reports were collected from participants during the study. These included the PANAS questionnaire to measure positive and negative emotions, the State-Trait Anxiety Inventory (STAI) to assess anxiety levels, the Self-Assessment Manikins (SAM) scale to assess emotional states, and the Short Stress State Questionnaire (SSSQ) to assess types of stress (anxiety, involvement, distress). Statistical analysis of the self-reports confirmed that the stress condition induced a significant increase in arousal and a decrease

in positivity, while entertainment had a smaller effect, enhancing the reliability of the protocol.

In the original work of WESAD dataset, the data processing followed the steps: noise filtering, signal smoothing and segmentation into analysis windows (60 seconds for physiological signals, 5 seconds for accelerometric data). Then, critical statistical and frequency features were extracted, such as heart rate (HR), heart rate variability (HRV), electrodermal activity (EDA) statistics, respiratory rate characteristics (RESP) and electromyogram (EMG), acceleration and temperature statistics as shown in

Table 15. In the next stage, machine learning models, including Decision Trees (DT), Random Forest (RF), AdaBoost (AB), Linear Discriminant Analysis (LDA) and k-Nearest Neighbors (kNN), were developed and evaluated to classify physiological signals in different conditions.Data analysis showed that the classifiers achieved up to 80% accuracy in the three-class problem (Baseline vs. Stress vs. Amusement) and up to 93% accuracy in the Stress vs. Non-Stress classification. Respiration (RESP) and electrocardiogram (ECG) were found to be the most important physiological variables for stress detection, while electrodermal activity (EDA) had greater discrimination when recorded by chest sensors. The results of the study indicate that the use of chest sensors provides higher classification accuracy, however, recordings from wearable wrist devices, such as the Empatica E4, remain promising due to their non-intrusive nature.

Overall, the WESAD dataset is a reliable and publicly available dataset for research into emotion recognition and stress detection via mobile devices. It provides a solid foundation for the development and evaluation of machine learning models and is expected to contribute significantly to the understanding of physiological responses to stress, enabling the development of practical applications in the field of health and mental well-being.

Labels and categorization

The dataset includes 4 different states that indicate different emotions in the participants.

Table 6 Label definitions for different physiological states in the dataset

Label	Situation
1	Baseline
2	Stress
3	Amusement
4	Meditation

Table 7 is a summary table listing the signals included in the WESAD dataset.

Table 7 Sensor specifications for physiological signal acquisition, including device type and sampling frequency.

Signal	Device	Frequency
ECG	Respiban	700 hz
RESP	Respiban	700 hz
EDA	Respiban	700 hz
ACC	Respiban	64 hz
TEMP	Respiban	700 hz
BVP	Empatica	64 hz
EDA	Empatica	4 hz
ACC	Empatica	32 hz
TEMP	Empatica	4 hz

The high sampling rates shown in Table 7 allow us to accurately record and analyze physiological signals.

4.1.1 Description of the tools and software used

For the analysis of the data, various tools and software were used for the processing and visualization of the physiological signals. Regarding their visualization and analysis, it was mainly carried out in MATLAB, meaning the processing of the data, the extraction of features, while for the classification python was used. More specifically, libraries such as NumPy for the management of numerical data, SciPy for the filtering for the statistical analysis of the signals as well as scikit_learn for the classification of the data and the training of machine learning models were used. The main environment where the algorithms were implemented was Jupiter Notebook. Regarding the toolboxes and the function of MATLAB, we used signal processing toolbox and statistics and machine learning toolbox.

4.2. Data processing process

Initially, since the data coming from the Empatica device does not contain labels, we had to synchronize the signals coming from respinan and Empatica. To analyze the temporal evolution of the labels, a time vector is created based on the data sampling rate (700 Hz). Then, the times when the participant's label changes are identified, and the intervals during which each label remains constant are calculated. This information is stored in a table that contains for each label the start time, the end time, and the corresponding label category .The resulting temporal information is stored in a .txt file, which includes three columns: the start time, the end time, and the corresponding label for this interval as shown in Figure 19.

Start Time	End Time	Label
0.000000	306.545714	0
306.547143	1450.545714	1
1450.547143	2273.544286	0
2273.545714	2888.544286	2
2888.545714	3160.545714	0
3160.547143	3225.545714	6
3225.547143	4097.545714	0
4097.547143	4488.544286	4
4488.545714	4763.545714	0
4763.547143	5125.545714	3
5125.547143	5269.545714	0
5269.547143	5333.545714	7
5333.547143	5496.545714	0
5496.547143	5873.545714	4
5873.547143	6078.998571	0

Figure 19 Time intervals table representing the times that the label is changing.

At the beginning of the process, the participants performed a double tap on the chest, which also signaled the beginning of the experiment. In Figure 20, we observe an anomaly at the beginning of the signal, which is due to the participant hitting the device.

Finally, the data is visualized. The EDA signal is plotted as a function of time in one diagram, while the labels are displayed separately in a second diagram. The times when the labels change is marked with red markers so that the transitions between the different states of the participant are visible. Figure 21

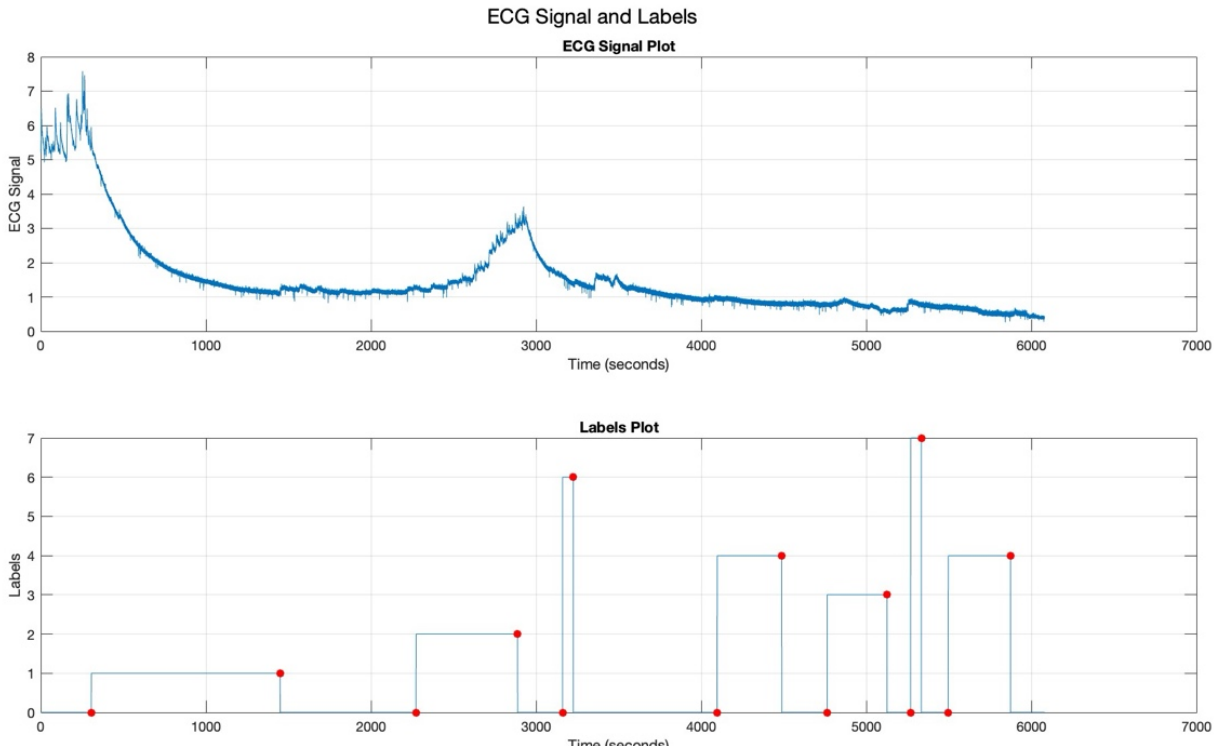


Figure 20 ECG signal over time (top) and corresponding labels (bottom)

At this stage of implementation, the physiological signals were synchronized with the time intervals of the corresponding labels, with the aim of better understanding the participant's states during the experiment. The signals used come from the sensor on the wrist and on the chest, and concern the following recordings: BVP, ECG, and EDA.

For each type of signal, the corresponding time vector was calculated, based on the sampling rate: 64 Hz for BVP, 700 Hz for ECG and 4 Hz for EDA. Then, the time intervals were loaded from the files containing the start and end information of each recorded state as well as the label that represents it

Then, the signals were plotted on graphs, to which red rectangles were added to mark the time intervals of each label, and numerical indications indicating the value of the label (e.g., 1 for baseline, 2 for stress, etc.).

This approach allowed for the visual connection between physiological changes and alternating psychological states, which significantly enhances the understanding of the

structure of the dataset and can contribute to targeted analyses or the development of classification algorithms. The results are presented in Figure 21 Figure 21 and Figure 22

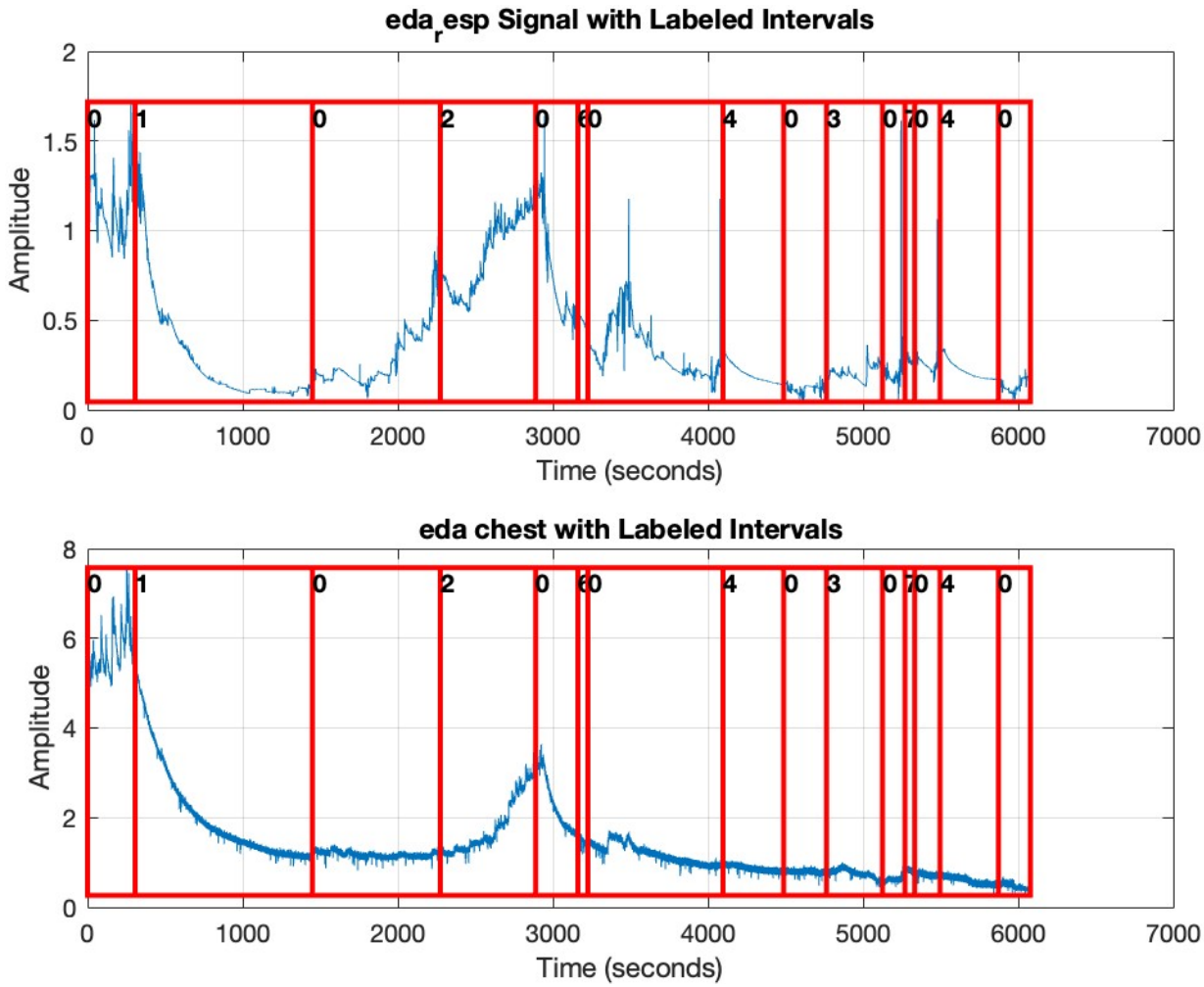


Figure 21 Electrodermal activity (EDA) signals from wrist (top) and chest (bottom) with labeled intervals

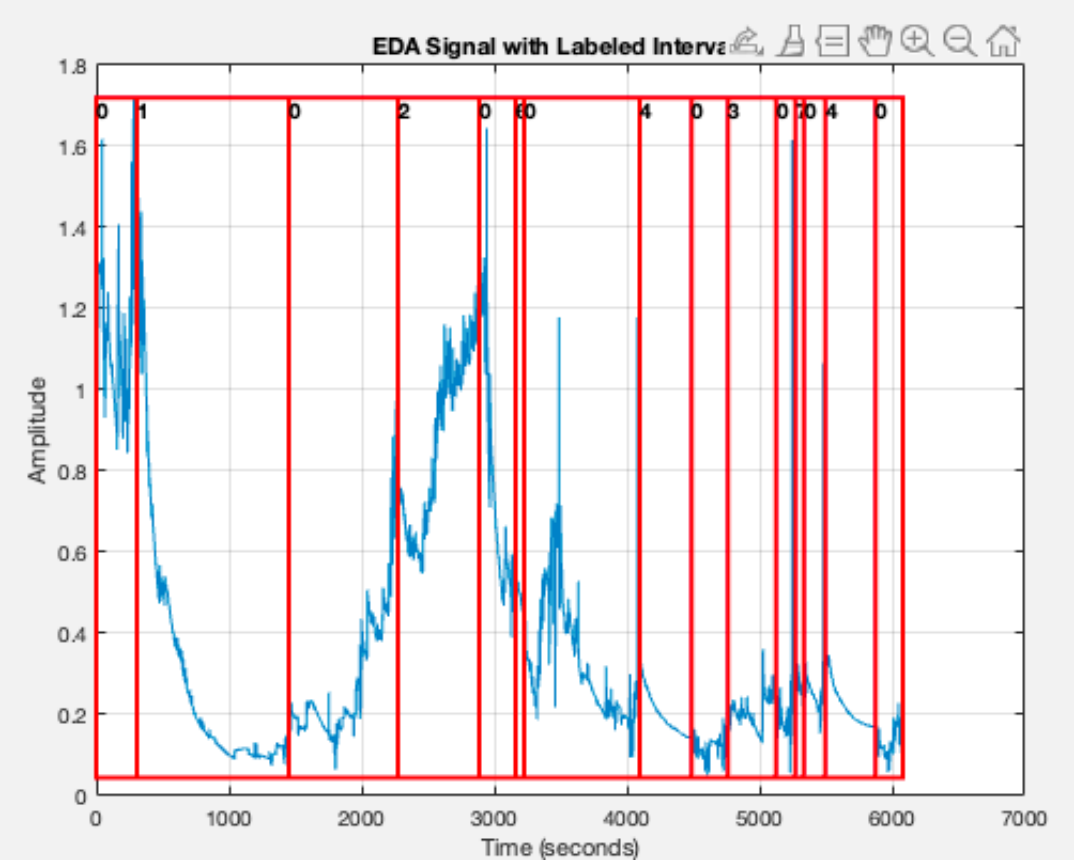


Figure 22 Electrodermal activity (EDA) signals from the wrist with labeled time intervals.

We then focused on the parallel visualization of the physiological signals recorded by the sensor on the participant's wrist, within the context of the WESAD experiment. The simultaneous visualization of these signals allows for the comparative observation and analysis of the participant's physiological reactions, providing a valuable tool for understanding their behavior in conditions of anxiety or other psychological states.

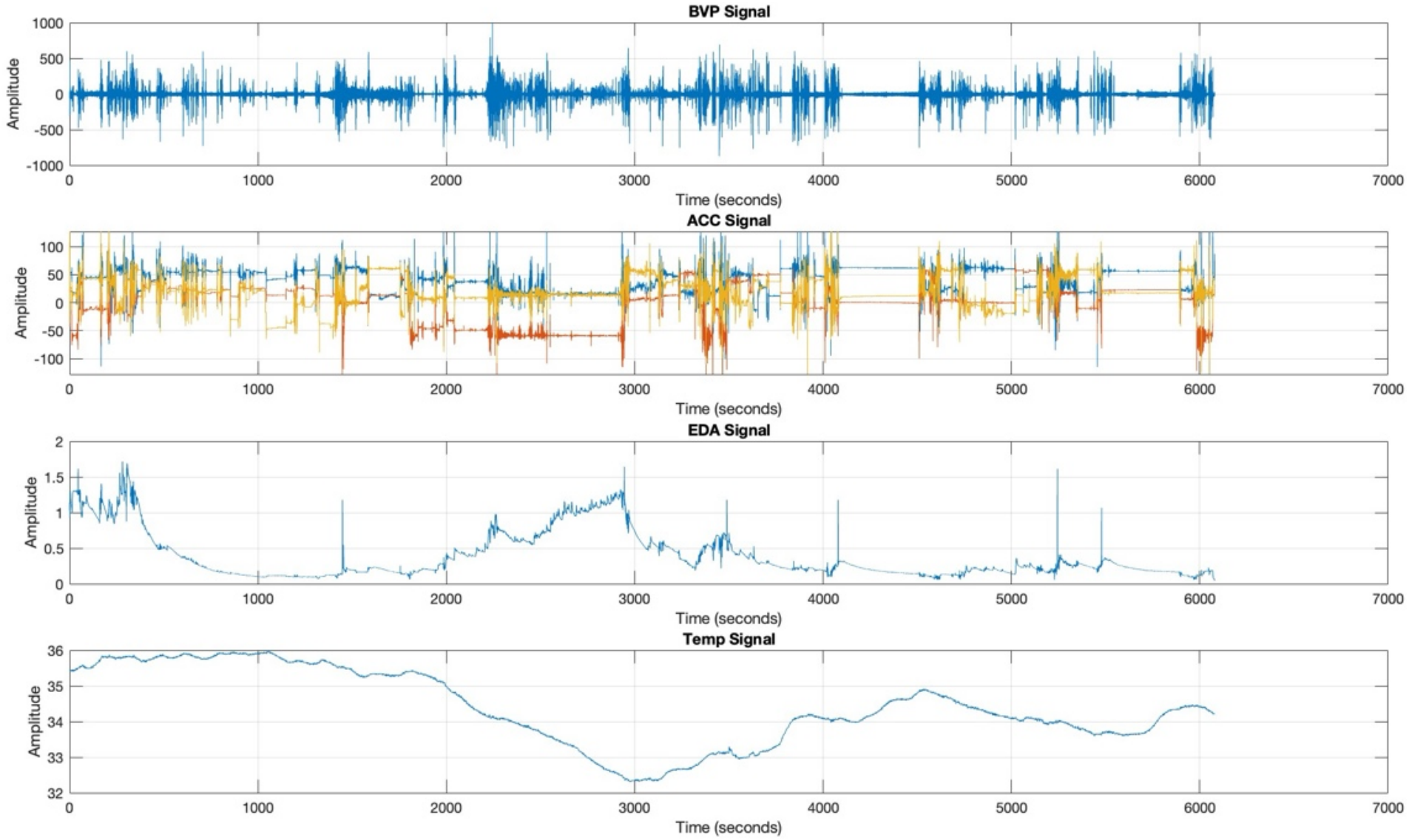


Figure 23 Time-series plots of physiological signals, including Blood Volume Pulse (BVP), Accelerometer (ACC), Electrodermal Activity (EDA), and Temperature (TEMP), illustrating variations over time.

Following the implementation, processing of the physiological signals was performed for all participants in the WESAD dataset, to temporally align the signals with the corresponding psychological states, as identified by the labels. The process was implemented through an automated loop, which sequentially processes the participants' data. For each participant, the recorded data is loaded, and four basic signals are extracted from the sensor on the wrist: BVP (pulsatile blood flow), EDA (electrodermal activity), TEMP (temperature), and ACC (three-dimensional acceleration). For each of the signals mentioned before, the corresponding time vector is calculated, based on the respective sampling rate, and then they are matched with the time-periods described in the label files of each participant (e.g. s2.txt). The signal points falling within each time interval are labeled with the corresponding state category, and the result is stored in .txt format files, organized by signal type and participant (e.g. BVP/BVP_S2.txt, EDA/EDA_S3.txt). This process provides fully synchronized and labeled data, which is suitable for future analysis, visualization, or use in machine learning algorithms.

The dataset creators specify that only the labels assigned numerical values are to be considered. That is ,the following labels should be considered: 1: Baseline, 2: Stress, 3: Amusement, 4: Meditation. For the rest, there is no reference to what exactly they identify. So, we were obliged in the files that we have extracted to keep only the labels with the numbers 1,2,3,4, for us to be able to create the machine learning modelSo, we reprocessed the original data and divided the data of each signal into 4 different categories based on the label.

For this purpose, the physiological signals were separated by label, to create separate files for each category of emotional state. The process was applied to BVP and EDA signals, which had previously been synchronized in time based on the corresponding time intervals of the labels. Initially, the data file of each participant is read and for each line of the file the value of the label is identified. They are classified into four categories, depending on the value of the label (1: baseline, 2: stress, 3: amusement, 4: meditation), stored in separate text files (e.g. This separation facilitates the independent processing and analysis of each data category, while at the same time determining the structure of the data more flexible and compatible with methodologies

Diagrams for EDA with the different labels are presented in Figure 24 that shows the baseline state, in Figure 25 that shows the stress state, in Figure 26 that shows the amusement state, and in Figure *27* that shows the meditation state.

Figure 24 depicts the EDA signal recorded during the baseline condition, showcasing the stable physiological state of the subject in the absence of stressors.

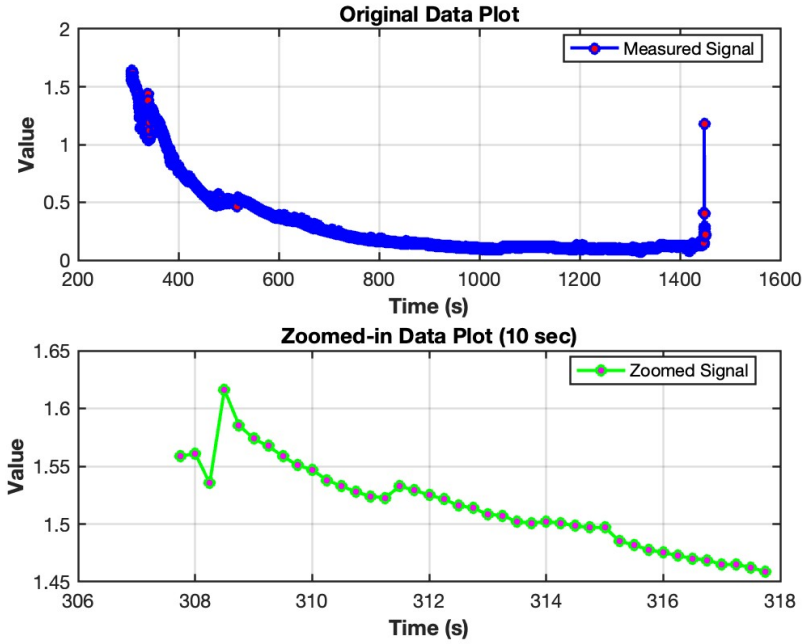


Figure 24 Eda baseline condition original and zoomed in

Figure 25 illustrates the EDA signal during the stress condition. The signal exhibits pronounced fluctuations and increased amplitude, which are characteristic of sympathetic nervous system activation.

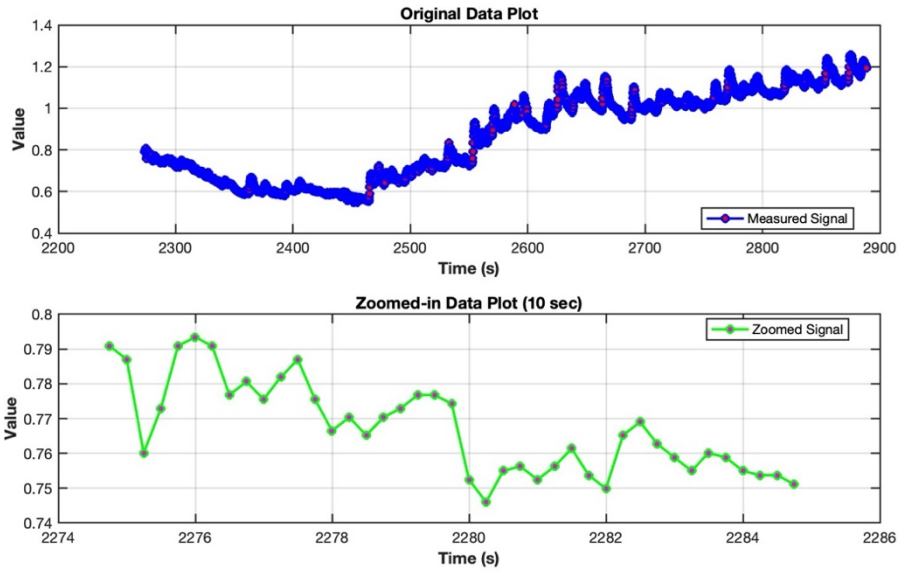


Figure 25 Eda stress condition and zoomed in.

In Figure 26, the EDA response corresponding to the amusement condition is presented. Compared to the baseline and stress states, the signal demonstrates moderate variation and a unique pattern of activation.

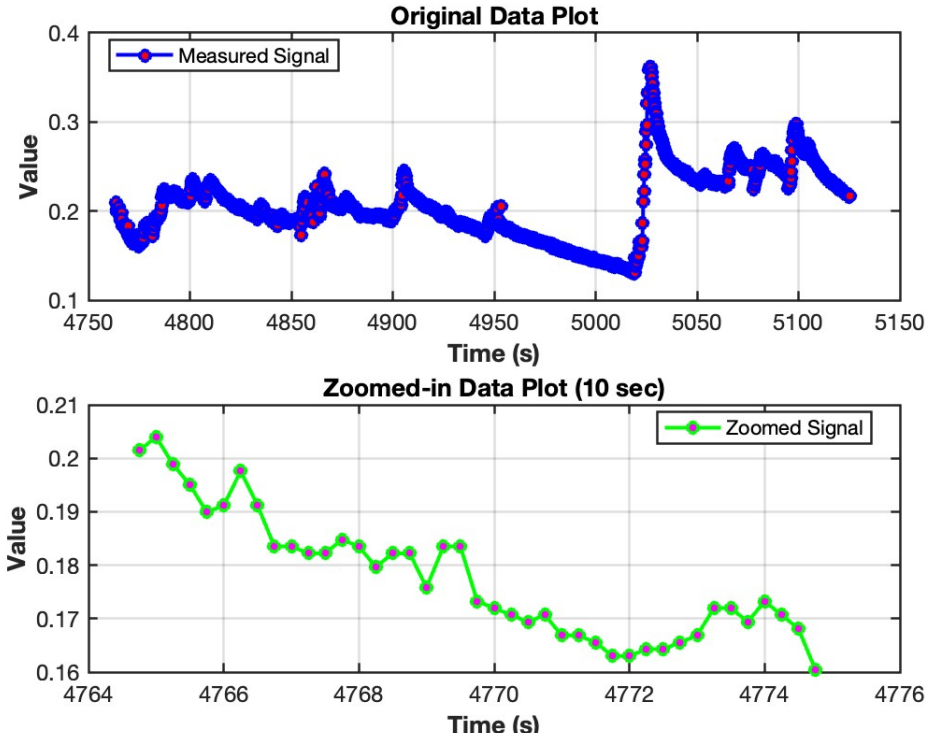


Figure 26 Eda amusement condition original and zoomed in

Finally, Figure 27 presents the EDA signal during the meditation condition. The signal is characterized by reduced activity and smoother transitions, indicative of a relaxed physiological state.

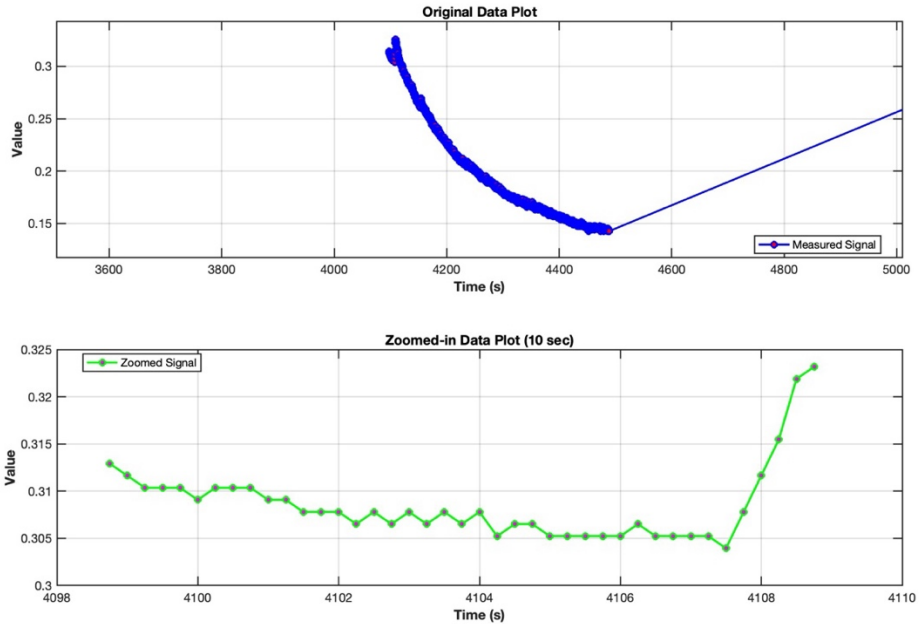


Figure 27 Eda meditation condition original and zoomed in

The Blood Volume Pulse (BVP) signal under different emotional states is illustrated in Figure 28 to Figure 31. Each figure displays both the original and a zoomed-in version of the signal, allowing for a more detailed observation of the signal morphology and its variation across emotional conditions.

Figure 28 presents the BVP signal during the baseline state. The waveform demonstrates stable cardiac activity, serving as a reference for comparison with the other conditions.

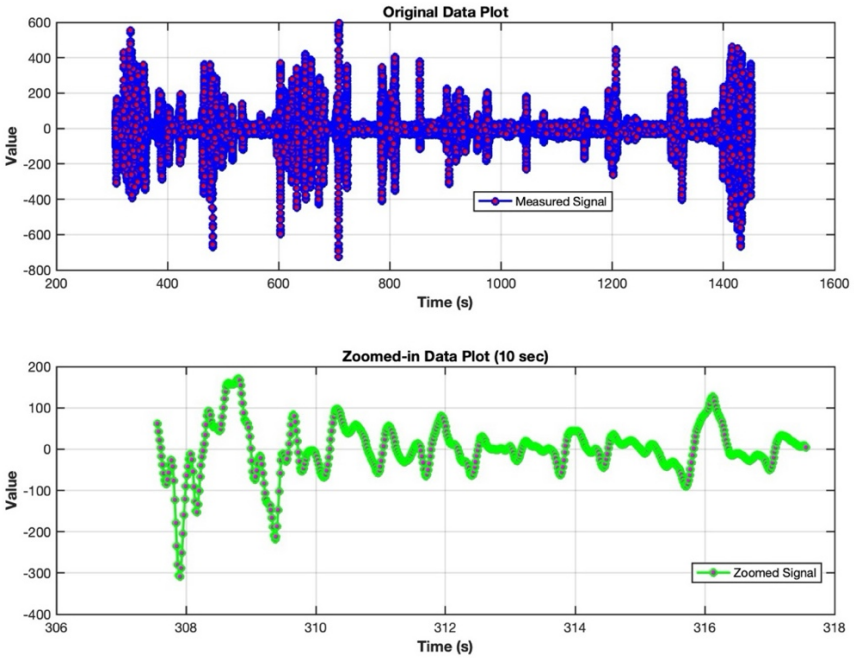


Figure 28 BVP baseline condition original and zoomed in

In Figure 29 the BVP signal under the **stress** condition, is presented. The waveform reveals increased irregularities and fluctuations, indicating heightened physiological arousal associated with stress responses.

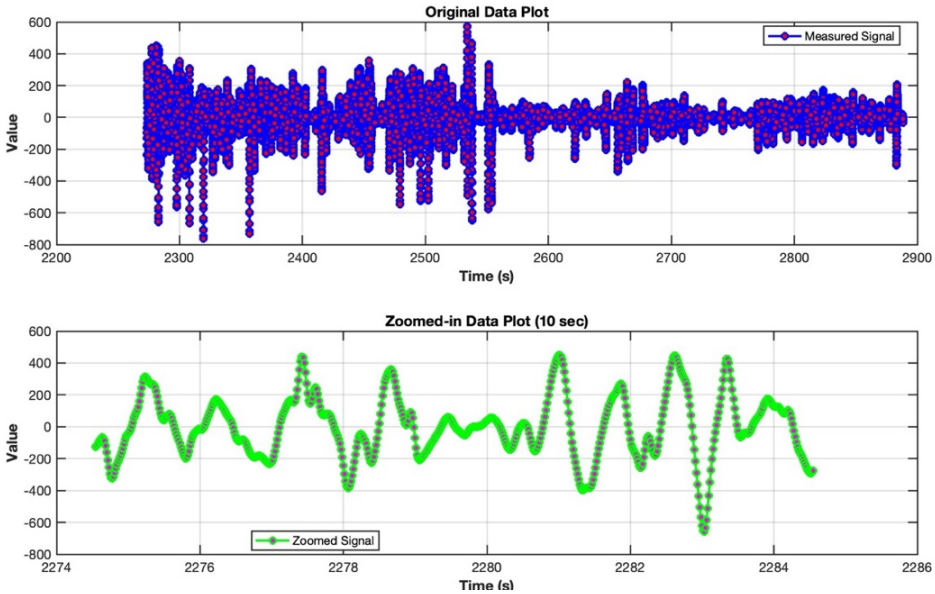


Figure 29 BVP stress condition original and zoomed in

In Figure 30, the BVP signal during the meditation state is illustrated. The signal appears smoother and more regular, reflecting a relaxed autonomic state consistent with meditative practices.

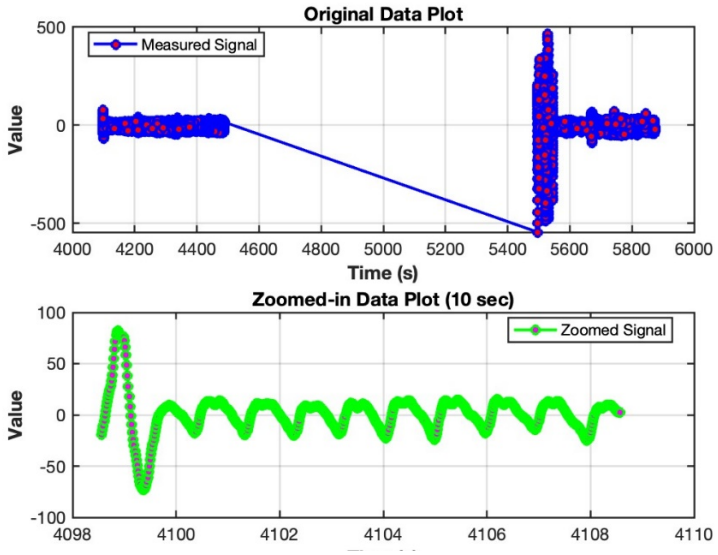


Figure 30 BVP meditation condition original and zoomed in

Lastly, Figure 31 displays the BVP signal in the context of **amusement**. The signal pattern shows moderate variability and rhythmic structure, which may reflect emotional engagement without the intensity of stress.

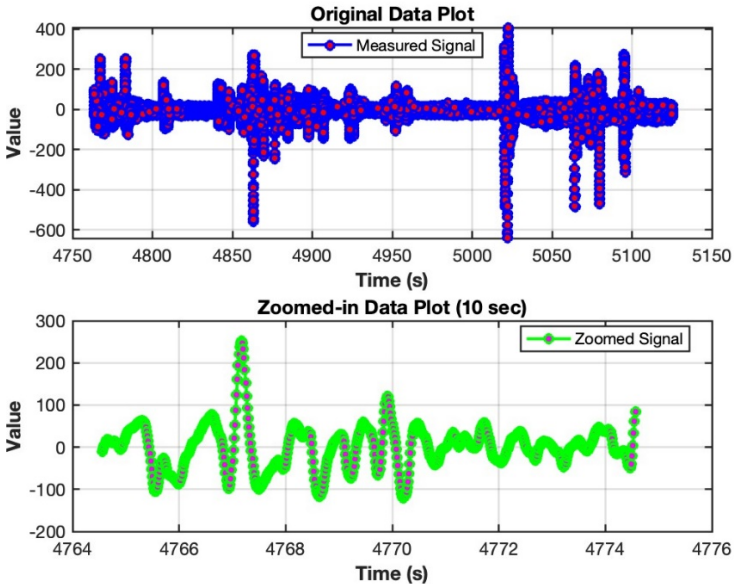


Figure 31 BVP amusement condition original and zoomed in

4.2.1. Signal filtering and normalization

Preprocessing of physiological signals is a critical step for extracting reliable features. The process includes filtering, noise removal, and normalization to ensure similarity between samples. Depending on the type and nature of the signal, different filters are used.

The EDA includes two main components. The tonic component represents the baseline level of skin conductance and reflects slow changes in the signal which are primarily influenced by factors such as hydration, skin properties, and general autonomic arousal. The Phatic component captures rapid, transient changes in skin conductance that occur in response to specific stimuli, such as emotional or cognitive events. These fast fluctuations, known as skin conductance responses (SCRs), are often linked to heightened sympathetic nervous system activity which identifies fast responses due to emotional reactions. To preprocess the EDA signal, we applied filtering techniques to remove noise and effectively separate the tonic and phasic components. First, a low-pass Butterworth filter with a cutoff frequency of 1.5 Hz was used to smooth the signal and eliminate high-frequency noise caused by movement artifacts or environmental factors. Next, a high-pass filter with a cutoff frequency of 0.05 Hz was applied to separate the slowly varying tonic component from the rapidly fluctuating phasic component. The cutoff frequencies were selected based on previous studies to ensure optimal decomposition of the EDA signal.

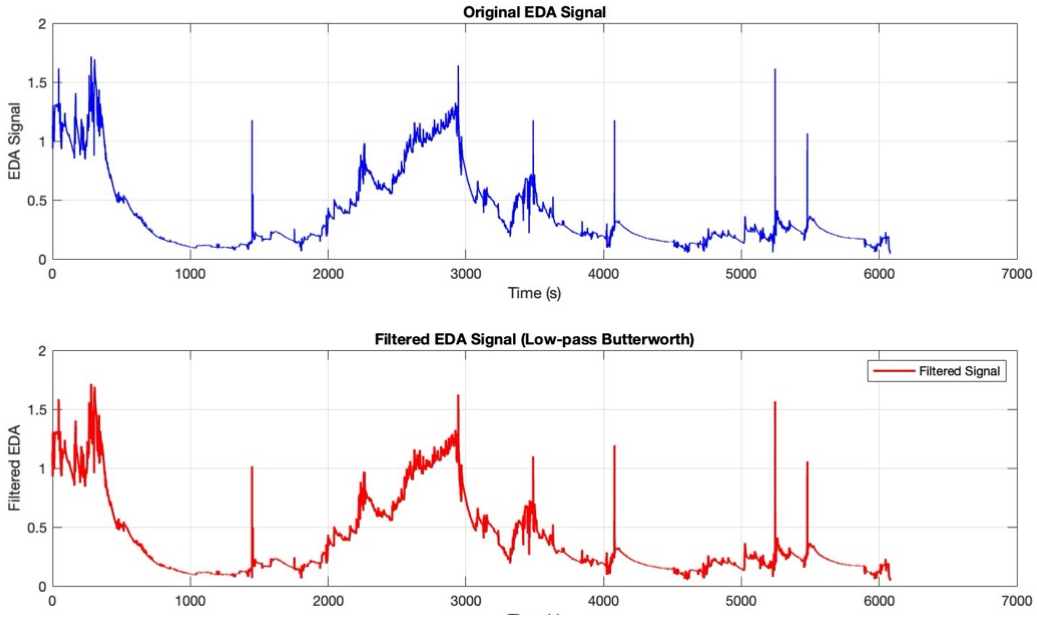


Figure 32 Comparison of the original EDA signal and the filtered EDA signal using a low-pass Butterworth filter (bottom)

The signal was then separated into its two main components: the phasic component, which reflects short-term responses of the autonomic nervous system, and the tonic

component, which expresses the general level of arousal. The separation was achieved using a moving average, considering the phasic component as the filtered signal and the tonic component as the remainder between the initial and filtered signal. The simultaneous display of these three signals (initial, tonic, and phasic) allows for an in-depth interpretation of the participant's physiological response at each time point as shown in Figure 33.

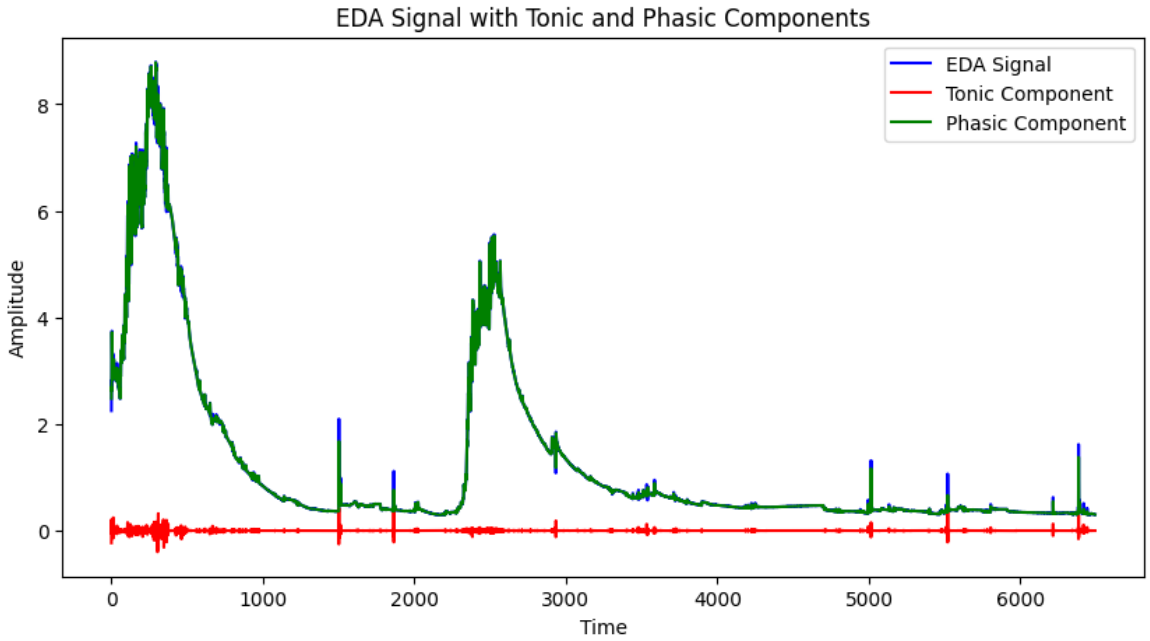


Figure 33 Decomposition of the Electrodermal Activity (EDA) signal into tonic and phasic components, highlighting slow baseline changes (tonic) and rapid responses (phasic).

Then, for the full EDA signal of a participant, a peak detection algorithm using a threshold was applied to identify moments of intense electrodermal activity. Figure 34

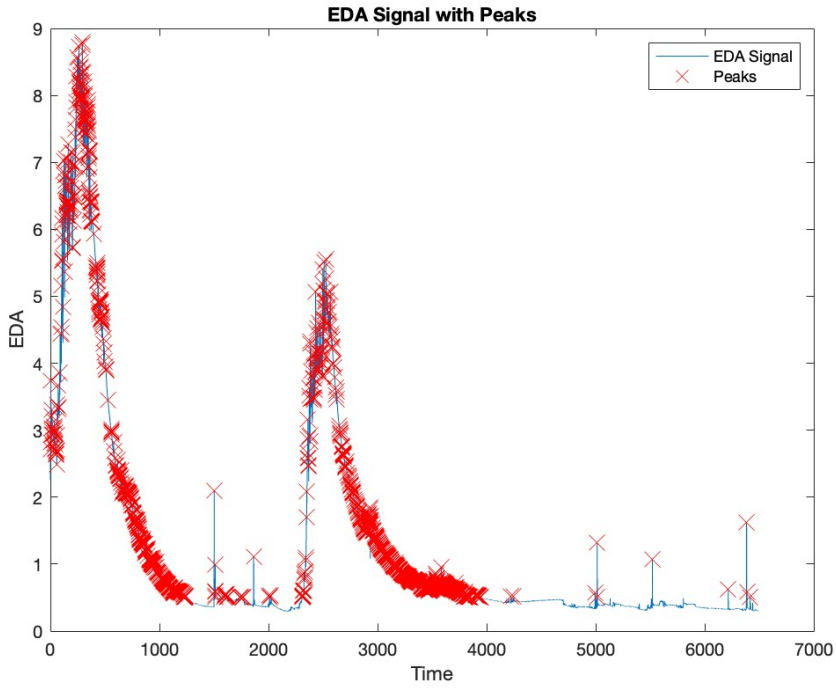


Figure 34 Electrodermal Activity (EDA) signal with detected peaks

Blood Volume Pulse (BVP) is a physiological signal derived from photoplethysmography (PPG) and is commonly used to assess heart rate (HR) and heart rate variability (HRV). To preprocess the BVP signal, we applied a band-pass filter with a frequency range of 0.5–5 Hz to isolate heart rate dynamics while removing low-frequency drift and high-frequency noise. This range was chosen because it effectively captures the dominant frequencies associated with cardiac activity, as heart rate typically falls within this range.

To further improve the signal and reduce noise, a moving average filter was applied to smooth out short-term fluctuations. This step helps improve peak detection by reducing small variations that may occur due to motion artifacts or environmental factors the results are presenting in Figure 35.

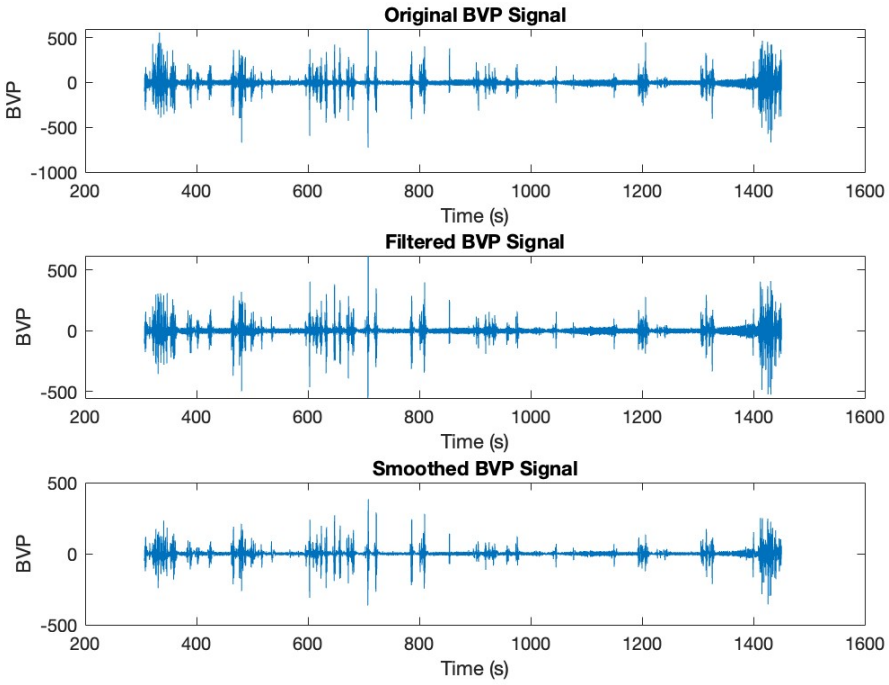


Figure 35 Processing of the Blood Volume Pulse (BVP) signal, showing the original signal), filtered signal, and smoothed signal to enhance signal quality.

After preprocessing, we applied various feature extraction algorithms to identify and visualize RR intervals - the time differences between consecutive R peaks as shown in Figure 36. Accurate detection of these peaks is essential for extracting heart rate variability measures, which provide information about autonomic nervous system activity. The extracted RR intervals were then used for further analysis, such as computing time-domain and frequency-domain HRV features.

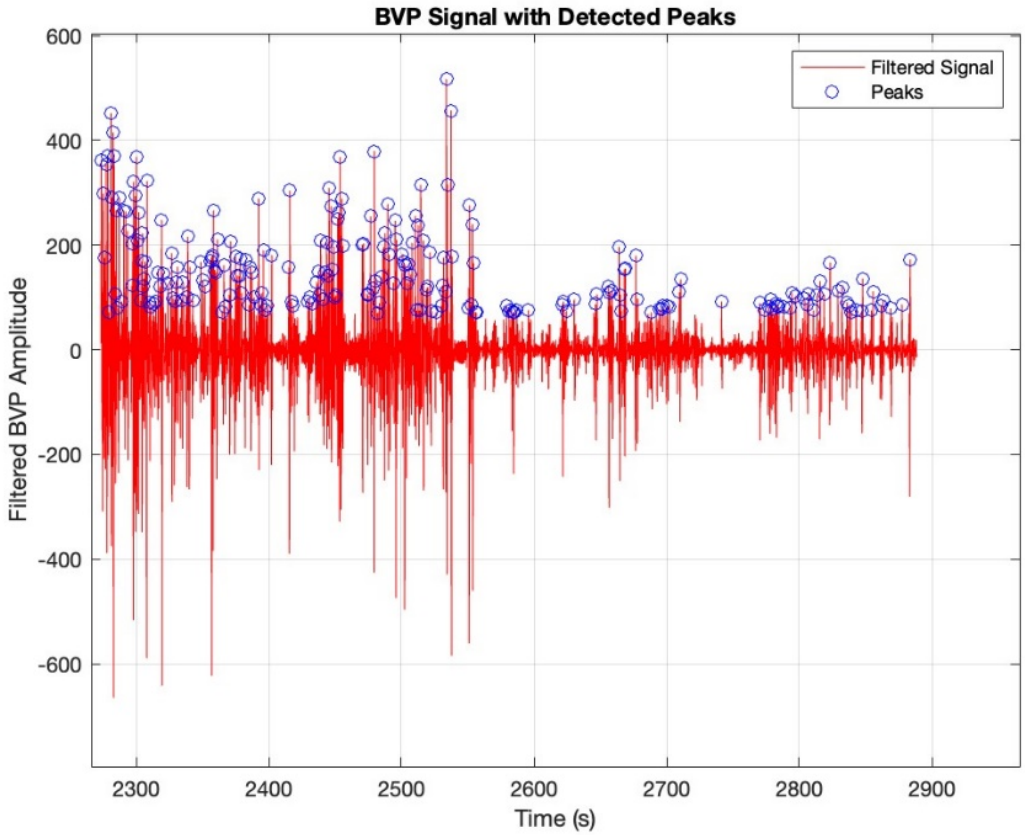


Figure 36 Filtered Blood Volume Pulse (BVP) signal with detected peaks

4.2.2. Feature extraction and selection from EDA and BVP

Segmentation and window size selection are among the most critical steps in the analysis of physiological signals such as EDA and BVP. This technique allows the extraction of features from fixed time intervals, improving the performance of classification and pattern recognition.

For a more targeted and detailed analysis of the EDA signal, a feature extraction technique using time windows was applied to identify local changes and dynamic responses at different time points. The signal was divided into 60-second windows, which were examined under three different settings: no overlap, 50% overlap, and 25% overlap. Each window was subjected to a low-pass Butterworth filter to remove noise and smooth the signal. Subsequently, numerical features were calculated in the time and spectral domains, such as mean, standard deviation, skewness, kurtosis, number and height of peaks, rise times, total area under the curve, average rate of change (derivative), spectral power, dominant frequency, spectral entropy, and spectral center of gravity. The following tables provide an analytical description of the features that we have extracted (Table 8, Table 9, and Table 10).

The use of different windowing settings allowed the evaluation of whether temporal granularity affects the quality and distinctiveness of the extracted features, thus contributing to the creation of a rich and time-sensitive dataset suitable for machine learning applications.

Table 8 Time Domain signal features

Feature	Description	Typical Range
Mean	Average amplitude of the EDA signal over the window. Reflects general arousal level.	0.1 – 10 µS
Standard Deviation (Std)	Variability of the signal. Higher in states of elevated emotional arousal.	0.01 – 1.5 µS
Skewness	Asymmetry of the signal distribution. Positive skew often appears in EDA.	> 0 (usually 0.5–3.0)
Kurtosis	Peaked Ness of the signal. High kurtosis implies sharp, infrequent peaks.	> 3 (often 4–10)
SMA	Mean absolute amplitude. Captures general skin conductance tone.	Similar to mean (0.1–10 µS)

Table 9 Time domain signal Dynamics

Feature	Description	Typical Range (Literature)
Peaks Count	Number of peaks detected in the window. Related to number of phasic responses.	0 – 20 peaks/min
Mean Peak Amplitude	Average amplitude of detected peaks. Reflects intensity of responses.	0.01 – 1.0 µS

Mean Rise Time	Average time between successive peaks. Measure's reaction dynamics.	1 – 10 s
Sum Peak Amplitude	Total energy of phasic responses in the window.	-
Sum Rise Time	Cumulative sum of time between peaks.	-
Sum Areas	Area under the curve. Reflects total EDA activity over the window.	-
Mean Derivative	Average rate of change. Indicates signal slope direction.	Close to 0
Mean Negative Derivative	Average negative slope. Relates to rate of decrease in EDA.	≤ 0

Table 10 Frequency domain Features

Feature	Description	Typical Range
Total Power	Total power of the signal's spectrum (via Welch's PSD). Indicates global energy.	Depends on scale
Spectral Power	Sum of all spectral components.	-
Dominant Frequency	Frequency with the highest power. May reflect typical SCR frequency.	0.045 – 0.25 Hz
Spectral Entropy	Complexity/disorder of frequency content. Higher in stress or random signals.	0 – 1 (normalized)
Spectral Centroid	Weighted average frequency. Indicates energy distribution in spectrum.	0.05 – 0.3 Hz
Min Spectral Power	Minimum power among frequency bins.	-
Max Spectral Power	Maximum power among frequency bins.	-

Signal Energy	Sum of squared signal values. Measures total energy in the signal.	Depends on signal scale

To extract features from the BVP signal, a procedure similar to that used for EDA was applied but adapted to the specificities of the cardiovascular signal. First, the signal was filtered with a Butterworth low-pass filter to remove high-frequency noise and preserve the heart rate signal. The signal was then divided into 60-second time windows — both without overlap and with 25% and 50% overlap — to allow detailed analysis of the signal at a local level. For each window, time and frequency domain features were calculated, such as mean, standard deviation, skewness, kurtosis, number and size of peaks, interpulse times, area of the curve, derivatives, and activity level. In the frequency domain, features such as total spectral power, dominant frequency, spectral entropy, and centroid of the spectrum were extracted. Tables with a more comprehensive picture of the features we extracted follow. The use of windows allows for localized observation of heart rate variability, while the extracted features make the BVP signal suitable for analysis or classification in conditions of different emotional states.

Table 11 Time Domain Statistical Features extracted for BVP signal

Feature	Description	Typical Range
Mean	Average value of the BVP signal over the window.	Varies (depends on preprocessing)
Standard Deviation	Measures variability in the BVP signal amplitude.	Varies
Skewness	Asymmetry of the BVP signal distribution.	Usually, > 0
Kurtosis	Peaked Ness of the BVP signal. High values indicate sharp pulse peaks.	Often > 3

Table 12 Time Domain Statistical Features extracted for BVP signal.

Feature	Description	Typical Range
Peaks Count	**Number of pulse peaks detected in the window.**	**60–100 per minute (resting)**
Mean Peak Amplitude	Average height of detected pulse peaks.	Varies
Mean Rise Time	Average time between consecutive peaks (related to heart rate).	0.6 – 1.2 s
Sum Peak Amplitude	Sum of amplitudes of all detected peaks.	-
Sum Rise Time	Sum of rise times between all peaks.	-
Sum Areas	Total area under the BVP signal curve.	-
Mean Derivative	Average rate of signal change.	Close to 0
Mean Negative Derivative	Mean rate of decrease. Related to pulse downstroke.	≤ 0

Table 13 Frequency - Domain Features extracted for BVP signal.

Feature	Description	Typical Range
Total Power	Total spectral power computed via Welch's method.	Varies
Spectral Power	Sum of all power spectral components.	Varies
Dominant Frequency	Frequency with highest spectral power (related to heart rate).	1 – 1.7 Hz (60–100 bpm)
Spectral Entropy	Degree of disorder in frequency content.	0 – 1 (normalized)
Spectral Centroid	Spectral energy center. Weighted average frequency.	Near dominant frequency
Min Spectral Power	Minimum power across frequencies.	-
Max Spectral Power	Maximum power across frequencies.	-
Signal Energy	Sum of squared amplitudes. Indicates signal strength.	Depends on scale

As part of the preprocessing and analysis of the EDA data of the WESAD dataset, a feature selection process was implemented with the aim of improving the performance of future classification models. Initially, the data extracted from the participants' EDA signals were organized in a window format, with each line representing a time snapshot and including calculated statistical and morphological features of the signal, as well as the corresponding emotional state label.

To evaluate the contribution of each feature to class discrimination, the SelectKBest technique was used in combination with the ANOVA F-test (f_classif) algorithm. This algorithm applies a one-way analysis of variance (ANOVA) and calculates the F statistic for each feature, examining whether the variation of the feature values between different classes is significantly greater than the variation within classes. The higher the F value, the more distinct the corresponding feature is considered.

Then, the 10 most important features were selected based on the highest F values, aiming to reduce the dimensionality of the problem and retain the most informative variables. Choosing the appropriate number of features (in this case k=10) is a crucial step, as it affects both the generalization of the model and the reduction of computational complexity.

Finally, a graph was created in which the F scores of the selected features were presented. This visualization provided a visual understanding of the relative importance of each of the features in the problem of classifying emotional states.

This methodology constitutes a representative example of supervised feature selection, where labels are actively used to estimate the contribution of each variable, appropriately preparing the dataset for training high-performance classifiers.

The Figure 37 depicts the importance of features for classifying emotional states, as calculated by different machine learning algorithms: Random Forest, Gradient Boosting, Logistic Regression, and SVM. The variation in the importance of features between algorithms highlights the different ways each method interprets the data. In particular, features such as Mean Negative Derivative and Mean Derivative show high importance in many models, suggesting that they may have strong discriminative power for classifying states. The lines corresponding to each algorithm provide a comparative overview of how each method identifies the most critical features.

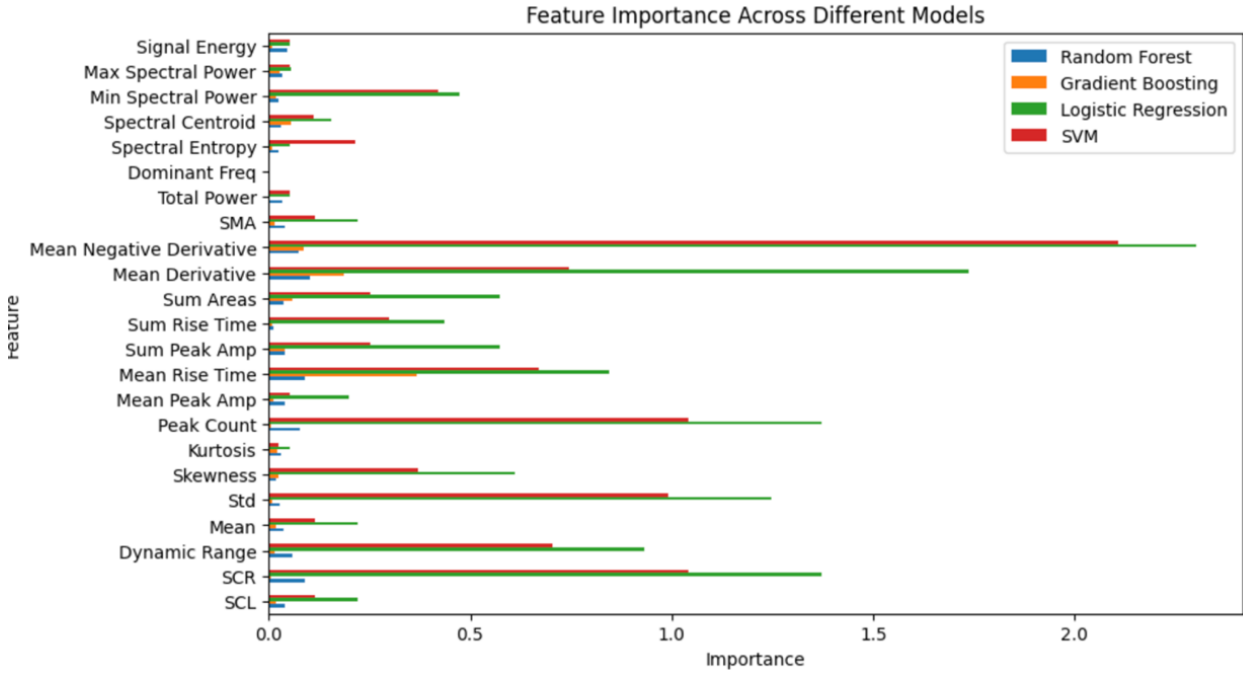


Figure 37 Feature importance comparison across different machine learning models for EDA signal

Figure 38 shows the overall importance of the features, as derived from the combined evaluation of the models. It is observed that the Mean Negative Derivative and the Mean Derivative have the highest importance, followed by the Mean Rise Time and the Standard Deviation (Std). These features are mainly related to the dynamic change of the EDA signal, which indicates that temporal variation plays an important role in the discrimination of emotional states. The classification of the features in descending order of importance allows the identification of the most informative variables, providing a strong basis for further optimization of the classification model.

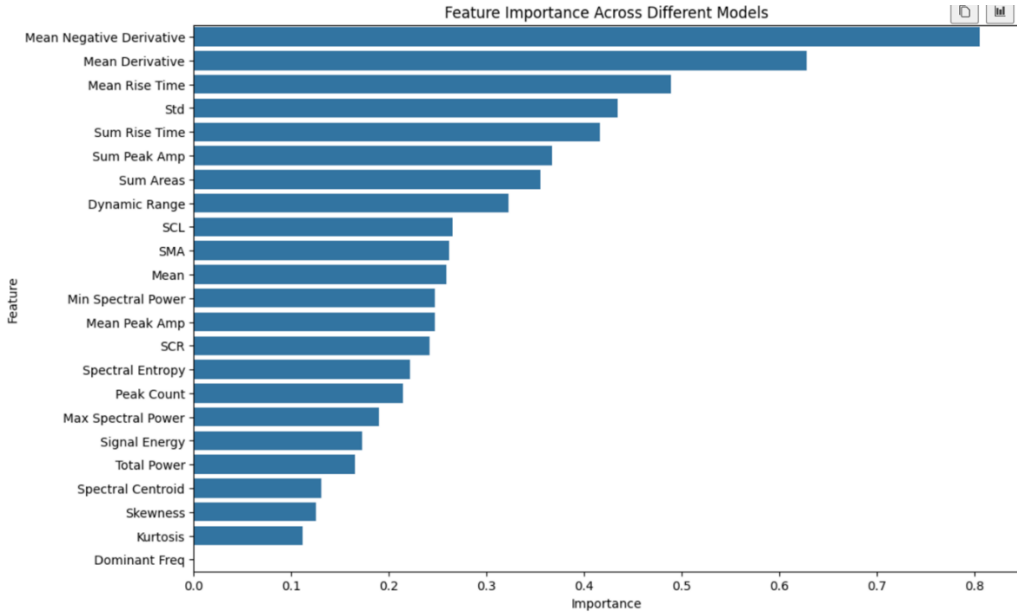


Figure 38 Overall feature importance across all models for EDA signal

In the context of the analysis of Blood Volume Pulse (BVP) signals, a combinatorial feature selection approach was applied to highlight the most discriminative variables for the classification of emotional states. Four different classifiers were used to estimate the importance of each feature: Random Forest, Gradient Boosting, Logistic Regression, and Support Vector Machine (Linear SVM). The use of multiple models allows for a more stable and reliable estimate of the contribution of each feature, limiting the dependence on the specificities of a single algorithm.

After normalization of the data, each model was trained, and the relative importance of the features was calculated. For the trees (Random Forest and Gradient Boosting) the intrinsic importance metrics (feature importances) were utilized, while for the linear models (Logistic Regression and SVM) the absolute values of the coefficients were used. Figure 39

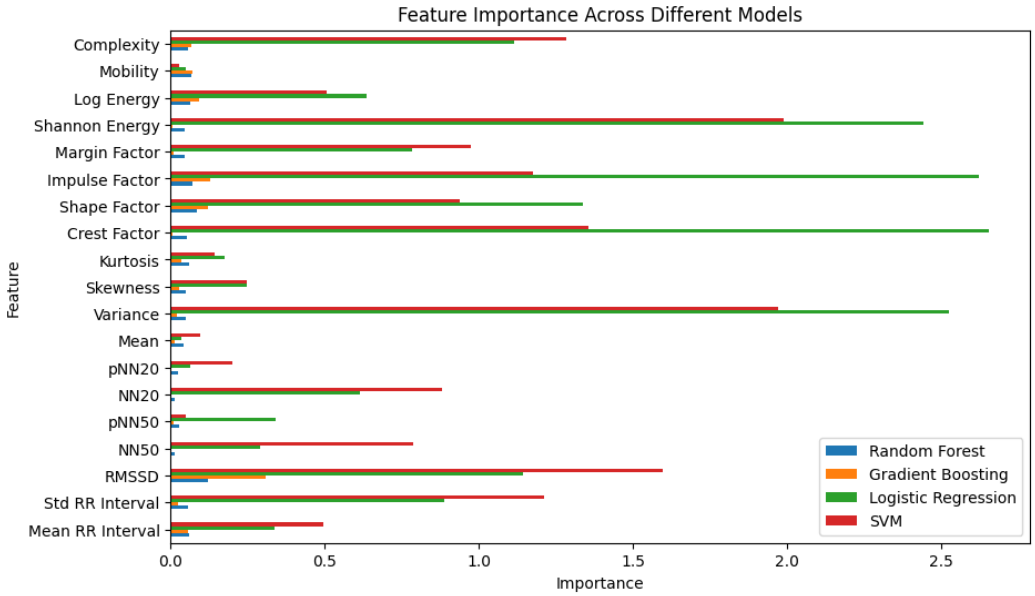


Figure 39 Feature importance comparison across different machine learning models for BVP signal

The combined classification revealed that the features with the highest average informative value were Variance, Mean RR Interval, Shannon Energy, RMSSD, and Std RR Interval as shown in Figure 40. These variables describe both the temporal variation and the energy behavior of the cardiovascular signal, which is crucial for the discrimination of emotional states. This approach contributed decisively to reducing the dimension of the problem and creating a more efficient set of features for the next classification phase.

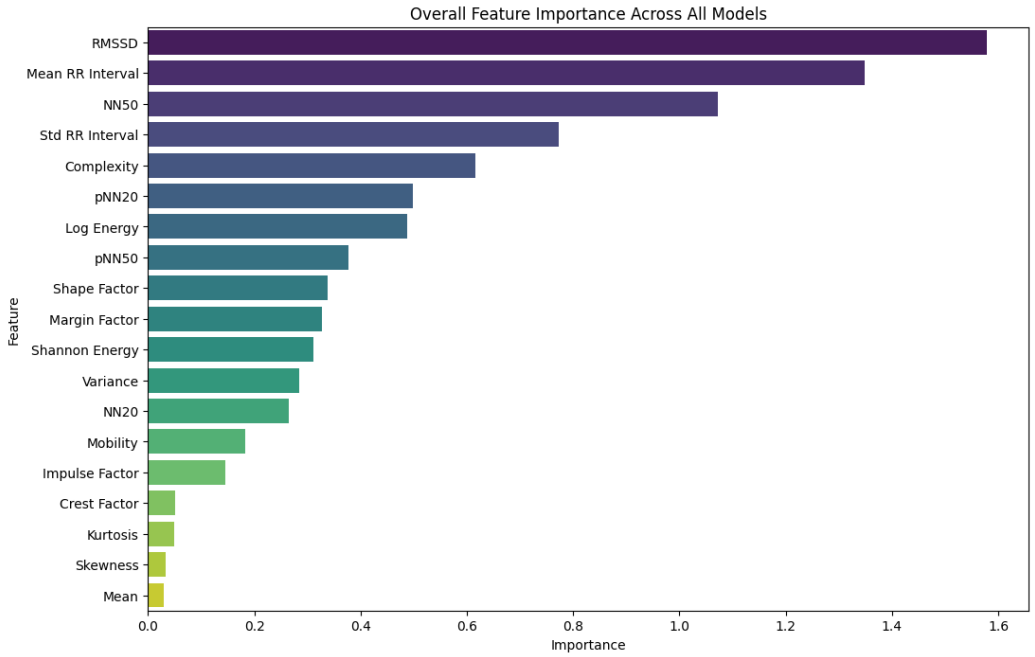


Figure 40 Overall feature importance across all models for BVP signal comment.

Based on the results obtained from the feature selection algorithms, we proceeded to create six separate datasets for each physiological signal. Specifically, for each overlap configuration, two datasets were generated: one containing the top 10 most important features and another containing only the top 5. This process enabled a comparative evaluation of model performance under varying dimensionalities and data segmentation strategies.

4.3. Implementation of machine learning algorithms

To analyze and classify emotional states in the WESAD database, machine learning algorithms were implemented on both Electrodermal Activity (EDA) and Blood Volume Pulse (BVP) signals. The process initially included the filtered selection of states: Baseline, Stress, amusement and meditation. The data were pre-processed through normalization and separation into training and testing sets. Then, six different supervised classifiers were trained and evaluated: Random Forest, Gradient Boosting, Logistic Regression, Support Vector Machine (SVM), k-Nearest Neighbors (KNN) and Naive Bayes, using as input the 5 or 10 most important features of each signal, as selected from a previous feature filtering process. The performance of each model was evaluated through key metrics — accuracy, precision and recall — while the results were represented with confusion matrices for better interpretation. The implementation of this multi-model approach on two different types of biomarkers allowed the comparative study of the effectiveness of the models on different physiological parameters, enhancing the generalizability of the findings for automatic stress recognition.

5.Results

In this section we report on the performance of different machine learning algorithms in classifying stress and non-stress states using features derived from EDA and BVP signals. The tests were conducted with different time window strategies and with different levels of overlap, and a detailed analysis of the results is presented.

The results indicate that using more features improves the performance of the classification algorithms compared to using only the 5 most important features. Specifically, it is observed that the accuracy of all classifiers is higher when 10 features are used compared to 5. The difference is more evident in multiclass classification where the use of more features offers better separability between categories. More specifically, using Eda in binary classification using labels one and two using 10 features, random forest achieved an accuracy of 94.31% with 50% overlap, while in the corresponding category using 5 features it achieved 92.16%. In a corresponding case in multiclass classification, gradient boosting showed an accuracy of 80.56% using 10 features while it decreased to 60.84% using 5 features.

Overall, the performance of the classifiers differs with random forest and gradient boosting showing consistently better results, especially in binary classifications.

Naïve Bayes presented the lowest performance in all classifications. This can be considered expected since their role is based on the assumption that the features are independent of each other, which is not true in biological signals. In multiclass classification, its accuracy did not exceed 0.50%.

Regarding the effect of temporal overlap, introducing 25% and 50% overlap seems to improve the classification of signals. This may be due to the fact that overlapping windows offer more context to the changes in the physiological signals.

In line with literature findings the classification between binary and multiclass classification differs greatly, revealing the challenge in recognizing different emotions. The difficulties in recognizing all classes are due to the lack of strong differences between the 4 emotional states, i.e. baseline vs amusement. More specifically, the accuracy of the classifiers drops by 15-20% in multiclass classification.

5.1 Eda classification results (with feature selection)

Regarding EDA the maximum value in binary currency was achieved by random forest with a value of 94.31% (60 sec, 50% Overlap, Top 10 Features). Regarding classification (1,3,4 as non-stress and 2 as Stress) maximum value: 93.12% by KNN (60 sec, 25% Overlap, Top 10 Features). End in multiclass classifications maximum value: 80.56% by Random Forest (60 sec, 50% Overlap, Top 10 Features).

in the following tables as label 1 we refer to the baseline state, as label 2 to the stress state, as label 3 to the amusement state and finally as label 4 to the meditation state

As shown in Table 14, the best results for the set of 10 features were achieved by the Gradient Boosting and Random Forest models, especially in the Binary classification of Stress (label 2) versus non-Stress conditions (labels 1, 3, and 4). scenario, where accuracy of 91.09% and 89.73% was recorded respectively. Multiclassification presented lower accuracy values, with Random Forest excelling, reaching 72.60%. Correspondingly, in

Table 15 it is observed that the use of the top 5 features leads to slightly reduced performance, especially in multiclassification, while simple binary classification (1 vs 2) remains at high levels, with KNN and Gradient Boosting reaching or exceeding 87%. These results highlight the importance of choosing both the appropriate number of features and the model, especially when the goal concerns complex classifications between multiple emotional states.

Table 14 Classification performance of different machine learning models in EDA signal using the top 10 features with a 60-second. (Label 1: Baseline, Label 2 : Stress, Label 3 : Amusement , Label 4: Meditation)

Model	1&2	1,3,4 as non-stress and 2 as stress	Multiclass
Random Forest	**92,13 %**	89.72%	71.23%
Gradient Boosting	91,01%	**91.09%**	**71.91%**
Logistic Regression	88,76%	86.93%	60.27%
SVM	89,88%	87.67%	54.79%
KNN	86,51%	88.35%	58.21%

Naïve Bayes	74,15%	77,39%	38.35%

Table 15 Classification performance of different machine learning models in EDA signal using the top 5 features with a 60-second. (Label 1: Baseline, Label 2 : Stress, Label 3 : Amusement , Label 4: Meditation)

Model	1&2	1,3,4 as non-stress and 2 as stress	Multiclass
Random Forest	87,64%	87.67%	68.49%
Gradient Boosting	88,76%	**89.72%**	**69.86%**
Logistic Regression	89,88%	85.61%	56.16%
SVM	**91,01%**	84.24%	56.84%
KNN	83,14%	86.98%	52.73%
Naïve Bayes	82,02%	84.93%	39.72%

In the second phase of the evaluation, a 60-second window with 25% overlap was applied in order to increase the number of available samples and to examine the impact of temporal overlap on the performance of the models. As in the previous case, two feature variants were used: the set of the 10 most important and the set of the 5 most important. The results are recorded in Table 16 and Table 17 respectively.

The results from Table 16 show a significant improvement in the performance of the classifiers compared to the case without overlap. Random Forest shows impressive performance in all three classification scenarios, reaching 92,3 % accuracy in "1 vs 2", 91.7% in Binary classification of Stress (label 2) versus non-Stress conditions (labels 1, 3, and 4). and 73.15% in multiclassification. Gradient Boosting also records very strong performance, especially in binary classification scenarios.

Table 16 Classification performance of different machine learning models in EDA signal using the top 10 features with a 60-second with 25% overlap. (Label 1: Baseline, Label 2 : Stress, Label 3 : Amusement , Label 4: Meditation)

Model	1&2	1,3,4 as non-stress and 2 as stress	Multiclass

Random Forest	**92.3%**	**91.7%**	**73.15%**
Gradient Boosting	88.9%	93.65%	73.01%
Logistic Regression	85.47%	86.77%	55.02%
SVM	87.17%	88.87%	55.55%
KNN	89.94%	91.12%	65.6%
Naïve Bayes	74.35%	79.36%	47.19%

Correspondingly, Table 17 confirms that even with a reduced number of features (Top 5), the models maintain high accuracy rates, especially in binary problems. Random Forest and Gradient Boosting remain the most reliable models, while in multiclassification Gradient boosting records a remarkable accuracy of 60.84%, despite the reduced dimensionality. In general, introducing temporal overlap seems to enhance the performance of the models, probably due to the more complete representation of emotional changes over time.

Table 17 Classification performance of different machine learning models in EDA signal using the top 5 features with a 60-second with 25% overlap. (Label 1: Baseline, Label 2 : Stress, Label 3 : Amusement , Label 4: Meditation)

Model	1&2	1,3,4 as non-stress and 2 as stress	Multiclass
Random Forest	85.47%	89.94%	55.55%
Gradient Boosting	**88.88%**	**92.06%**	**60.84%**
Logistic Regression	85.47%	85.18%	58.2%
SVM	88.03%	89.94%	54.49%
KNN	83.76%	89.93%	57.14%
Naïve Bayes	76.92%	80.95%	45.4%

In the third and final phase of the EDA signal experiments, a 60-second window with 50% overlap was applied, which offers even higher sample density compared to the previous cases. The evaluation process followed the structure of the previous scenarios, using the 10 and 5 most important features and the accuracy metric as the main evaluation criterion. The results are summarized in Table 18 and Table 19.

As shown in Table 18, the Gradient Boosting and Random Forest models continue to excel, with accuracies reaching 94.31% and 93.8% respectively in the “1 vs 2” scenario, while maintaining excellent performance in the Binary classification approach distinguishing the "Stress" class (label 2) from all other classes ("Baseline", "Amusement", and "Meditation" –

labels 1, 3, and 4 respectively) scenario with values close to 90–92%. Their performance in multiclass remains the highest among all models, with accuracies of 80.56%, confirming the stability and power of these algorithms in complex classification problems.

Table 18 Classification performance of different machine learning models in EDA signal using the top 10 features with a 60-second with 50% overlap. (Label 1: Baseline, Label 2 : Stress, Label 3 : Amusement , Label 4: Meditation)

Model	1&2	1,3,4 as non-stress and 2 as stress	Multiclass
Random Forest	**94.31%**	91.99%	**80.56%**
Gradient Boosting	93.18%	90.93%	73.85%
Logistic Regression	86.93%	85.51%	59.01%
SVM	87.5%	85.86%	56.53%
KNN	92.61%	**92.22%**	68.18%
Naïve Bayes	72.15%	80.91%	46.99%

In Table 19, where only the top 5 features were used, an expected decrease in the accuracy of the models is observed, mainly in multiclass. However, both Random Forest and Gradient Boosting maintain satisfactory performance, with accuracies of 73.49% and 74.91% respectively in multiclass. The general picture suggests that high temporal overlap improves the generalizability of models in binary scenarios, while requiring more information (more features) for efficient multiclassification.

Table 19 Classification performance of different machine learning models in EDA signal using the top 5 features with a 60-second with 50% overlap. (Label 1: Baseline, Label 2 : Stress, Label 3 : Amusement , Label 4: Meditation)

Model	1&2	1,3,4 as non-stress and 2 as stress	Multiclass
Random Forest	**92.16%**	**91.51%**	73.49%
Gradient Boosting	91.17%	91.16%	**74.91%**
Logistic Regression	85.22%	82.33%	55.47%
SVM	87.5%	84.45%	55.43%
KNN	89.2%	90.45%	65.37%
Naïve Bayes	76.0%	80.91%	42.04%

In Figure 41,Figure 42,Figure 43 are presented the confusing numbers with the highest performance for the EDA signal.

Figure 41 shows the overall performance of different machine learning algorithms using the 10 most important features of the EDA signal, as derived from 60-second windows with 50%

overlap. The visualization focuses on the scenario with the highest overall performance, allowing for a direct comparison between models in terms of classification accuracy. Random Forest recorded the highest accuracy in this scenario, demonstrating high generalization ability even with a limited number of features.

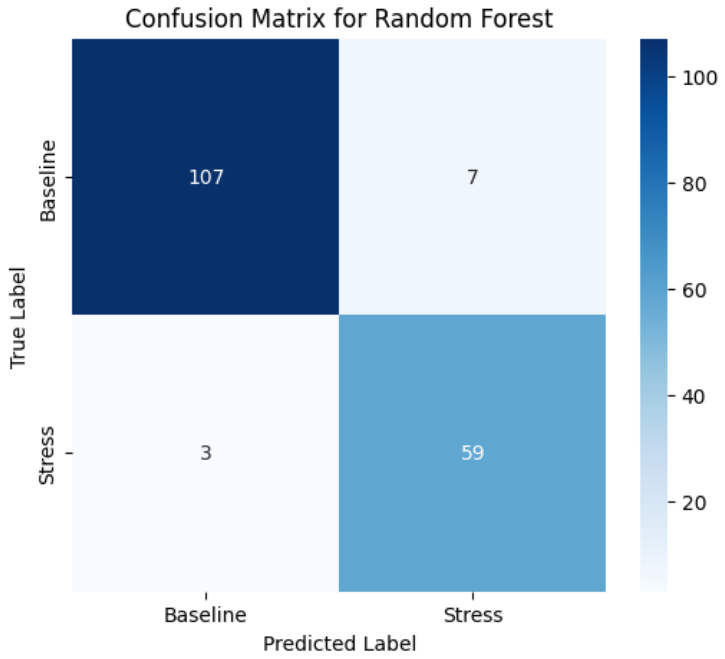


Figure 41 Confusion matrix for binary classification using labels 1 & 2 stress non-stress for Eda signal.

Figure 42 depicts the confusion matrix for binary classification where class 2 (stress) is distinguished against all the rest (1, 3, 4), which are considered non-stress. The analysis is based on the EDA signal and aims to evaluate the model's ability to accurately identify stress states. As can be seen from the illustration, the model achieves a high percentage of correct classifications for the stress category, while incorrect predictions remain limited, indicating reliable behavior in real psychophysiological stress recognition scenarios.

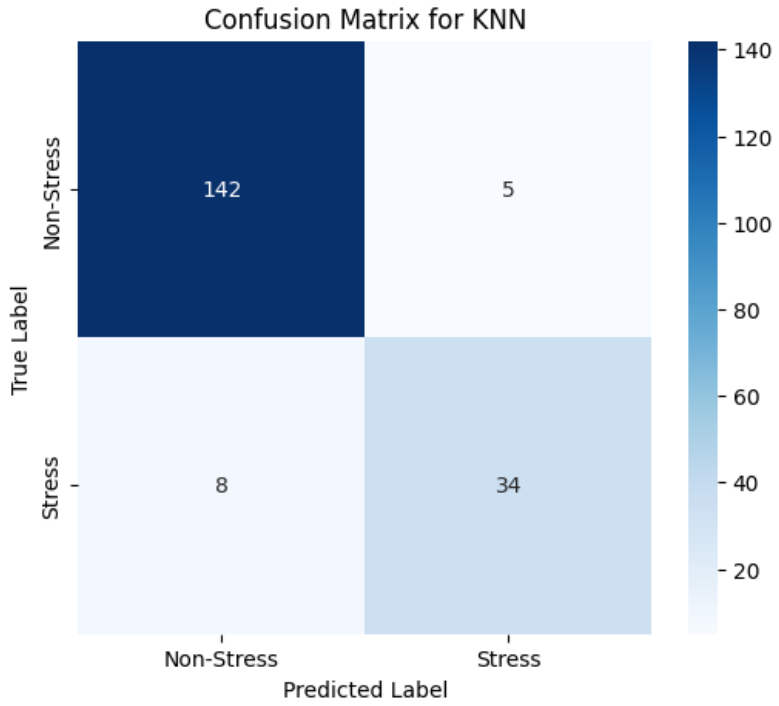


Figure 42 Confusion matrix for binary classification using labels 1,3,4 as non-stress and 2 as stress for EDA signal.

In Figure 43 we are presenting the confusion matrix for the Random Forest model, which was applied to the multiclassification of the four emotional states of the EDA signal: Baseline (1), Stress (2), Amusement (3) and Meditation (4). The model records high accuracy for the Baseline (210 correct predictions) and Meditation (132 correct predictions) categories, while the discrimination of the Stress category (115 correct predictions, 8 errors) remains consistently satisfactory. In contrast, the Amusement (3) class presents the greatest difficulty in classification, with 12 cases classified as Baseline and 5 as Stress, indicating a possible overlap in the characteristics of these states.

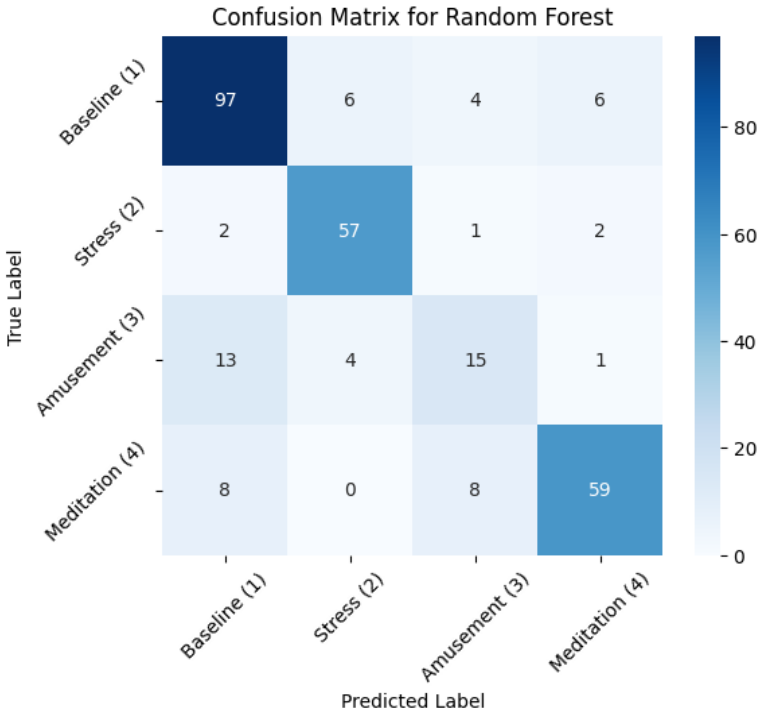


Figure 43 Confusion matrix for multiclass classification for EDA signal.

5.2 BVP classification results (with feature selection)

Regarding BVP, random forest once again had consistently higher performance. In binary classification (1&2 - Baseline vs Stress) The best performance was achieved by GB (93.65 %) at (60 sec, 25% Overlap, Top 10 Features). In classification (1,3,4 as non-stress και 2 as Stress) highest accuracy was achieved by RF (93.83%) at (60 sec, 0% Overlap, Top 10 Features). End in multiclass classification Gradient Boosting (78.45%) in (60 sec, 50% Overlap, Top 10 Features).

Table 20 presents the classification results using the 5 most important features of the BVP signal with a 60-second non-overlapping window. The evaluation was performed in the three classification scenarios: (a) binary classification (1 vs 2), (b) discrimination of stress (2) vs all others (1, 3, 4), and (c) multiclassification among the four categories. It is observed that using only 5 features leads to generally lower accuracy compared to the corresponding experiments using 10 features. In binary classification (1 vs 2), GB recorded the highest accuracy (86.27%), marginally surpassing Random Forest (84.31%). In the Binary classification of Stress (label 2) versus non-stress conditions (labels 1, 3, and 4). classification, the best results were recorded by Random Forest (92.69%) and Gradient Boosting (92.59%), confirming that the linear separability of the problem can be maintained even with a limited number of features. In multiclassification, Random Forest recorded the highest accuracy (72.94%), followed by Gradient Boosting (72.84%) and KNN (58.02 %).

These results suggest that ensemble models (Random Forest, Gradient Boosting) still have an advantage in multiclassification, even when the available features are fewer.

Table 20 Classification performance of different machine learning models in BVP signal using the top 5 features with a 60-second window. (Label 1: Baseline, Label 2 : Stress, Label 3 : Amusement, Label 4: Meditation)

Model	1&2	1,3,4 as non-stress and 2 as stress	Multiclass
Random Forest	84.31%	**92.69%**	**72.94%**
Gradient Boosting	**86.27%**	92.59%	72.84%
Logistic Regression	78.43%	81.48%	45.68%
SVM	84.11%	82.72%	44.44%
KNN	84.30%	83.95%	58.02%
Naïve Bayes	68.63%	33.33%	45.68%

In Table 21, the classification results using the 10 most important features of the EDA signal with a 60-second window without overlap are presented. Using 10 features led to an improvement in the accuracy of most models compared to the corresponding classification with 5 features Table 20. In binary classification (1 vs 2), Gradient boosting had the highest accuracy (90.20%), followed by SVM (88.24%) and Random Forest (86.27%). The increase in the available features enhanced the discriminative ability of the models, allowing for more reliable predictions. In the Binary classification of Stress (label 2) versus non-Stress conditions (labels 1, 3, and 4) scenario, Random Forest and Gradient Boosting showed excellent performance with 93.83 % and 91.36% accuracy respectively, demonstrating that decision trees models work well when sufficient information is available. In multiclassification, the highest performance was recorded by KNN (72.84 %), followed by RF (71.60 %) and Gradient Boosting (69.14%). It is observed that the performance of multiclassification is still lower compared to binary problems, indicating that different emotional states have greater overlap in their features.

Table 21 Classification performance of different machine learning models in BVP signal using the top 10 features with a 60-second. (Label 1: Baseline, Label 2 : Stress, Label 3 : Amusement , Label 4: Meditation)

Model	1&2	1,3,4 as non-stress and 2 as stress	Multiclass
Random Forest	86.27%	**93.83%**	71.60%
Gradient Boosting	**90.20%**	91.36%	69.14%
Logistic Regression	82.35%	83.95%	56.79%

SVM	88.24%	90.02%	59.26%
KNN	84.31%	90.12%	**72.84%**
Naïve Bayes	86.27%	40.74%	56.79%

The classification results across the three different labeling scenarios demonstrate notable variations in model performance. In the binary classification between labels 1 and 2, Gradient boosting achieved the highest accuracy (88.89%), followed closely by RF (88.03%) and SVM (88.13%). When combining labels 1, 3, and 4 as non-stress versus label 2 as stress, the performance generally improved, with Gradient Boosting again leading at 91.01% accuracy, and Random Forest (90.48%) and SVM (82.54%) also showing strong results. In the more challenging multiclass scenario, all models exhibited lower accuracy, with Random Forest achieving the highest performance at 71.96%, and Gradient Boosting following closely at 69.84%. These results indicate that while models perform robustly under binary classification, their effectiveness diminishes when classifying multiple stress-related states, highlighting the increased complexity of multiclass stress detection. The results are presenting in

Table 22

Table 22 Classification performance of different machine learning models using the top 5 features with a 60-second with 25% overlap. (Label 1: Baseline, Label 2 : Stress, Label 3 : Amusement , Label 4: Meditation)

Model	1&2	1,3,4 as non-stress and 2 as stress	Multiclass
Random Forest	88.03%	90.48%	**71.96%**
Gradient Boosting	**88.89%**	**91.01%**	69.84%
Logistic Regression	87.18%	82.54%	60.32%
SVM	88.13%	83.07%	56.08%
KNN	83.76%	80.95%	54.50%
Naïve Bayes	76.07%	30.69%	50.79%

Compared to

Table 22 (Top 5 Features), the use of 10 features led to an overall improvement in performance across all classification scenarios, confirming that retaining more information contributes to the accuracy of the models. In the second and third scenarios, which concern the discrimination of stress from the other states, the Random Forest and Gradient Boosting algorithms recorded the highest classification accuracy (93.65% and 92.06% respectively), which suggests that they are particularly effective in detecting stress states based on selected features. The SVM method also achieved high performance (92.59%), while KNN

showed similar stability (91.53%). In contrast, the Naïve Bayes method showed a significant drop in accuracy (33.86%) when classifying stress versus non-stress, which may be due to a violation of the statistical assumptions of the method or its unsuitability for this problem. In the first scenario (1 vs. 2), the highest performance was recorded by Gradient Boosting (93.65%), while SVM and Logistic Regression showed lower rates (82.01% and 92.06% respectively). Overall, the results highlight the superiority of ensemble methods (Random Forest, Gradient Boosting) and the need for careful algorithm selection depending on the classification goal and data characteristics.

Table 23 Classification performance of different machine learning models using the top 10 features with a 60-second with 25% overlap. (Label 1: Baseline, Label 2 : Stress, Label 3 : Amusement , Label 4: Meditation)

Model	1&2	1,3,4 as non-stress and 2 as stress	Multiclass
Random Forest	87.18%	**93.65%**	**93.65%**
Gradient Boosting	**93.65%**	92.06%	92.06%
Logistic Regression	92.06%	82.01%	82.01%
SVM	82.01%	92.59%	92.59%
KNN	92.59%	91.53%	91.53%
Naïve Bayes	91.53%	33.86%	33.86%

In In the stress discrimination scenario from the other states (2 vs. 1,3,4), the Random Forest and Gradient Boosting algorithms prevailed in accuracy with 89.25% and 89.05% respectively, which demonstrates the reliability of ensemble models even with a limited number of features. Their performance remains consistently high, although the reduction of features slightly affected the results compared to Table 23. Logistic Regression and SVM presented reduced performance (80.21% and 79.86% respectively), while the Naïve Bayes method recorded the lowest accuracy (32.16%), indicating its unsuitability for this type of problem under the specific constraints.In the multi-class scenario, a significant drop in accuracy is observed in all algorithms. The best performance is recorded by Random Forest with 74.91%, followed by Gradient Boosting with 72.44%, while the other methods are at low levels, with Naïve Bayes scoring only 48.41%. This result suggests that the reduction of features in combination with the increased complexity of multiclassification negatively affects the accuracy of the models. Overall, the findings reinforce the position that ensemble models, and especially Random Forest, maintain high performance even with a limited number of features, offering an effective solution to emotional or physiological state recognition applications with limited computational cost. In contrast, models such as Naïve

Bayes seem to be significantly affected by the reduction of discriminative information in the data.

Table 24, the classification results are presented using the 5 most important features of the BVP signal, with a 60-second window and 50% overlap. In the stress discrimination scenario from the other states (2 vs. 1,3,4), the Random Forest and Gradient Boosting algorithms prevailed in accuracy with 89.25% and 89.05% respectively, which demonstrates the reliability of ensemble models even with a limited number of features. Their performance remains consistently high, although the reduction of features slightly affected the results compared to Table 23**.** Logistic Regression and SVM presented reduced performance (80.21% and 79.86% respectively), while the Naïve Bayes method recorded the lowest accuracy (32.16%), indicating its unsuitability for this type of problem under the specific constraints.In the multi-class scenario, a significant drop in accuracy is observed in all algorithms. The best performance is recorded by Random Forest with 74.91%, followed by Gradient Boosting with 72.44%, while the other methods are at low levels, with Naïve Bayes scoring only 48.41%. This result suggests that the reduction of features in combination with the increased complexity of multiclassification negatively affects the accuracy of the models. Overall, the findings reinforce the position that ensemble models, and especially Random Forest, maintain high performance even with a limited number of features, offering an effective solution to emotional or physiological state recognition applications with limited computational cost. In contrast, models such as Naïve Bayes seem to be significantly affected by the reduction of discriminative information in the data.

Table 24 Classification performance of different machine learning models using the top 5 features with a 60-second with 50% overlap. (Label 1: Baseline, Label 2 : Stress, Label 3 : Amusement , Label 4: Meditation)

Model	1&2	1,3,4 as non-stress and 2 as stress	Multiclass
Random Forest	85.23%	**89.25%**	**74.91%**
Gradient Boosting	83.52%	89.05%	72.44%
Logistic Regression	86.93%	80.21%	56.89%
SVM	**87.50%**	79.86%	59.01%
KNN	84.09%	77.74%	54.06%
Naïve Bayes	69.89%	32.16%	48.41%

Continuing to present our results in Table 25, the classification results are presented using the 10 most important features of the BVP signal, with a 60-second window and 50% overlap. In the “stress vs. non-stress” scenario (Label 2 vs. 1, 3, 4), Random Forest recorded the highest accuracy (92.62%), followed closely by Gradient Boosting (92.58%) and KNN

(91.52%), confirming once again the potential of ensemble and instance-based methods in accurately distinguishing psychophysiological stress states. Logistic Regression and SVM recorded lower performances (80.92% and 89.05% respectively), while Naïve Bayes dropped significantly (35.34%), revealing limited generalization ability in this particular problem. Regarding multiclassification, the overall performance decreased, which is expected due to the increased complexity. The highest accuracy was recorded by Random Forest (78.65%) and Gradient Boosting (78.45%), while KNN also performed at satisfactory levels (73.14%). In contrast, Logistic Regression and SVM recorded more moderate performances (64.66% and 67.49% respectively), while Naïve Bayes remained below the 55% threshold (53.71%). This comparative evaluation demonstrates the importance of choosing the appropriate classification algorithm, with ensemble models maintaining an advantage in accuracy and stability, especially when rich and discriminative features are used. At the same time, the use of 50% overlap in combination with 60-second windows seems to enhance the classification ability of the models, allowing for more effective capture of dynamic changes in physiological signals.

Table 25 Performance of different machine learning models using the top 10 features with a 60-second with 50% overlap. (Label 1: Baseline, Label 2 : Stress, Label 3 : Amusement , Label 4: Meditation)

Model	1&2	1,3,4 as non-stress and 2 as stress	Multiclass
Random Forest	**87.50%**	92.62%	78.65%
Gradient Boosting	86.93%	92.58%	**78.45%**
Logistic Regression	85.80%	80.92%	64.66%
SVM	86.93%	89.05%	67.49%
KNN	89.20%	91.52%	73.14%
Naïve Bayes	76.14%	35.34%	53.71%

In **Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε.,Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε.,Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε.** are presenting the confusing matrixes with the highest performance for the BVP signal

Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε. shows the confusion matrix for binary classification between Baseline (1) and Stress (2) states, using the BVP (Blood Volume Pulse) signal. The matrix records the frequency of correct and incorrect classifications, allowing the evaluation of the model performance.

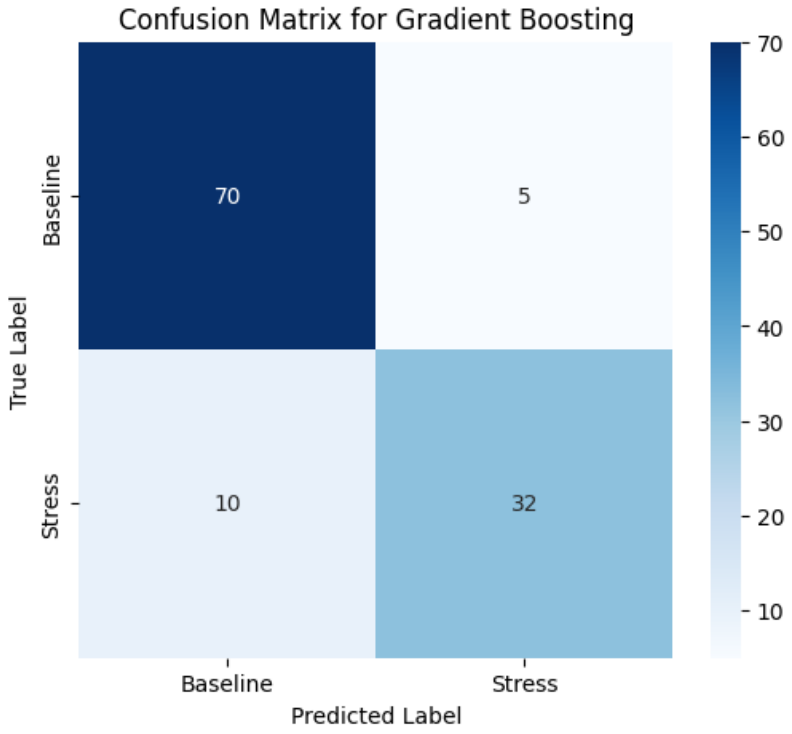


Figure 44 Confusion matrix for binary classification using labels 1 & 2 for BVP signal.

Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε. depicts the confusion matrix for the Random Forest model, which was applied to the binary classification between non-stress (1,3,4) and stress (2) states, using the BVP (Blood Volume Pulse) signal. The analysis suggests that, although the model has high accuracy in identifying non-stress states, it has greater difficulty in correctly classifying stress samples, which may be due to an overlap of BVP features between the two categories.

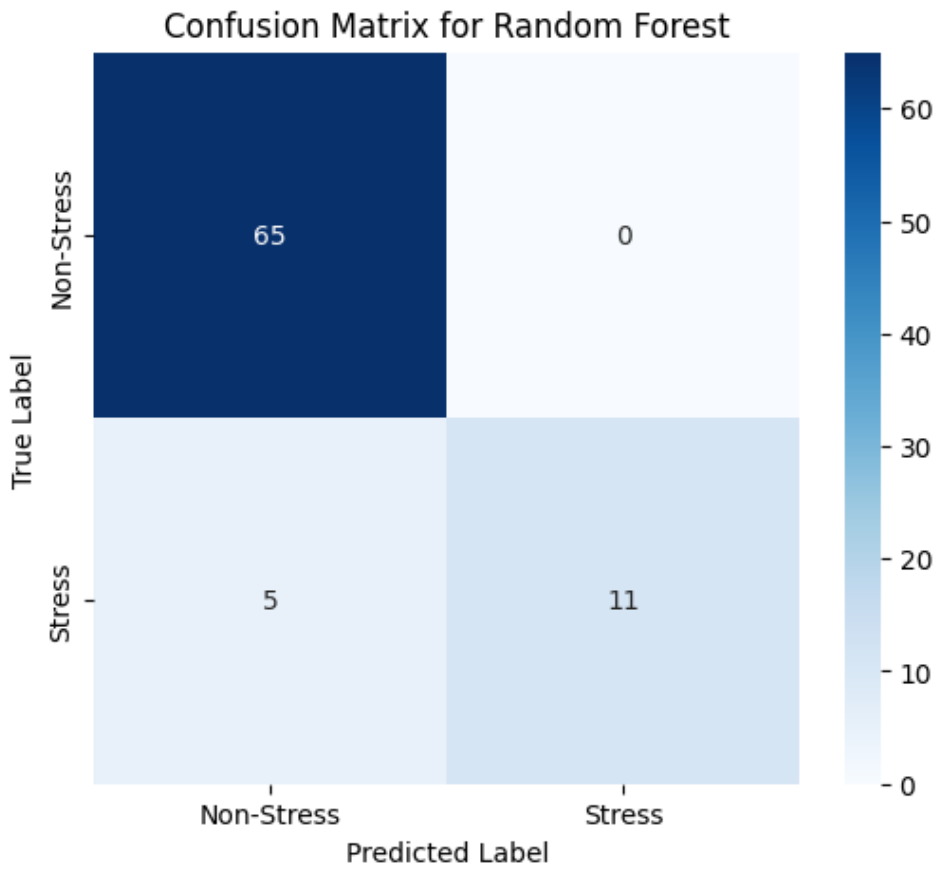


Figure 45 Confusion matrix for binary classification using labels 1,3,4 as non-stress and 2 as stress for BVP signal.

Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε. presents the confusion matrix for the Gradient Boosting model, which was applied to the multiclassification of the four emotional states (Baseline, Stress, Amusement, Meditation) using the BVP signal.

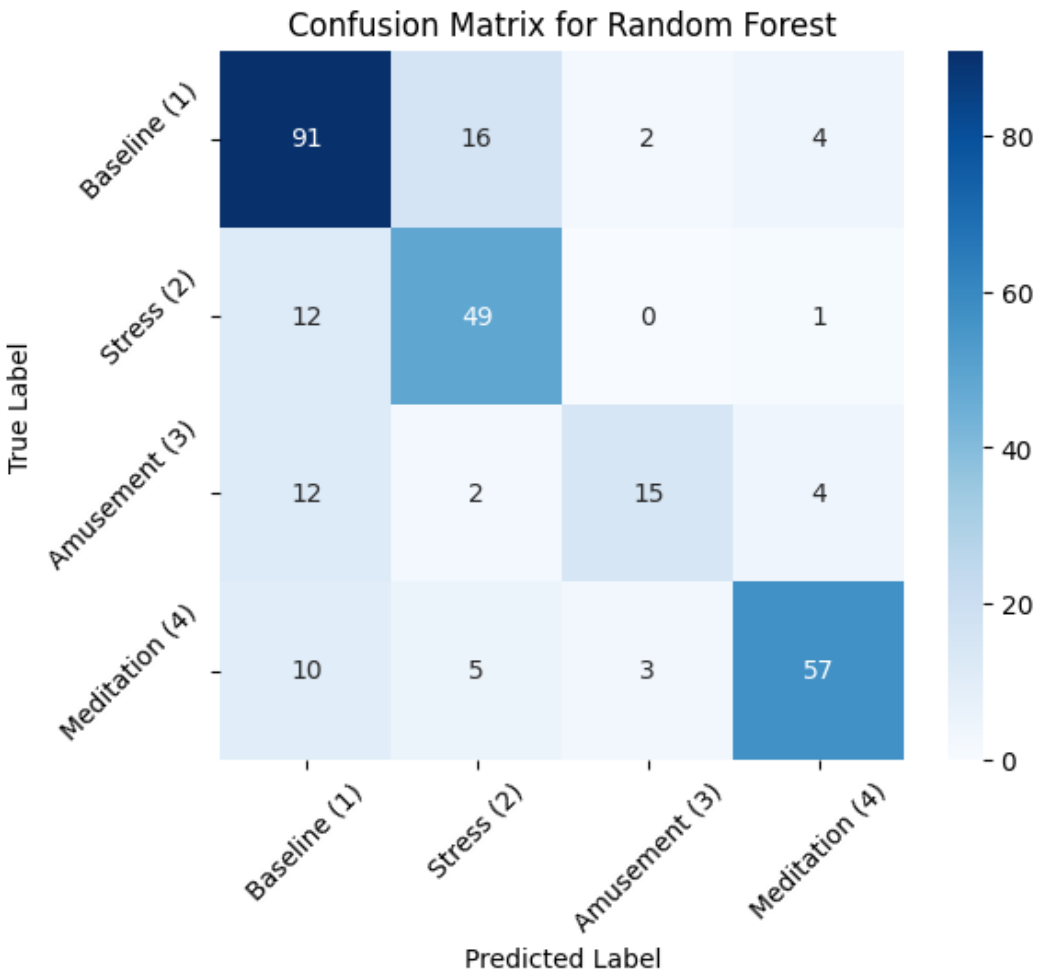


Figure 46 Confusion matrix for multiclass classification for BVP signal.

The strong confusion observed by classification algorithms between the Baseline, Amusement, and Meditation states—despite the fact that they belong to the same general category of "non-anxious" conditions—highlights a critical property of both the dataset and the nature of the physiological signals being recorded. Specifically, signals derived from these states (such as electrodermal activity and heart rate) often exhibit overlapping or similar patterns, resulting in reduced discriminability between the respective categories. For example, during both the Baseline and Meditation states, the individual is in a state of rest, with reduced mobility and stabilized cardiorespiratory activity. Furthermore, in the Amusement state which normally is characterized by emotional arousal, in the explored dataset is not accompanied by a strong physiological response for all participants and therefore, leads to physiological signs that resemble those of Baseline (e.g., transient fluctuations in EDA or heart rate without systematic change).

This intragroup similarity between the “non-stress” states makes it particularly difficult for algorithms to detect clear boundaries between them. Thus, when the Baseline, Amusement, and Meditation categories are unified into a general “non-stress” category and compared only with the Stress category (label 2), the classification becomes strikingly more accurate. This is due to the fact that the stress state presents a clearly differentiated physiological profile, such as increased electrodermal activity, tachycardia, and reduced heart rate variability — changes that create distinct boundaries in the feature space relative to the other conditions.

Therefore, the observed increase in accuracy in classification scenarios type 1,3,4 vs 2, even when more data and more complex groups are involved, is not a statistical anomaly, but rather confirms that the psychophysiological distance between stress and the remaining states is more pronounced than the corresponding one between the individual non-stress states. In other words, the improvement in performance does not come from the reduction of the complexity of the problem, but from the increased physiological differentiation that allows the model to create cleaner and more distinct decision surfaces.

5.3 EDA and BVP classification results (with all features)

At this stage of the analysis, the performance of various machine learning algorithms was evaluated using all available features, i.e. without a feature selection process. The evaluation was performed on BVP and EDA signals with 60-second windows and 0%, 25% and 50% overlap, under three different classification scenarios: (a) labels 1 and 2 (Baseline vs Stress), (b) stress vs the remaining states (1,3,4 vs 2), and (c) multiclassification of all labels.

Regarding BVP signals, the Random Forest model showed excellent performance, with an accuracy of 90.22% in scenario 1&2 without overlap and 91.45% with 25% overlap, while with 50% overlap it reached 86.93%. In the stress vs. non-stress scenario, the highest accuracy was achieved with 93.83% by Gradient Boosting (0% overlap) and 93.29% by Random Forest (50% overlap). In multiclassification, the highest accuracy was again achieved by Random Forest with 80.57%, followed by Gradient Boosting with 79.51%. Naïve Bayes had a significantly lower performance, ranging from 49.12% to 53.09% in multiclassification.

Correspondingly, in EDA signals, Random Forest prevailed in all scenarios, recording 96.02% in scenario 1&2 with 50% overlap, 94.71% in the stress vs non-stress scenario and 80.92% in multiclassification. Gradient Boosting also presented very good results with accuracies of 94.61% and 75.62% respectively. The remaining models performed at moderate levels (e.g.

KNN with 90.60% for 1&2 and 66.08% for multiclass), while the Naïve Bayes method again presented the lowest performance (42.76% in multiclass with 50% overlap).

The results in Table 26 Table 27 demonstrate that ensemble models (Random Forest and Gradient Boosting) offer consistently high performance even without feature selection, making them suitable for the analysis of physiological signals such as BVP and EDA. Furthermore, it is observed that the use of larger overlap (e.g. 50%) enhances the generalization ability of the models, especially in multiclass, where the greatest differences between the algorithms are observed.

Table 26 BVP classification results without feature selection methods

Overlap	Model	Baseline vs Stress (1&2)	non-Stress vs Stress (1,3,4 vs 2)	Multiclass (%)
0%	Random Forest	90.22%	91.36%	72.84%
	Gradient Boosting	88.24%	93.83%	72.64%
	Logistic Regression	82.35%	85.19%	60.49%
	SVM	78.43%	86.42%	59.26%
	KNN	90.20%	87.65%	71.60%
	Naïve Bayes	84.31%	51.85%	53.09%
25%	Random Forest	91.45%	92.06%	76.19%
	Gradient Boosting	87.18%	91.01%	74.07%
	Logistic Regression	88.03%	84.13%	63.49%
	SVM	88.89%	89.75%	66.14%
	KNN	85.47%	89.95%	69.31%
	Naïve Bayes	82.05%	41.27%	51.32%
50%	Random Forest	86.93%	93.29%	80.57%
	Gradient Boosting	88.64%	90.81%	79.51%
	Logistic Regression	87.50%	83.75%	65.02%
	SVM	87.60%	89.40%	66.43%
	KNN	88.64%	89.75%	71.73%
	Naïve Bayes	76.14%	44.88%	49.12%

Table 27 EDA classification results without feature selection.

Overlap	Model	Baseline vs Stress (1&2)	non-Stress vs Stress (1,3,4 vs 2)	Multiclass (%)
0%	Random Forest	89.89%	91.10%	73.29%
	Gradient Boosting	92.13%	89.73%	68.49%
	Logistic Regression	88.76%	86.79%	57.53%
	SVM	88.56%	86.99%	55.48%
	KNN	91.01%	90.41%	60.27%
	Naïve Bayes	71.91%	71.92%	40.41%
25%	Random Forest	91.45%	94.71%	73.02%
	Gradient Boosting	89.74%	94.61%	73.54%
	Logistic Regression	90.60%	88.89%	60.85%
	SVM	88.03%	89.42%	60.85%
	KNN	90.60%	91.53%	64.02%
	Naïve Bayes	70.09%	77.78%	45.50%
50%	Random Forest	96.02%	94.35%	80.92%
	Gradient Boosting	93.18%	92.23%	75.62%
	Logistic Regression	86.93%	85.87%	58.66%
	SVM	85.80%	86.22%	58.66%
	KNN	88.64%	90.46%	66.08%
	Naïve Bayes	67.61%	77.39%	42.76%

6. Conclusions and Future Work

Identification and Assessment of Physiological Stress Responses

Identifying and assessing an individual's physiological response to stress is a complex and multidimensional process that requires the integration of signal processing techniques, mechanical features, and machine learning methodologies. The first critical step in this process is the accurate classification of physiological states, as this is used as the basis for physiological signal analysis and growth calculations for stress detection. Given the dynamic and multifactorial nature of stress responses, the development of reliable and adaptive classification systems remains an active area of research in emotional computing and biomedical signal processing.

Among the most used physiological signals for stress detection, Electrodermal Activity (EDA) and Blood Volume Pulse (BVP) provide valuable information about the response of this nervous system to internal and external stressors. EDA measures changes in skin conductance, which are influenced by sweat gland activity and reflect sympathetic nervous system activation. In contrast, BVP, extracted from photoplethysmography (PPG) sensors, provides information about heart rate (HR) and heart rate variability (HRV), both of which are commonly used indicators of physiological arousal and stress. These biomarkers, when subjected to appropriate preprocessing and feature extraction, allow the development of robust models capable of detecting and classifying demand-related states with accuracy.

The Role of Wearable Sensors in Stress Monitoring

The increasing availability of wearable sensors has revolutionized physiological signal monitoring, providing a practical, non-intrusive, and continuous method for assessing an individual’s physiological state in real-world settings. Traditional laboratory methods for stress assessment often rely on controlled environments with rigorous experimental protocols, limiting their application to real-life scenarios. In contrast, wearable devices allow for real-time data collection in naturalistic settings, allowing researchers and practitioners to analyze stress responses in everyday activities. This shift toward continuous and ambulatory monitoring enhances the ecological validity of stress detection systems, making them more relevant for applications in healthcare, workplace stress management, and personal well-being monitoring.

Despite these advantages, developing reliable computational methodologies for stress detection remains a challenge. A key factor affecting the performance of classification models is the selection of features derived from physiological signals. The effectiveness of a stress detection system largely depends on the quality and relevance of the extracted features, as well as on the signal processing techniques used to improve data interpretation.

Typical preprocessing steps include noise reduction, artifact removal, filtering, and signal normalization, which play a critical role in improving the reliability of the extracted features.

Feature Selection and Machine Learning Approaches to Stress Classification

The feature engineering process involves identifying significant statistical, temporal, and frequency domain features that effectively characterize stress-related changes in physiological signals. In the case of EDA, the features typically extracted include tonic and phasic components, skin conductance level (SCL), and skin conductance responses (SCRs). For BVP-derived signals, features such as heart rate (HR), heart rate variability (HRV), low-frequency (LF) and high-frequency (HF) power, and interpulse intervals (IBI) are widely used for stress classification.

Selecting the most informative features is a critical step in improving classifier performance, as redundant or irrelevant features can lead to overfitting, increased computational complexity, and reduced model generalization. Various feature selection techniques, such as recursive feature elimination (RFE), principal component analysis (PCA), and information-based mutual selection, are commonly used to optimize the feature set.

Once an optimal feature set is determined, machine learning algorithms are applied to classify different stress states. Traditional classifiers such as Support Vector Machines (SVMs), k-Nearest Neighbors (k-NN), Decision Trees, and Random Forests (RF) have been widely used for trend classification. More recently, deep learning approaches, including Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have shown promise in automatically learning feature representations from physiological signals. However, deep learning models require large amounts of labeled data and high computational resources, which can pose challenges for real-time stress detection applications.

6.1 Future work

Although the study showed that machine learning models can achieve high accuracy in binary classifications, multiclass classification remains a challenge. Future research could focus on advanced feature selection methods such as feature clustering or the use of PCA dimensionality reduction techniques to better represent physiological signals.[121][122]

In addition, the extension of deeper neural network models, such as Long Short-Term Memory (LSTM) networks or Convolutional Neural Networks (CNNs), could improve the results of classifiers by considering the temporal interdependence of physiological signals. These algorithms have proven to be extremely efficient in modeling complex nonlinear

correlations between signals that can offer significant improvements in model generalization [121][123][124].

Another direction is to integrate additional data sources such as information from mobile devices that record behavioral and environmental characteristics in order to create a holistic model and stress value. Event analysis, which is based on the correlation of different events, could be a powerful tool for understanding stress mechanisms in real conditions.

Overall, the study highlighted the effectiveness of machine learning algorithms but also the challenges that exist in a possible generalization of models. Improvements in classifiers and adaptive techniques could drive future research in developing complex features that better capture the physiological stress response and integrating data from multiple sensors to increase the accuracy and reliability of classifications.

References

[1] B. S. Mcewen and E. Stellar, “Stress and the Individual Mechanisms Leading to Disease,” 1993. [Online]. Available: http://archinte.jamanetwork.com/

[2] G. P. Chrousos, “Stress and disorders of the stress system,” Jul. 2009. doi: 10.1038/nrendo.2009.106.

[3] 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.

[4] H. S. Choi, “Emotion Recognition Using a Siamese Model and a Late Fusion-Based Multimodal Method in the WESAD Dataset with Hardware Accelerators,” *Electronics (Switzerland)*, vol. 14, no. 4, Feb. 2025, doi: 10.3390/electronics14040723.

[5] E. Oliver and S. Dakshit, “Cross-Modality Investigation on WESAD Stress Classification,” Feb. 2025, [Online]. Available: http://arxiv.org/abs/2502.18733

[6] D. Shirley Benita, A. Shamila Ebenezer, L. Susmitha, M. S. P. Subathra, and S. Jeba Priya, “Stress Detection Using CNN on the WESAD Dataset,” in *ESIC 2024 - 4th International Conference on Emerging Systems and Intelligent Computing, Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 308–313. doi: 10.1109/ESIC60604.2024.10481604.

[7] M. A. Al Aleem, R. Mubarak, N. M. Salem, and I. Sadek, “A Deep Learning Approach Using WESAD Data for Multi-Class Classification with Wearable Sensors,” in *NILES 2024 - 6th Novel Intelligent and Leading Emerging Sciences Conference, Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 194–197. doi: 10.1109/NILES63360.2024.10753228.

[8] *2020 12th International Conference on Computational Intelligence and Communication Networks (CICN)*. IEEE, 2020.

[9] R. C. Kessler *et al.*, “Lifetime Prevalence and Age-of-Onset Distributions of DSM-IV Disorders in the National Comorbidity Survey Replication.”

[10] W. Wang, L. Wang, and C. Huang, “Event-Triggered Control for Guaranteed-Cost Bipartite Formation of Multi-Agent Systems,” *IEEE Access*, vol. 10, pp. 18338–18351, 2022, doi: 10.1109/ACCESS.2021.3086404.

[11] N. Rashid, L. Chen, M. Dautta, A. Jimenez, P. Tseng, and M. A. Al Faruque, “Feature Augmented Hybrid CNN for Stress Recognition Using Wrist-based Photoplethysmography Sensor,” Aug. 2021, [Online]. Available: http://arxiv.org/abs/2108.03166

[12] T. Islam and P. Washington, “Personalized Prediction of Recurrent Stress Events Using Self-Supervised Learning on Multimodal Time-Series Data,” Jul. 2023, [Online]. Available: http://arxiv.org/abs/2307.03337

[13] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, “CvxEDA: A convex optimization approach to electrodermal activity processing,” *IEEE Trans Biomed Eng*, vol. 63, no. 4, pp. 797–804, Apr. 2016, doi: 10.1109/TBME.2015.2474131.

[14] M. A. Fauzi, B. Yang, and P. Yeng, “Improving Stress Detection Using Weighted Score-Level Fusion of Multiple Sensor,” in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Nov. 2022, pp. 65–71. doi: 10.1145/3568231.3568242.

[15] H. F. Posada-Quintero and K. H. Chon, “Innovations in electrodermal activity data collection and signal processing: A systematic review,” Jan. 02, 2020, *MDPI AG*. doi: 10.3390/s20020479.

[16] F. Shaffer and J. P. Ginsberg, “An Overview of Heart Rate Variability Metrics and Norms,” Sep. 28, 2017, *Frontiers Media S.A*. doi: 10.3389/fpubh.2017.00258.

[17] R. Castaldo, P. Melillo, U. Bracale, M. Caserta, M. Triassi, and L. Pecchia, “Acute mental stress assessment via short term HRV analysis in healthy adults: A systematic review with meta-analysis,” 2015, *Elsevier Ltd*. doi: 10.1016/j.bspc.2015.02.012.

[18] G. E. Billman, “The LF/HF ratio does not accurately measure cardiac sympatho-vagal balance,” 2013. doi: 10.3389/fphys.2013.00026.

[19] G. Georgieva-Tsaneva, E. Gospodinova, M. Gospodinov, and K. Cheshmedzhiev, “Cardio-diagnostic assisting computer system,” *Diagnostics*, vol. 10, no. 5, May 2020, doi: 10.3390/diagnostics10050322.

[20] M. Brennan, M. Palaniswami, and P. Kamen, “Do Existing Measures of Poincaré Plot Geometry Reflect Nonlinear Features of Heart Rate Variability?,” 2001.

[21] V.-T. Ninh *et al.*, “An Improved Subject-Independent Stress Detection Model Applied to Consumer-grade Wearable Devices,” Mar. 2022, [Online]. Available: http://arxiv.org/abs/2203.09663

[22] E. Zhou, M. Soleymani, and M. J. Matarić, “Investigating the Generalizability of Physiological Characteristics of Anxiety,” Jan. 2024, doi: 10.1109/BIBM58861.2023.10385292.

[23] *2015 International Conference on Advances in Biomedical Engineering (ICABME)*. IEEE, 2015.

[24] K. Nkurikiyeyezu, A. Yokokubo, and G. Lopez, “Effect of person-specific biometrics in improving generic stress predictive models,” 2020, *M Y U Scientific Publishing Division*. doi: 10.18494/SAM.2020.2650.

[25] B. Szakonyi, I. Vassányi, E. Schumacher, and I. Kósa, “Efficient methods for acute stress detection using heart rate variability data from Ambient Assisted Living sensors,” *Biomed Eng Online*, vol. 20, no. 1, Dec. 2021, doi: 10.1186/s12938-021-00911-6.

[26] P. Garg, J. Santhosh, A. Dengel, and S. Ishimaru, “Stress detection by machine learning and wearable sensors,” in *International Conference on Intelligent User Interfaces, Proceedings IUI*, Association for Computing Machinery, Apr. 2021, pp. 43–45. doi: 10.1145/3397482.3450732.

[27] S. Campanella, A. Altaleb, A. Belli, P. Pierleoni, and L. Palma, “A Method for Stress Detection Using Empatica E4 Bracelet and Machine-Learning Techniques,” *Sensors*, vol. 23, no. 7, Apr. 2023, doi: 10.3390/s23073565.

[28] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams, “Continuous stress detection using a wrist device - in laboratory and real life,” in *UbiComp 2016 Adjunct - Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Association for Computing Machinery, Inc, Sep. 2016, pp. 1185–1193. doi: 10.1145/2968219.2968306.

[29] A. Sano and R. W. Picard, “Stress recognition using wearable sensors and mobile phones,” in *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, 2013, pp. 671–676. doi: 10.1109/ACII.2013.117.

[30] J. A. Healey and R. W. Picard, “Detecting stress during real-world driving tasks using physiological sensors,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156–166, Jun. 2005, doi: 10.1109/TITS.2005.848368.

[31] J. Hernandez, R. R. Morris, and R. W. Picard, “Call center stress recognition with person-specific models,” in *Lecture Notes in Computer Science (including subseries*

Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2011, pp. 125–134. doi: 10.1007/978-3-642-24600-5_16.

[32] *Proceedings of the 10th ACM/IEEE International Conference on Information Processing in Sensor Networks : April 12-14, 2011, Chicago, IL, USA, IPSN '11*. Association for Computing Machinery, 2011.

[33] S. Koldijk, M. Sappelli, S. Verberne, M. A. Neerincx, and W. Kraaij, "The Swell knowledge work dataset for stress and user modeling research," in *ICMI 2014 - Proceedings of the 2014 International Conference on Multimodal Interaction*, Association for Computing Machinery, Inc, Nov. 2014, pp. 291–298. doi: 10.1145/2663204.2663257.

[34] N. El Haouij, J. M. Poggi, S. Sevestre-Ghalila, R. Ghozi, and M. Jadane, "AffectiveROAD system and database to assess driver's attention," in *Proceedings of the ACM Symposium on Applied Computing*, Association for Computing Machinery, Apr. 2018, pp. 800–803. doi: 10.1145/3167132.3167395.

[35] S. Katsigiannis and N. Ramzan, "DREAMER: A Database for Emotion Recognition Through EEG and ECG Signals from Wireless Low-cost Off-the-Shelf Devices," *IEEE J Biomed Health Inform*, vol. 22, no. 1, pp. 98–107, Jan. 2018, doi: 10.1109/JBHI.2017.2688239.

[36] S. Koelstra *et al.*, "DEAP: A database for emotion analysis; Using physiological signals," *IEEE Trans Affect Comput*, vol. 3, no. 1, pp. 18–31, Jan. 2012, doi: 10.1109/T-AFFC.2011.15.

[37] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras, "AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups," *IEEE Trans Affect Comput*, vol. 12, no. 2, pp. 479–493, Apr. 2021, doi: 10.1109/TAFFC.2018.2884461.

[38] J. Chan, A. C. Miller, and E. B. Fox, "Representing and Denoising Wearable ECG Recordings," Nov. 2020, [Online]. Available: http://arxiv.org/abs/2012.00110

[39] A. Nieto-Castanon, "Preparing fMRI Data for Statistical Analysis."

[40] A. Iranfar, A. Arza, and D. Atienza, "ReLearn: A Robust Machine Learning Framework in Presence of Missing Data for Multimodal Stress Detection from Physiological Signals," Apr. 2021, [Online]. Available: http://arxiv.org/abs/2104.14278

[41] W. Boucsein *et al.*, "Publication recommendations for electrodermal measurements," *Psychophysiology*, vol. 49, no. 8, pp. 1017–1034, 2012, doi: 10.1111/j.1469-8986.2012.01384.x.

[42] A. Scippa, L. Sallese, N. Grossi, and G. Campatelli, “Improved dynamic compensation for accurate cutting force measurements in milling applications,” *Mech Syst Signal Process*, vol. 54, pp. 314–324, Mar. 2015, doi: 10.1016/j.ymssp.2014.08.019.

[43] N. V Thakor, Y.-S. Zhu, and S. Momher, “Applications of Adaptive Filtering to ECG Analysis : Noise Cancellation and Arrhythmia Detection.”

[44] K. C. Liu, C. Y. Hsieh, H. Y. Huang, S. J. P. Hsu, and C. T. Chan, “An Analysis of Segmentation Approaches and Window Sizes in Wearable-Based Critical Fall Detection Systems with Machine Learning Models,” *IEEE Sens J*, vol. 20, no. 6, pp. 3303–3313, Mar. 2020, doi: 10.1109/JSEN.2019.2955141.

[45] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, *Activity Recognition using Cell Phone Accelerometers*. 2010.

[46] A. K. Bourke and G. M. Lyons, “A threshold-based fall-detection algorithm using a bi-axial gyroscope sensor,” *Med Eng Phys*, vol. 30, no. 1, pp. 84–90, Jan. 2008, doi: 10.1016/j.medengphy.2006.12.001.

[47] Y. Nawal, M. Oussalah, B. Fergani, and A. Fleury, “New incremental SVM algorithms for human activity recognition in smart homes,” *J Ambient Intell Humaniz Comput*, vol. 14, no. 10, pp. 13433–13450, Oct. 2023, doi: 10.1007/s12652-022-03798-w.

[48] S. Chambon, V. Thorey, P. J. Arnal, E. Mignot, and A. Gramfort, “A deep learning architecture to detect events in EEG signals during sleep,” Jul. 2018, [Online]. Available: http://arxiv.org/abs/1807.05981

[49] X. Zhou, F. Ghorbani, V. Roessner, B. Hommel, A. Prochnow, and C. Beste, “The metacontrol of event segmentation—A neurophysiological and behavioral perspective,” *Hum Brain Mapp*, vol. 45, no. 11, Aug. 2024, doi: 10.1002/hbm.26727.

[50] C. Joung *et al.*, “Deep learning based ECG segmentation for delineation of diverse arrhythmias,” *PLoS One*, vol. 19, no. 6 June, Jun. 2024, doi: 10.1371/journal.pone.0303178.

[51] R. Redina, J. Hejc, M. Filipenska, and Z. Starek, “Analyzing the performance of biomedical time-series segmentation with electrophysiology data,” *Sci Rep*, vol. 15, no. 1, Dec. 2025, doi: 10.1038/s41598-025-90533-y.

[52] A. Dehghani, O. Sarbishei, T. Glatard, and E. Shihab, “A quantitative comparison of overlapping and non-overlapping sliding windows for human activity recognition using inertial sensors,” *Sensors (Switzerland)*, vol. 19, no. 22, Nov. 2019, doi: 10.3390/s19225026.

[53] W. Al-Salman, Y. Li, and P. Wen, “Detection of EEG K-complexes using fractal dimension of time frequency images technique coupled with undirected graph features,” *Front Neuroinform*, vol. 13, May 2019, doi: 10.3389/fninf.2019.00045.

[54] A. Waheed Awan, I. Taj, S. Khalid, S. Muhammad Usman, A. Shariq Imran, and M. Usman Akram, “Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000. Advancing Emotional Health Assessments: A Hybrid Deep Learning Approach Using Physiological Signals for Robust Emotion Recognition”, doi: 10.1109/ACCESS.2017.DOI.

[55] L. Wang, “Computational reduction for noninvasive transmural electrophysiological imaging,” *Comput Biol Med*, vol. 43, no. 3, pp. 184–199, Mar. 2013, doi: 10.1016/j.compbiomed.2012.12.003.

[56] I. Stancin, M. Cifrek, and A. Jovic, “A review of eeg signal features and their application in driver drowsiness detection systems,” Jun. 01, 2021, *MDPI AG*. doi: 10.3390/s21113786.

[57] Saeid. Sanei and Jonathon. Chambers, *EEG signal processing*. John Wiley & Sons, 2009.

[58] L. Shu *et al.*, “A review of emotion recognition using physiological signals,” Jul. 01, 2018, *MDPI AG*. doi: 10.3390/s18072074.

[59] A. K. Singh and S. Krishnan, “ECG signal feature extraction trends in methods and applications,” Dec. 01, 2023, *BioMed Central Ltd*. doi: 10.1186/s12938-023-01075-1.

[60] A. K. Kirby, S. Pancholi, Z. Anderson, C. Chesler, T. H. Everett, and B. S. Duerstock, “Time and frequency domain analysis of physiological features during autonomic dysreflexia after spinal cord injury,” *Front Neurosci*, vol. 17, 2023, doi: 10.3389/fnins.2023.1210815.

[61] “Probability and Statistics for Engineering and the Sciences 9th Edition pdf”.

[62] T. Chai, “Root Mean Square,” 2022, pp. 1–3. doi: 10.1007/978-3-030-26050-7_280-1.

[63] D. Iacobucci, S. Román, S. Moon, and D. Rouziès, “A Tutorial on What to Do With Skewness, Kurtosis, and Outliers: New Insights to Help Scholars Conduct and Defend Their Research,” *Psychol Mark*, 2025, doi: 10.1002/mar.22187.

[64] “Digital Signal Processing Second Edition.” [Online]. Available: www.DSPguide.com

[65] H. F. Posada-Quintero, J. P. Florian, A. D. Orjuela-Cañón, T. Aljama-Corrales, S. Charleston-Villalobos, and K. H. Chon, “Power Spectral Density Analysis of

Electrodermal Activity for Sympathetic Function Assessment," *Ann Biomed Eng*, vol. 44, no. 10, pp. 3124–3135, Oct. 2016, doi: 10.1007/s10439-016-1606-6.

[66] A. R. Kiselev, O. M. Posnenkova, A. S. Karavaev, V. A. Shvartz, M. Yu. Novikov, and V. I. Gridnev, "Frequency-Domain Features and Low-Frequency Synchronization of Photoplethysmographic Waveform Variability and Heart Rate Variability with Increasing Severity of Cardiovascular Diseases," *Biomedicines*, vol. 12, no. 9, p. 2088, Sep. 2024, doi: 10.3390/biomedicines12092088.

[67] "Entropy and Information Theory."

[68] M. M. Kansal, "A REVIEW OF RESEARCH PAPERS ON FOURIER TRANSFORMS & STATISTICAL FOURIER ANALYSIS," 2018. [Online]. Available: www.jetir.org

[69] S. Shenbaga Ezhil, "Real time application of Fourier transforms," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 8, no. 2, pp. 574–577, Nov. 2017, doi: 10.11591/ijeecs.v8.i2.pp574-577.

[70] C. Gantt and H. Yue, "The Fourier Transform and Signal Processing."

[71] N. Ji, L. Ma, H. Dong, and X. Zhang, "EEG signals feature extraction based on DWT and EMD combined with approximate entropy," *Brain Sci*, vol. 9, no. 8, Aug. 2019, doi: 10.3390/brainsci9080201.

[72] H. K. Lee and Y. S. Choi, "Application of continuous wavelet transform and convolutional neural network in decoding motor imagery brain-computer Interface," *Entropy*, vol. 21, no. 12, Dec. 2019, doi: 10.3390/e21121199.

[73] I. Krak and O. Khorozov, "Physiological Signals Analysis, Recognition and Classification Using Machine Learning Algorithms," 2020.

[74] · Book, "A Wavelet Tour of Signal Processing," 1999. [Online]. Available: https://www.researchgate.net/publication/215659315

[75] J. Rafiee and P. W. Tse, "Use of autocorrelation of wavelet coefficients for fault diagnosis," *Mech Syst Signal Process*, vol. 23, no. 5, pp. 1554–1572, Jul. 2009, doi: 10.1016/j.ymssp.2009.02.008.

[76] T. Takami, U. D. Nielsen, and J. J. Jensen, "Estimation of autocorrelation function and spectrum density of wave-induced responses using prolate spheroidal wave functions," *Journal of Marine Science and Technology (Japan)*, vol. 26, no. 3, pp. 772–791, Sep. 2021, doi: 10.1007/s00773-020-00768-9.

[77] . José S. Murguía and . Haret C. Rosu, *Discrete Wavelet Analyses for Time Series*. INTECH Open Access Publisher, 2011.

[78] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction," 2022, *Frontiers Media SA*. doi: 10.3389/fbinf.2022.927312.

[79] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers and Electrical Engineering*, vol. 40, no. 1, pp. 16–28, Jan. 2014, doi: 10.1016/j.compeleceng.2013.11.024.

[80] *Rule-Based Evolutionary Online Learning Systems: A Principled Approach to LCS Analysis and Design*. Scholars Portal, 2019.

[81] I. Guyon and A. M. De, "An Introduction to Variable and Feature Selection André Elisseeff," 2003.

[82] R. Jain and W. Xu, "Artificial Intelligence based wrapper for high dimensional feature selection," *BMC Bioinformatics*, vol. 24, no. 1, Dec. 2023, doi: 10.1186/s12859-023-05502-x.

[83] A. Hassan, J. H. Paik, S. Khare, and S. A. Hassan, "PPFS: Predictive Permutation Feature Selection," Oct. 2021, doi: 10.1016/j.patcog.2024.111069.

[84] J. Li, H. Zhang, J. Zhao, X. Guo, W. Rihan, and G. Deng, "Embedded Feature Selection and Machine Learning Methods for Flash Flood Susceptibility-Mapping in the Mainstream Songhua River Basin, China," *Remote Sens (Basel)*, vol. 14, no. 21, Nov. 2022, doi: 10.3390/rs14215523.

[85] M. J. JimÉnez-Navarro, M. MartÍnez-Ballesteros, I. S. Brito, F. MartÍnez-Álvarez, and G. Asencio-CortÉs, "Embedded feature selection for neural networks via learnable drop layer," *Log J IGPL*, Jun. 2024, doi: 10.1093/jigpal/jzae062.

[86] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," 2005.

[87] H. Deng and G. Runger, "Feature Selection via Regularized Trees," Jan. 2012, [Online]. Available: http://arxiv.org/abs/1201.1587

[88] J. A. Mendivil Sauceda, B. Y. Marquez, and J. J. Esqueda Elizondo, "Emotion Classification from Electroencephalographic Signals Using Machine Learning," *Brain Sci*, vol. 14, no. 12, Dec. 2024, doi: 10.3390/brainsci14121211.

[89] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, “Machine learning: A review of classification and combining techniques,” *Artif Intell Rev*, vol. 26, no. 3, pp. 159–190, Nov. 2006, doi: 10.1007/s10462-007-9052-3.

[90] S. B. Kotsiantis, “Supervised Machine Learning: A Review of Classification Techniques,” 2007.

[91] T. Jiang, J. L. Gradus, and A. J. Rosellini, “Supervised Machine Learning: A Brief Primer,” *Behav Ther*, vol. 51, no. 5, pp. 675–687, Sep. 2020, doi: 10.1016/j.beth.2020.05.002.

[92] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, “Comparing different supervised machine learning algorithms for disease prediction,” *BMC Med Inform Decis Mak*, vol. 19, no. 1, Dec. 2019, doi: 10.1186/s12911-019-1004-8.

[93] O. N. Manjrekar and M. P. Dudukovic, “Identification of flow regime in a bubble column reactor with a combination of optical probe data and machine learning technique,” *Chemical Engineering Science: X*, vol. 2, May 2019, doi: 10.1016/j.cesx.2019.100023.

[94] S. S. Olofintuyi, “BREAST CANCER DETECTION WITH MACHINE LEARNING APPROACH,” *FUDMA JOURNAL OF SCIENCES*, vol. 7, no. 2, pp. 216–222, Apr. 2023, doi: 10.33003/fjs-2023-0702-1392.

[95] S. I. Dimitriadis and D. Liparas, “How random is the random forest? Random forest algorithm on the service of structural imaging biomarkers for Alzheimer’s disease: From Alzheimer’s disease neuroimaging initiative (ADNI) database,” Jun. 01, 2018, *Wolters Kluwer Medknow Publications*. doi: 10.4103/1673-5374.233433.

[96] S. Nosratabadi *et al.*, “Data Science in Economics: Comprehensive Review of Advanced Machine Learning and Deep Learning Methods,” Oct. 13, 2020. doi: 10.20944/preprints202010.0263.v1.

[97] T. Zhang *et al.*, “Improving Convection Trigger Functions in Deep Convective Parameterization Schemes Using Machine Learning,” *J Adv Model Earth Syst*, vol. 13, no. 5, May 2021, doi: 10.1029/2020MS002365.

[98] Dewi Widyawati and Amaliah Faradibah, “Comparison Analysis of Classification Model Performance in Lung Cancer Prediction Using Decision Tree, Naive Bayes, and Support Vector Machine,” *Indonesian Journal of Data and Science*, vol. 4, no. 2, pp. 80–89, Jul. 2023, doi: 10.56705/ijodas.v4i2.76.

[99] Z. N. Sultani and S. A. Yousif, “Hybrid feature selection based on mutual information and auc for parkinson’s disease classification,” *Article in Journal of Theoretical and*

Applied Information Technology, vol. 96, no. 18, 2018, [Online]. Available: https://www.researchgate.net/publication/328146770

[100] D. Kumar Ghosh, "Perspective Chapter: Linear Regression and Logistic Regression Models," in *Recent Advances in Biostatistics*, IntechOpen, 2024. doi: 10.5772/intechopen.1003183.

[101] S. Naeem, A. Ali, S. Anam, and M. M. Ahmed, "An Unsupervised Machine Learning Algorithms: Comprehensive Review," *International Journal of Computing and Digital Systems*, vol. 13, no. 1, pp. 911–921, 2023, doi: 10.12785/ijcds/130172.

[102] *ICIEA 2018 : 2018 5th International Conference on Industrial Engineering and Applications : April 26-28, 2018, Singapore*. IEEE Press, 2018.

[103] J. G. Dy and C. E. Brodley, "Feature Selection for Unsupervised Learning," 2004.

[104] C. M. Eckhardt *et al.*, "Unsupervised machine learning methods and emerging applications in healthcare," Feb. 01, 2023, *Springer Science and Business Media Deutschland GmbH*. doi: 10.1007/s00167-022-07233-7.

[105] E. Maida *et al.*, "Identifying definite patterns of unmet needs in patients with multiple sclerosis using unsupervised machine learning," *Neurological Sciences*, vol. 45, no. 7, pp. 3333–3345, Jul. 2024, doi: 10.1007/s10072-024-07416-9.

[106] V. Mohan, S. Vijayarani, M. S. M. Sylviaa, and M. P. Scholar, "DIMENSIONALITY REDUCTION-A STUDY," 2016. [Online]. Available: http://www.ijeast.com

[107] W. Jia, M. Sun, J. Lian, and S. Hou, "Feature dimensionality reduction: a review," *Complex and Intelligent Systems*, vol. 8, no. 3, pp. 2663–2693, Jun. 2022, doi: 10.1007/s40747-021-00637-x.

[108] F. L. Gewers *et al.*, "Principal Component Analysis: A Natural Approach to Data Exploration," Apr. 2018, doi: 10.1145/3447755.

[109] D. De, T. Nayak, S. Chowdhury, and P. K. Dhal, "Insights of Host Physiological Parameters and Gut Microbiome of Indian Type 2 Diabetic Patients Visualized via Metagenomics and Machine Learning Approaches," *Front Microbiol*, vol. 13, Jul. 2022, doi: 10.3389/fmicb.2022.914124.

[110] T. T. Cai and R. Ma, "Theoretical Foundations of t-SNE for Visualizing High-Dimensional Clustered Data," May 2021, [Online]. Available: http://arxiv.org/abs/2105.07536

[111] L. Van Der Maaten and G. Hinton, "Visualizing Data using t-SNE," 2008.

[112] D. Zhang *et al.*, “Supporting Clustering with Contrastive Learning,” Mar. 2021, [Online]. Available: http://arxiv.org/abs/2103.12953

[113] S. Raschka, “Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning,” Nov. 2018, [Online]. Available: http://arxiv.org/abs/1811.12808

[114] C. Miller, T. Portlock, D. M. Nyaga, and J. M. O’Sullivan, “A review of model evaluation metrics for machine learning in genetics and genomics,” *Frontiers in Bioinformatics*, vol. 4, Sep. 2024, doi: 10.3389/fbinf.2024.1457619.

[115] L. A. Yates, Z. Aandahl, S. A. Richards, and B. W. Brook, “Cross validation for model selection: A review with examples from ecology,” *Ecol Monogr*, vol. 93, no. 1, Feb. 2023, doi: 10.1002/ecm.1557.

[116] F. H. Fadel and S. F. Behadili, “A Comparative Study for Supervised Learning Algorithms to Analyze Sentiment Tweets,” *Iraqi Journal of Science*, vol. 63, no. 6, pp. 2712–2724, 2022, doi: 10.24996/ijs.2022.63.6.36.

[117] A. Geroldinger, L. Lusa, M. Nold, and G. Heinze, “Leave-one-out cross-validation, penalization, and differential bias of some prediction model performance measures—a simulation study,” *Diagn Progn Res*, vol. 7, no. 1, May 2023, doi: 10.1186/s41512-023-00146-0.

[118] D. Gholamiangonabadi, N. Kiselov, and K. Grolinger, “Deep Neural Networks for Human Activity Recognition with Wearable Sensors: Leave-One-Subject-Out Cross-Validation for Model Selection,” *IEEE Access*, vol. 8, pp. 133982–133994, 2020, doi: 10.1109/ACCESS.2020.3010715.

[119] O. Rainio, J. Teuho, and R. Klén, “Evaluation metrics and statistical tests for machine learning,” *Sci Rep*, vol. 14, no. 1, Dec. 2024, doi: 10.1038/s41598-024-56706-x.

[120] J. Opitz, “A Closer Look at Classification Evaluation Metrics and a Critical Reflection of Common Evaluation Practice”, doi: 10.1162/tacl.

[121] Y. Badr, U. Tariq, F. Al-Shargie, F. Babiloni, F. Al Mughairbi, and H. Al-Nashash, “A review on evaluating mental stress by deep learning using EEG signals,” Jul. 01, 2024, *Springer Science and Business Media Deutschland GmbH*. doi: 10.1007/s00521-024-09809-5.

[122] S. Pichandi, G. Balasubramanian, and V. Chakrapani, “Hybrid deep models for parallel feature extraction and enhanced emotion state classification,” *Sci Rep*, vol. 14, no. 1, p. 24957, Dec. 2024, doi: 10.1038/s41598-024-75850-y.

[123] Y. Acikmese and S. E. Alptekin, “Prediction of stress levels with LSTM and passive mobile sensors,” in *Procedia Computer Science*, Elsevier B.V., 2019, pp. 658–667. doi: 10.1016/j.procs.2019.09.221.

[124] M. Jain and A. Srihari, “Comparison of Machine Learning Models for Stress Detection from Sensor Data Using Long Short-Term Memory (LSTM) Networks and Convolutional Neural Networks (CNNs),” *International Journal of Scientific Research and Management (IJSRM)*, vol. 12, no. 12, pp. 1775–1792, Dec. 2024, doi: 10.18535/ijsrm/v12i12.ec02.