NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF PHYSICS

INÊS ALEXANDRA VIEIRA BOTELHO

BSc in Biomedical Engineering Sciences

EMOTION RECOGNITION USING MULTIMODAL TIME SERIES

DISSERTATION ON BIOMEDICAL ENGINEERING

MASTER IN BIOMEDICAL ENGINEERING

NOVA University Lisbon
October, 2023

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTMENT OF PHYSICS

EMOTION RECOGNITION USING MULTIMODAL TIME SERIES

DISSERTATION ON BIOMEDICAL ENGINEERING

INÊS ALEXANDRA VIEIRA BOTELHO

BSc in Biomedical Engineering Sciences

Adviser: Hugo Gamboa

Associate Professor, NOVA School of Science and Technology

MASTER IN BIOMEDICAL ENGINEERING

NOVA University Lisbon
October, 2023

Emotion Recognition using Multimodal Time Series
Dissertation on Biomedical Engineering

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v6.10.17) [1].

Acknowledgements

Firstly, I express my heartfelt gratitude to Professor Hugo Gamboa for providing me with the invaluable opportunity to conduct my thesis at the esteemed Fraunhofer Institute. This experience allowed me to collaborate with like-minded individuals who share my passion for this field. I am especially thankful to Pedro Matias, whose unwavering support, valuable insights, and willingness to assist have been pivotal in the successful completion of my work.

I would also like to extend my appreciation to Nova School of Science and Technology and the dedicated faculty members who have guided me throughout this academic journey. Their influence has greatly contributed to the enriching years of my education.

Last but certainly not least, my profound gratitude goes to my family for their unwavering belief in my abilities and for making it possible for me to pursue a college degree. My parents have always stood by my side, understanding my need to work from home. I would also like to express my thanks to Inês, who has been my constant companion and a pillar of support. Her unwavering confidence and strength have been a constant source of inspiration.

I am also grateful for the friendships I have cultivated along this journey and for my fellow members at AEFCT and NuPride, who have made these five years truly special by allowing me to explore new horizons and experiences.

Abstract

Emotions play an important role in Human-Computer Interaction (HCI), specially in healthcare, as they can provide psychological feedback on patients' status. Wearable sensor data such as electrodermal activity (EDA), electrocardiography (ECG), and Voice activity may enhance emotion recognition, but their integration remains a challenge. This work aimed to address this by exploring cross-modal synergies.

In pursuit of such goals, data collections have been conducted, encompassing recordings from these three modalities, during a sequential protocol. Diverse stimulus types were employed covering four quadrants of the 2D Valence-Arousal spectrum. Signals were pre-processed with state-of-the-art methodologies and meaningful feature were extracted. Optimization was extensively conducted using Machine Learning techniques, as well as preliminary Deep Learning experiments. Multimodal fusion approaches were evaluated through (1) early and (2) late fusion techniques and validated in the (a) collected (public) and (b) external (private) datasets.

In unimodal analyses, the Voice signal showed higher performance on public datasets (simulated emotions) compared to our protocol. Models averaged 69% (public) and 51% (private) balanced accuracy (BA). Physiological signals had similar performance, 47% (public) and 44% (private) BA for EDA, and 48% (public) and 52% (private) BA for ECG. Combining data modalities using late fusion consistently outperformed unimodal strategies, achieving 57% (public) and 60% (private) BA in a 4-class classification problem. Annotator agreement significantly affected emotion detection, as models trained on consistent labels performed better. For stimuli-wise, audiovisual, and acting tasks triggered better model performance.

Emotion recognition remains challenging due to inherent uncertainty. Combining modalities enhances model performance, emphasizing the need for reliable assessment tools in healthcare scenarios.

Keywords: Emotion Recognition, Physiological Signals, Voice Activity, Multimodal Data Fusion, Machine Learning

Resumo

As emoções desempenham um papel fundamental na *Human Computer Interaction (HCI)*, especialmente na área da saúde, pois são capazes de fornecer *feedback* psicológico do estado dos pacientes. Dados de sensores *wearable* como *electrodermal activity (EDA)*, *electrocardiography (ECG)* e atividade vocal podem melhorar o reconhecimento de emoções, mas a sua junção permanece um desafio. Este trabalho visa abordar esse desafio explorando sinergias entre diferentes modalidades.

Desta forma, foram realizadas recolhas destes dados, seguindo um protocolo sequencial. Foram usados vários tipos de estímulos que incluem os quatro quadrantes do espetro *2D Valence-Arousal*. Os sinais foram pré-processados com metodologias do estado da arte e *features* relevantes foram extraídas. Foi realizada uma extensa otimização usando técnicas de *Machine Learning*, bem como experiências preliminares de *Deep Learning*. Abordagens de fusão multimodal foram avaliadas por meio de técnicas de fusão (1) precoce e (2) tardia e validadas em *datasets* (a) coletados (públicos) (b) e externos (privados).

Nas análises unimodais, o sinal de Voz mostrou um desempenho superior nos *datasets* públicos (emulações de emoções) em comparação com o protocolo proposto. Os modelos alcançaram uma média de 69% (público) e 51% (privado) de *balanced accuracy (BA)*. Os sinais fisiológicos tiveram desempenho semelhante, de 47% (público) e 44% (privado) de BA para o EDA e 48% (público) e 52% (privado) de BA para o ECG. A combinação de modalidades de dados usando a fusão tardia superou consistentemente, alcançando 57% (público) e 60% (privado) de BA em um problema de classificação de quatro classes. O acordo entre os anotadores afetou significativamente a deteção de emoções. Modelos treinados com *labels* consistentes tiveram um desempenho melhor. Quanto aos estímulos, tarefas audiovisuais e de atuação provocaram um melhor desempenho dos modelos.

O reconhecimento de emoções ainda é um desafio devido à incerteza inerente. A combinação de modalidades aprimora o desempenho do modelo, destacando a necessidade de ferramentas de avaliação confiáveis em cenários de saúde.

Palavras-chave: Reconhecimento de Emoções, Sinais Fisiológicos, Atividade Vocal, Fusão de Dados Multimodais, Aprendizado de Máquina

Contents

List of Figures

List of Tables

Acronyms

AB	Ada Boost
AI	Artificial Intelligence
AMIGOS	A Dataset for Affect, Personality and Mood Research on Individuals and Groups
AUC	Area Under the ROC Curve
BA	Balanced Accuracy
BN	Baseline Normalization
BPM	Beats Per Minute
CNN	Convolutional Neural Network
CPP	Cepstral Peak Prominence
dB	Decibel
DL	Deep Learning
DT	Decision Tree
ECA	Embodied Conversational Agents
ECG	Electrocardiography
EDA	Electrodermal Activity
EMG	Electromyography
EMOUERJ	An Emotional Speech Database In Portuguese
EPEDD	European Portuguese Emotional Discourse Database
F0	Fundamental Frequency
FCN	Full Column Normalization
FN	False Negative
FP	False Positive
FT	Fourier Transform

HCI Human-Computer Interaction
HF High Frequency
HNR Harmonic to Noise Ratio
HR Heart Rate
HRV Heart Rate Variability
HVHA High Valence High Arousal
HVLA High Valence Low Arousal

IBI Inter-Beat Interval
ICANS Intelligent Customer and Advanced Natural Systems
IDE Integrated Development Environment
IQR Interquartile

KNN K-Nearest Neighbour

LDA Linear Discriminant Analysis
LF Low Frequency
LGBM Light GBM
LPCC Linear Prediction Cepstral Coefficients
LSTM Long Short-Term Memory
LVHA Low Valence High Arousal
LVLA Low Valence Low Arousal

M MinMax Scaling
MAD Mean Absolute Deviation
MFCC Mel Frequency Cepstral Coefficients
ML Machine Learning

NB Naive Bayes
NLP Natural Language Processing

PCA Principal Component Analysis
PPG Photoplethysmography

QDA Quadratic Discriminant Analysis

R Robust Scaling
RF Random Forest
RFE Recursive Feature Elimination

RMS Root Mean Square
RN Rest Normalization
ROC Receiver Operating Characteristic Curve
RRPL RR Poincaré Length
RRPW RR Poincaré Width
RSA Respiratory Sinus Arrhythmia

S Standard Scaling
SBN Subject-Based Normalization
SCL Skin Conductance Level
SCR Skin Conductance Response
SD Successive Differences of NN Intervals
ST Skin Temperature
STD Standard Deviation
SVM Support Vector Machine

t-SNE t-distributed Stochastic Neighbor Embedding
TN True Negative
TP True Positive

UMAP Uniform Manifold Approximation and Projection

VLF Very Low Frequency

YAAD Young Adult's Affective Data

ZCR Zero Crossing Rate

1

Introduction

This chapter describes the motivation behind this work, delineate its ultimate goals, and conclude with an overview of the document's structure.

1.1 Motivation

Emotions are complex phenomena that involve both psychological and physiological responses to stimuli, whether they are consciously perceived or not [2]. In Human-Computer Interaction (HCI), emotions are an essential aspect to consider, as they provide valuable feedback on whether users' needs and expectations are being met and allow for dynamic and intelligent adaptation of interfaces or services [3]. HCI is a multidisciplinary field that focuses on creating software that is user-friendly and accessible to everyone, by studying the relationship between people and computers and finding ways to make their interaction easy and enjoyable [4].

In recent times, the healthcare industry field has undergone reasonable technological advancements, opening the floor to decentralized treatments using wearable sensors [5]. This progress has expanded the potential applications of assessing human emotions, particularly in scenarios where understanding emotional states is advantageous, for example, in psychological assessments [6].

Nonetheless, existing work on emotion recognition has not yet fully exploited the potential of combining data from multiple sources, such as Electrocardiography (ECG), Electrodermal Activity (EDA), and Voice Activity, in order to improve the reliability of end-to-end systems [7]. Multiple studies were found using distinct data modalities but bringing them together might bridge the gap to achieve a robust system. Therefore, this work aims to investigate the benefits of multimodal data fusion for emotion recognition, and to explore different techniques for combining and analyzing data from multiple sensors.

1.2 Main Goals

The main objectives of this work are outlined below:

- Develop tools for emotion recognition using physiological and Voice activity data;
- Investigate the complementary nature of ECG, EDA, and Voice data. Demonstrate the benefits of integrating them heading a more reliable and accurate emotion recognition assessment;
- Demonstrate that simple, cost-effective Machine Learning (ML) techniques can provide trustworthy solutions to address complex tasks such as the presented;
- Study the influence of low-intensity stimuli on an emotion recognition task.

1.3 Outline

This thesis comprises five main chapters, alongside annexes and appendices cited when necessary. In chapter 2, fundamental concepts are introduced as a basis for the subsequent work. This chapter covers the significance of physiological signals (EDA, ECG, and Voice), their signal processing characteristics, and their roles in multimodal emotion recognition. It also explores Artificial Intelligence (AI) topics, including ML and Deep Learning (DL) algorithms, types of normalization, data fusion techniques, dimensionality reduction methods, and evaluation metrics. The discussion extends to emotions in HCI and the 2D Valence-Arousal model. The next chapter, chapter 3 has an overview of the current state of the art. Next, chapter 4 addresses data sources, processing steps, model development, and explains the emotion recognition task, covering both unimodal and multimodal approaches. In chapter 5, results and corresponding discussions are presented, including insights from a low intensity scenario, unimodal and multimodal outcomes for EDA, ECG, and Voice signals, and other relevant results. Finally, chapter 6 summarizes the research's core findings and outlines potential directions for future research.

The research work described in this dissertation was carried out in accordance with the norms established in the ethics code of Universidade Nova de Lisboa. The work described and the material presented in this dissertation, with the exceptions clearly indicated, constitute original work carried out by the author

2

Theoretical Background

This chapter provides a theoretical background of the topics under study. Theory behind the EDA, ECG, and Voice signals, as well as ML and DL background, and the role of emotions in the HCI field, are some of the topics covered.

2.1 Physiological Signals

2.1.1 Electrodermal Activity

The Autonomic Nervous System regulates the primary function of the sweat glands, which is the regulation of temperature. Additionally, some sweat glands, especially those located at the palms and soles of the feet, are also linked to psychological stimuli and evoked emotions at an arousal level [8]. Sweat is composed mostly of water but also contains minerals, lactic acid, and urea, making it a good conductor of electricity [8].

This signal is comprised of two distinct components: a tonic component (Skin Conductance Level (SCL)), ranging within [0, 0.16] Hz interval (lower frequency) [9], and a phasic component (Skin Conductance Response (SCR)), ranging within [0.16, 2.1] Hz interval (higher frequency) [9]. Figure 2.1 illustrates an SCR pulse, whose characteristics are described as follows:

- **Rise Time:** time interval between the onset of the SCR pulse and its peak [10];
- **Latency:** interval between the stimuli and the response onset [10];
- **Amplitude:** amplitude difference between the onset and the SCR pulse peak [10];
- **Decay time:** time between the peak of the SCR pulse and its half recovery [10].

Before performing any analysis of an EDA signal, it is required to do some pre-processing steps, mostly to separate both signal components and reduce the high-frequency noise effects. This way, some common pre-processing steps are described as follows:

- **Downsampling:** it optimizes memory usage and data processing without losing meaningful information in the signal [12];

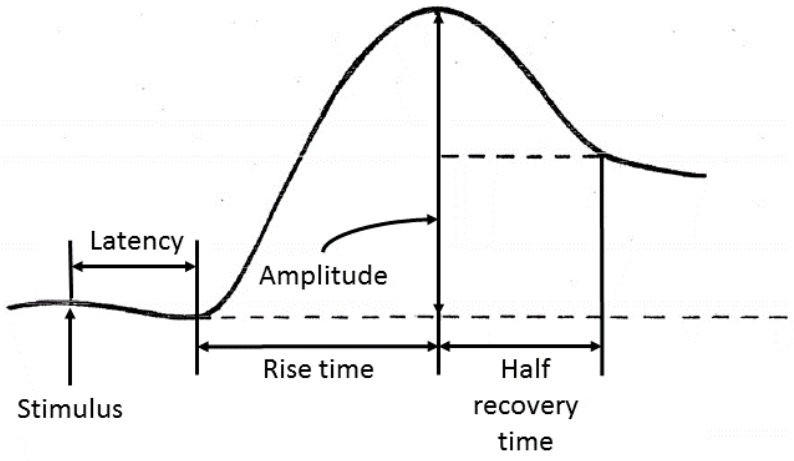


Figure 2.1: Illustration of a typical SCR pulse morphology, along with annotations indicating its main morphological characteristics [11].

- **Low-Pass Filter:** identical to a moving average, it smooths the signals reducing high-frequency noise levels. Cut-off usually settled at 5 Hz [12];
- **Decomposition:** detaches the signal into tonic (trend) and phasic (pulses) components using filtering, bayesian [13], or deconvolution [14] methods;
- **Normalization:** it standardizes the signal amplitude to get comparable amongst different people and devices.

Upon signal cleaning, some features are ready to be extracted, depending on its domain:

- **Amplitude:** describes the response intensity to stimuli. Descriptive statistics over SCR pulses or SCL trend may apply [12];
- **Time:** captures temporal relationships from SCR pulses (e.g., latency, rise/decay times) [12].

2.1.2 Electrocardiography

The human cardiac muscle, also known as myocardium, requires a source of energy and oxygen to function properly [15]. Its pumping action is regulated by an electrical conduction system that coordinates the contraction of its different chambers. This coordination is necessary to ensure a proper functioning [16]. So, the ECG signal results from capturing the variations of the myocardium's electrical depolarization over time.

Next, there are described the ECG waves and their associated characteristics:

- **P Wave:** represents atrial depolarization [17] and has a frequency range of [8, 10] Hz [18];

- **QRS Complex:** represents ventricular depolarization in three different phases. First, the Q wave represents the depolarization of the interventricular septum, then the R wave the depolarization of the main mass of the ventricles and finally the S wave, the depolarization of the base [17]. The frequency range of this complex is [10, 40] Hz [18]. The time interval between two successive R waves makes a beat, typically denoted as RR or NN to emphasize that the heartbeats are normal and consistent;

- **T Wave:** represents the ventricular repolarization [17] and has a frequency range of [5, 8] Hz [18].

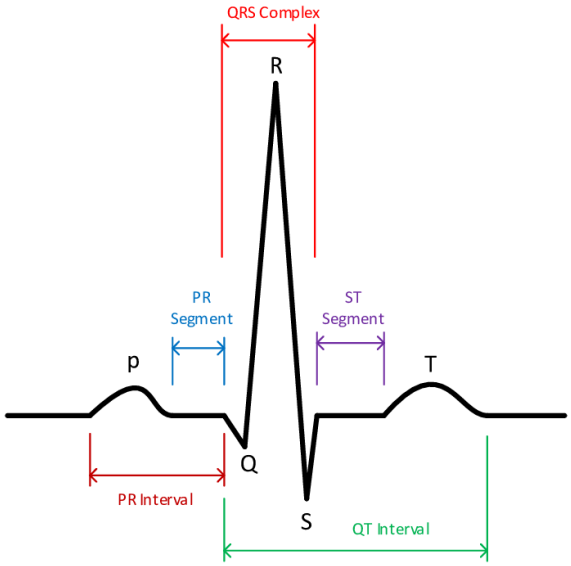


Figure 2.2: Illustration of the morphology of a standard ECG heartbeat (lead II). Intervals and segments terminology are also denoted [19].

In terms of the signal pre-processing, several cleaning strategies apply. Many techniques focus on preserving the frequency range of the signal and eliminating what's not included in that interval, typically the noise. Although the highest wave can reach 40Hz (QRS complex), some researchers keep a larger frequency range to avoid losing the waveform detail [20]. Different levels of detail should be applied depending on the task at handle [20]. Some examples of pre-processing techniques are described below:

- **Downsampling:** it optimizes memory usage and data processing without losing meaningful information in the signal [12];

- **Bandpass Filter:** the combination of a lowpass, to remove high-frequency content (e.g., noise), and a highpass filter, to attenuate low-frequency content (e.g., baseline wander). A common cut-off bandwidth is placed within [0.1, 100] Hz interval [21];

- **Notch Filter:** it attenuates the effect of a particular frequency. Usually used to remove the powerline frequency content (50/60 Hz) [21];

- **Normalization:** used to standardize the signal's amplitude amongst different users and/or devices.

The extraction of ECG features may differ depending on the domains where they are computed (amplitude, temporal, spectral). Each domain offers unique information about the signal, for example:

- **Temporal:** important domain for capturing periodicity-related content from the signal, parameters as the time interval between RR intervals (IBI) and their successive differences (SD) are analyzed. These metrics enable the extraction of important features related to Heart Rate (HR)/Beats Per Minute (BPM) and Heart Rate Variability (HRV) measures [21]. Additionally, it is also possible to extract features, such as the width and length in a Poincaré plot (RRPW and RRPL), by considering the RR intervals [22];

- **Spectral:** by computing the Fourier Transform (FT), the frequency spectrum provides information about the most prominent frequency bands, from which some features can be calculated, such as the centroid, roll-off, and the density of very low (VLF), low (LF), and high-frequency (HF) spectral bands [23].

2.1.3 Voice Activity

A Voice signal, illustrated in Figure 2.3, typically recorded as an audio signal, is a combination of periodical, evenly-shaped waveforms with different frequencies. The Fundamental Frequency (F0) of the vocal fold depends on many factors but the normal frequency range stands within [175, 245] Hz for a female adult and within [105, 160] Hz for a male adult [24]. Although the frequency of speech ranges within [200, 10000] Hz, most of the energy is focused within [200, 3500] Hz interval [25].

As every time series, a set of pre-processing methods apply to Voice (audio) signals before feature extraction takes place. Some examples are described below:

- **Pre-emphasis:** enhances high-frequency content via a first-order differences filter [26];

- **Framing**: splits continuous audio signals into fixed-length overlapping segments, where features are extracted from [26];

- **Low-Pass Filter:** a undesired high-frequency content via a low-pass filter (6 kHz cut-off) [27];

In audio signals, features can be extracted from multiple domains. In fact, the temporal domain does not offer as relevant information as some others. Audio signal patterns are not easy to detect over time. That is why it is common to transform signals into both frequency and time-frequency domains. The most common spectral representation of an

audio signal is called spectrogram [28]. In Figure 2.3, a voice (audio) signal is shown along with its spectrogram representation.

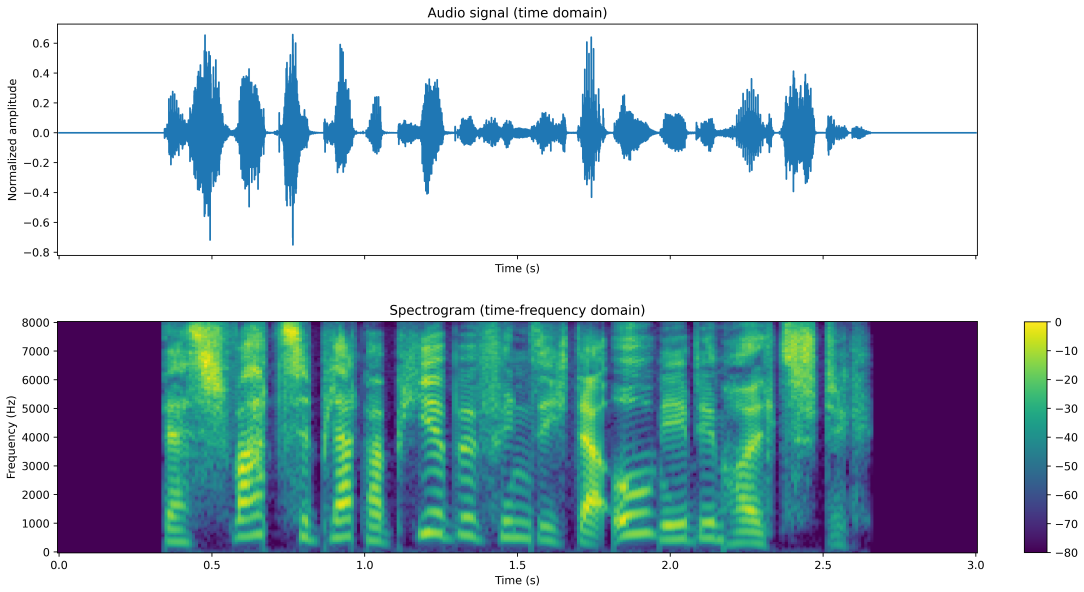


Figure 2.3: Illustration of voice activity captured by an audio signal. The waveform is represented over time (top) and time-spectral (bottom) domains. Intensity is given in Decibel (dB).

Some common features that can be extracted from a Voice signal are described as follows:

- **Spectral**: features extracted from the frequency (or time-frequency) domain, for example, the spectrum's roll-off, bandwidth, centroid, spread [29], period, energy [30];

- **Cepstral:** cepstrum is a type of spectral space computed through the inverse FT of the frequency spectrum's logarithm of the input waveform [31]. Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Cepstral Coefficients (LPCC) [29] and Cepstral Peak Prominence (CPP) [32] are some examples of features;

- **Temporal:** offer an indirect measure of frequency-related content, such as Zero Crossing Rate (ZCR) [29], and jitter [33];

- **Amplitude:** it measures oscillation levels in the amplitude domain, which can be tracked by shimmer [33].

2.2 Artificial Intelligence

AI involves methods that allow computers to mimic human behavior and surpass human decision-making in solving difficult tasks, either on their own or with limited human involvement [34]. ML is a subset of AI and refers a computer program's ability to enhance its performance through experience on a particular set of tasks and performance metrics [35]. As a sub-area of ML, DL uses multiple layers of nonlinear processing to automatically extract and transform features from data. This hierarchical structure results in a strong representation of features and makes DL ideal for processing and uncovering insights from large quantities of data from different sources [36].

2.2.1 Machine Learning Algorithms

In the ML field, algorithms can be divided into (1) supervised and (2) unsupervised, according with the availability of expected target labels. Supervised approaches can be further split to handle (1) classification or (2) regression tasks, whether the expected values are categorical or continuous, correspondingly. As supervised methods play an essential role in this work, a few of them are described below:

- **Decision Tree (DT):** algorithm that uses a tree-like model to make decisions based on a set of conditions. It works by recursively partitioning the input data based on the feature values and creating a tree of decisions that can be used to classify new data (threshold-based) [37];
- **Naive Bayes (NB):** algorithm based on Bayes Theorem, assuming independence between features (threshold-based) [37];
- **Support Vector Machine (SVM):** it separates classes by creating boundaries between them, maximizing the distance between their boundaries (distance-based) [37];
- **K-Nearest Neighbour (KNN):** it assigns an unknown input sample based on the distance to the k nearest labeled samples (distance-based) [38];
- **Linear Discriminant Analysis (LDA):** it involves identifying the hyperplane of projection that minimizes inter-class variance and maximizes the distance between projected means of classes (distance-based) [39];
- **Quadratic Discriminant Analysis (QDA):** it models the decision boundary between classes using quadratic functions of the predictors [40];

Ensemble algorithms are also part of this category, since they combine multiple weak learners to improve their generalization capabilities [41]. A few examples are:

- **Random Forest (RF):** combines the decisions of multiple independent DTs whose output is aggregated in a voting approach [42];

- **Light GBM (LGBM):** propagates gradient between multiple weak learners (DTs) to optimize their performance [43]. Prone to overfit;

- **Ada Boost (AB):** assigns different sample weights to the data that will feed the next learner, in an iterative fashion [44]. Also prone to overfit.

2.2.2 Deep Learning Algorithms

DL algorithms can be divided in (a) discriminative or (b) generative model whether they try to (a) draw boundaries in the data space (classification or regression) or (b) build an algorithm to model the inner data distribution, respectively [45].

An example of a DL architecture is the Convolutional Neural Network (CNN). This model utilizes sequential convolutional operations for spatio-temporal relationship analysis (see Figure 2.4). Convolutional layers condense input data, preserving meaningful features. It's applied in Natural Language Processing (NLP), speech processing, and computer vision [46]. Additionally, an Long Short-Term Memory (LSTM) network captures sequential patterns in data, valuable in various fields, such as speech recognition and sentiment analysis [46].

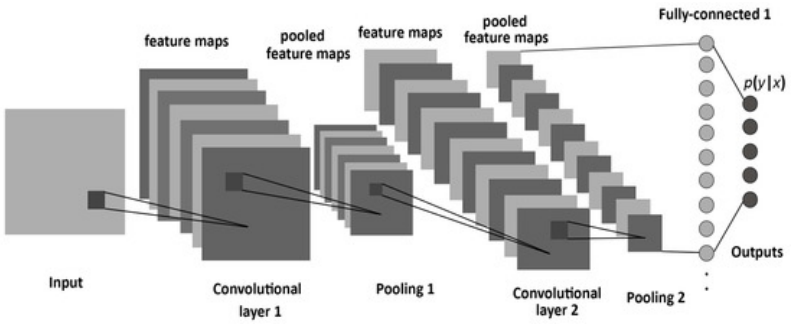


Figure 2.4: The structure of a CNN, consisting of convolutional, pooling, and fully-connected layers [47].

2.2.3 Feature Normalization

After extracting features and before feeding them for training models, it is common practice to normalize the feature sets. This aims to constrain each feature's value range and optimize ML algorithms' learning process. A few types are described as follows:

- **MinMax Scaling (M):** it scales data by their minimum and maximum values, being rather sensitive to outliers. It is described as:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{2.1}$$

where X is the original feature value, X_{min} its minimum, X_{max} its maximum, and X_{scaled} its scaled value;

- **Robust Scaling (R):** it scales data by their median and Interquartile (IQR) values, being, thus, more robust to outliers. It is formulated by:

$$X_{scaled} = \frac{X - median(X)}{IQR(X)} \tag{2.2}$$

where X is the original feature value, $median(X)$ its median, $IQR(X)$ its interquartile X_{min} its minimum, X_{max} its maximum, and X_{scaled} its scaled value;

- **Standard Scaling (S):** it scales the data to become zero mean, unit variance distributed. It is defined as:

$$X_{scaled} = \frac{X - \mu}{\sigma} \tag{2.3}$$

where X is the original feature value, μ its mean, σ its Standard Deviation (STD), and X_{scaled} its scaled value.

2.2.4 Data Fusion Techniques

The complementary and richer information obtained from multiple sensors has led to improved results compared to a single input. Although monomodal learning systems have made significant progress in accuracy and robustness, they still have limitations and inaccuracies. Researchers have turned to deep multimodal learning techniques to enhance model performance. To work with multimodal approaches, capturing and fusing informative features from multiple modalities is crucial and can be categorized into two main different techniques [48]:

- **Late Fusion:** classifying features from different modalities and then fuse to create the final prediction, but this can decrease the overall performance of the integration process [48]. Some examples of late fusion techniques are described above:
  - **Voting:** combines predictions from multiple classifiers by selecting the mode (most frequent prediction) across all individual classifiers for a given instance, and it can be described by the followed equation:

$$\hat{y}_j = mode(y_{1,j}, y_{2,j}, ..., y_{M,j}) \tag{2.4}$$

where M is the number of classifiers, $y_{i,j}$ the predicted class label by the i-th classifier for the j-th instance, and $\hat{y}_j$ the class predicted by the voting technique;

  - **Averaging:** does the predictions by taking the average of predicted class labels from multiple individual predictors, and it's described as follows:

$$\hat{y}_j = \frac{1}{M} \sum_{i=1}^{M} \hat{y}_{i,j} \tag{2.5}$$

where M is the number of classifiers, $\hat{y}_{i,j}$ the predicted class label by the i-th classifier for the j-th instance, and $\hat{y}_j$ the class predicted by the averaging technique;

- **Max Voting:** determinates the class by selecting the one with the highest combined probability across all individual modalities, and the equation that describes this is:

$$\hat{y}_j = \max_k(\max_i(\mathrm{p}_{i,j,k})) \tag{2.6}$$

where M is the number of classifiers, $\mathrm{p}_{i,j,k}$ be the probability predicted by the i-th modality for the j-th instance belonging for the k-th class, and $\hat{y}_j$ the class predicted by the voting technique.

- **Weighted Averaging:** gets the predictions by assigning weights to each predictor's prediction from multiple individual predictors. The equation is described as follows:

$$\hat{y}_j = \frac{\sum_{i=1}^{M} \mathrm{w}_i . \hat{y}_{i,j}}{\sum_{i=1}^{M} \mathrm{w}_i} \tag{2.7}$$

where M is the number of classifiers, $\hat{y}_{i,j}$ the predicted class label by the i-th classifier for the j-th instance, w_i the weight assigned by the i-th classifier, and $\hat{y}_j$ the class predicted by the weighted averaging technique.

- **Early Fusion:** features are combined at an early stage and fed into a single classifier. In this case, the diverse nature of data modalities and the rapid growth of features dimensionality may potentially overfit models and lead to inaccurate predictions [48].

2.2.5 Dimensionality Reduction

Dimensionality reduction simplifies high-dimensional data into a more compact, meaningful representation. Its aim is to capture data complexity with fewer parameters, known as intrinsic dimensionality. This technique addresses the curse of dimensionality, improving computational efficiency, data visualization, classification accuracy, and compression. It streamlines complex data for enhanced manageability and insights [49]. It exists various techniques to perform dimensionality reduction, such as Uniform Manifold Approximation and Projection (UMAP), that reduces the dimensionality by preserving the local and global structure of the data, optimized through a topological strategy which minimizes discrepancies in pairwise similarities between the original and lower feature spaces [50]. Others techniques are the Principal Component Analysis (PCA) [51], and t-distributed Stochastic Neighbor Embedding (t-SNE) [52].

2.2.6 Evaluation Metrics

The performance of trained models in classification problems can be evaluated globally through various metrics, being this a very important step on the model selection process. The confusion matrix, produced from the outcomes of the classification, is a widely used representation since every evaluation metric are derived from it [21]. In Table 2.1, a representation of a default confusion matrix is depicted.

Table 2.1: Confusion matrix for binary classification.

		Predicted	
		True	**False**
Actual	**True**	True Positive (TP)	False Negative (FN)
	False	False Positive (FP)	True Negative (TN)

The different values from each cell in a confusion matrix represent different characteristics. From Table 2.1, TP is the number of instances that are correctly predicted as positive, TN the number of instances that are correctly predicted as negative, FP the number of instances that are incorrectly predicted as positive, and FN the number of instances that are incorrectly predicted as negative.

Besides this, multiple metrics can be extracted from the confusion matrix and characterize the model's performance with single, absolute values. A few examples are given by:

- **Accuracy:** it measures the proportion of correctly classified samples to the total number of samples that were classified [21];

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.8}$$

- **Precision:** this metric is used to measure the proportion of true positive samples in the total number of samples that are predicted as positive [21];

$$Precision = \frac{TP}{TP + FP} \tag{2.9}$$

- **Recall:** this is a metric that measures how many of the actually positive samples are predicted as positive [21];

$$Recall = \frac{TP}{TP + FN} \tag{2.10}$$

- **F-Measure:** it calculates the harmonic mean between the recall and precision [21].

$$F - Measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{2.11}$$

But it is important to reinforce that the formulas of these performance metrics are designed for binary classification. When dealing with problems involving more than two classes, it becomes necessary to compute the metrics for each class individually and subsequently apply a specific type of averaging. There are three commonly used averaging techniques in multi-class scenarios, micro average, weighted average, and macro average, that performs an arithmetic mean of a given score over all classes. Being widely used to report results in imbalanced datasets, it is formulated as:

$$X_{macro} = \frac{1}{N} \sum_{i=1}^{N} X_i \tag{2.12}$$

where N is the number of classes, X_i the score of class i, and X_{macro} the macro-averaged score.

The Receiver Operating Characteristic Curve (ROC) curve, illustrated in Figure 2.5, provides an insightful visualization of the delicate balance between Sensitivity and Specificity scores. It offers a glimpse into the model's predictive confidence, revealing the interplay between its ability to correctly identify positive cases while avoiding false positives. One prominent metric derived from the ROC curve is Area Under the ROC Curve (AUC), which falls within the range of 0 to 1. This metric serves as a valuable gauge of how well the model distinguishes between the confidence levels of Positive and Negative class predictions. A higher AUC suggests a stronger discriminatory power, signifying the model's effectiveness in distinguishing between these two classes with confidence.

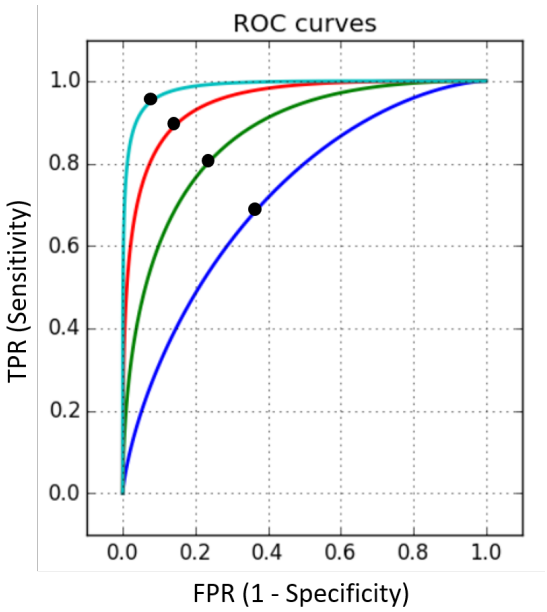


Figure 2.5: Representation of several ROC curve shapes. Each one corresponds to a different model performance [53]. The sharper the curve, the higher will be its AUC. The light-blue line achieved the best performance (as it has the highest area under its curve), while the dark-blue line obtained the worst one.

2.3 Emotions on Human-Computer Interaction

After a long period of time of not considering emotions in a HCI, it is now understood that they play a critical role in this type of environments. In fact, advancements in high-speed and high-quality signal processing have enabled technology to accurately assess a user's emotional state in real-time [54] as it has been demonstrated that a person's emotions are always reflected in their actions [55].

2.3.1 2D Valence-Arousal Model of Emotion

It is also important to know the type of emotions that exist so to machines have the capability to classify them when analyzing a human behaviour. On example of a model of emotion is the 2D Valence-Arousal, that describes emotions as a pair of valence and arousal intensity. Valence and arousal can go from low/negative to high/positive. So, the emotions can be viewed as a part of a 2D graph, where sentiments can be placed in four different quadrants, combining low and high arousal with low and high valence [56]: the first quadrant is called High Valence High Arousal (HVHA), the second one High Valence Low Arousal (HVLA), the third one Low Valence High Arousal (LVHA) and the fourth one Low Valence Low Arousal (LVLA). A visual representation of this model with examples of some emotions can be visualized in Figure 2.6. Some researchers also use a third dimension, the dominance, that represents the degree of control that a person has over their affective states [57].

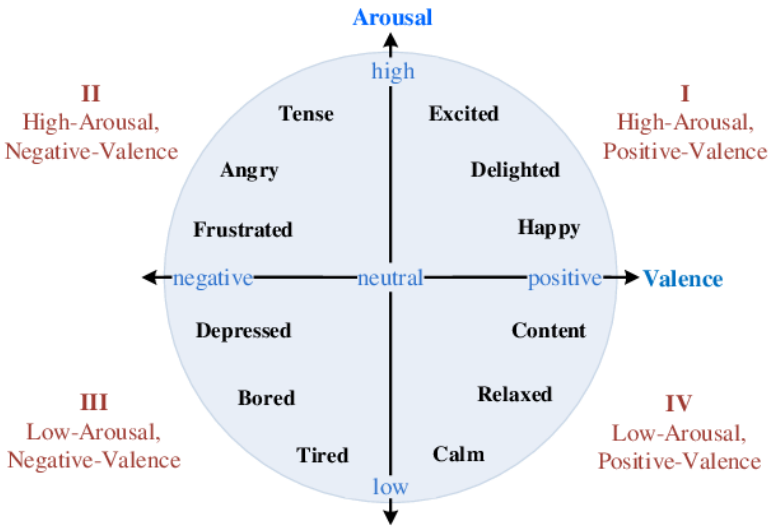


Figure 2.6: The 2D Valence-Arousal Model of Emotion [56].

2.3.2 Physiological Assessment of Emotions

The integration of the EDA, ECG and Voice signals births a new paradigm in emotion recognition. Each modality brings its unique lens, capturing facets of emotions unseen by the others. EDA emerges as an invaluable window into emotional arousal. This modality discerns arousal levels by measuring SCL changes in response to emotional stimuli, that is influenced by the automatic nervous system [58]. The ECG signal, by capturing the heart rate variations, it can measure the variations of the sympathetic and parasympathetic nervous systems as they influence it. In fact, it is known that an increase in the HR can be cause by an increase of the arousal of the emotions. On the other hand, emotions of negative valence cause the deceleration of it [59]. For the last, modifications in some features of the voice signal can be derived from both differences in the arousal as well

as the valence dimension. For example, high arousal emotions are linked with higher F0 mean and range, quicker pitch rises and falls, and increased vocal intensity. On the other hand, negative emotions are associated with specific acoustic variations, such as longer pauses, quicker F0 falls, and increased vocal intensity [60].

2.4 Biomedical Applications

Emotion assessment plays a important in various aspects, including mental health diagnosis and treatment for conditions such as depression, anxiety, and post-traumatic stress disorder [61]. In conditions such as autism spectrum disorder and neurodegenerative diseases (e.g., Huntington's disease, Parkinson's disease, and multiple sclerosis), it helps comprehend deficits in emotional processing and better tailor interventions to enhance emotional well-being and improve patients' quality of life [62, 63]. Emotion recognition has plenty of option for potentially offering insights to complement medical procedures (e.g, diagnoses, treatment outcomes) [64]. This understanding empowers healthcare professionals to tailor their communication and care strategies to align with the emotional needs of each patient [65]. Emotion recognition tools stand to benefit a substantial portion of the population [66]. Around 10% of the population exhibits Alexithymia trait, described as a limited capacity to self-describe their emotional state [66]. This percentage may raise up to 40% in case people already suffer from a prior mental disorder [66].

In healthcare, Tivatansaku and Ohkura (2013) developed a system utilizing augmented reality, emotional detection, and physiological monitoring to guide users in deep breathing for stress reduction [67]. Zenonos et al. (2016) introduced a novel mood recognition solution in workplace environments, identifying eight emotions using physiological signals namely, ECG, Photoplethysmography (PPG), Skin Temperature (ST), associated with a complementary app called *HealthyOffice* for optimizing mental health and productivity [68]. Anusha et al. (2018) conducted a workplace study, focusing on binary emotional classification (stress vs. non-stress) with ECG, EDA, and ST [69]. Tacconi et al. (2008) contributed to early mental disorder diagnosis using the P-cube platform, detecting six emotions (anger, boredom, desperation, disgust, happiness, pride) via smartphone or laptop [70]. Pioggia et al. (2010) addressed stress assessment with Interreality-based management, merging real and virtual worlds for effective treatment [71]. Ma et al. (2012) introduced mobile-based mood assessment, offering an objective method for daily mood tracking [72].

3

State of The Art

Emotion recognition has been a longstanding pursuit, with research spanning various modalities. Particularly, physiological and voice data have been integral in developing emotion classification tools. Many studies have focused on biosignals, excluding speech, to create robust emotion classifiers. For instance, one noteworthy study by Domínguez-Jiménez et al. (2020) leveraged both PPG and EDA signals to classify emotions into valence categories (positive, negative, neutral). Participants were exposed to video clips, and features were extracted from time and frequency domains. Employing an early fusion approach, and combining features from both signals, the researchers achieved an impressive 100% accuracy using Recursive Feature Elimination (RFE) for feature selection and SVM for classification [59].

Another avenue of research explores ECG signals. For example, Aziz et al. (2023) captured ECG and PPG signals from audiovisual stimuli to rigger and capture specific emotions. Several ML techniques, including SVM, KNN, NB, DT, RF, apart from DL approaches, were employed. Notably, SVM achieved an accuracy of 69% for arousal, 59% for valence (both using 2 classes), and KNN achieved 32% for dimensional emotion classification (in four quadrants of the 2D Valence-Arousal model) [73]. In Hristova, Grinberg and Lalev (2008), ECG, EDA, PPG, and Electromyography (EMG) signals were explored. A unique aspect of this study was calibrating the classifier using images and then testing it within an Embodied Conversational Agents (ECA) interaction scenario. LDA emerged as the top-performing algorithm for the ECG signal, achieving an averaged accuracy of 38% using negative, positive and neutral valence [74]. Some other studies regarding emotion recognition uses EDA signal such as [75–77].

An emerging trend involves combining ECG (or related signals such as PPG or HR) with EDA, as showcased in Ali et al. (2018). In this study, data collection was conducted in a subject-independent manner, with the model evaluation performed on subjects separate from those used during the training phase. Using CNN, an early fusion approach combining ECG and EDA signals resulted in a remarkable accuracy of 83% using four classes (from each quadrant of the 2D Valence-Arousal model) [78]. Moreover, Jang et al. (2014) employed an early fusion approach using EDA, ECG, ST, and PPG signals with

extended emotional stimuli from long film clips. Features were extracted, encompassing SCR, HRV, and HR statistics. Various machine learning algorithms were tested, with NB achieving the highest accuracy of 54%, with seven categorical emotions [79]. A similar approach was adopted in Xie, Xu and Shu (2018), which incorporated ECG, SCL, and EMG from a public dataset. The study aimed to classify five basic emotions, which were later condensed into two categories (positive and negative valence). Results showed that late or early fusion techniques, combined with specific feature selection methods, improved the unimodal results significantly. [80]. Besides this, there are a plenty of other studies that work with ECG signals such [75, 81, 82].

The Voice modality was also a focus of research, with Christ et al. (2022) analyzing that signal, along with text, and video recordings of soccer coaches during press conferences. A distinct approach here was the use of Martin's Humor Style to categorize emotions as positive or negative and self or other-directed, resulting in a 2D emotion space. Late fusion techniques were utilized, resulting in improved results compared to using a single modality in isolation. Combining audio, text, and video features achieved an AUC of 61% for emotion direction and 70% for their identification, as reported in [83]. Notably, Bo et al. (2014), using a hierarchical fusion approach with Voice signals, got an accuracy of 47% with seven classes [84]. Some other studies that work with voice signals are [85–87].

4

Methodology

In this chapter, the systematic technical framework comprising the proposed emotion recognition pipeline is described in several building blocks, from (a) datasets description, (b) data processing, (c) model development, and (d) data fusion strategies.

4.1 Data Sources

In this study, a range of databases were employed, including several publicly available ones that have been utilized in previous emotion recognition projects.

The Intelligent Customer and Advanced Natural Systems (ICANS) dataset, stemming from the detailed acquisition protocol outlined in subsection 4.1.1, stands as the most frequently employed. Additionally, three other public datasets are utilized: Young Adult's Affective Data (YAAD) and A Dataset for Affect, Personality and Mood Research on Individuals and Groups (AMIGOS) containing EDA and ECG signals, and European Portuguese Emotional Discourse Database (EPEDD) and An Emotional Speech Database In Portuguese (EMOUERJ), both encompassing Voice signals.

The YAAD dataset encompasses the EDA and ECG signals, spanning six core emotions (surprise, anger, fear, happiness, sadness, disgust, with an added neutral state), obtained from 12 participants interacting with 21 stimulus videos distributed across three sessions[88]. The AMIGOS dataset comprises physiological recordings (EDA, ECG). There were two different experiments, where the subjects were alone (and watched 16 short affective video clips) or in a group (and watched 4 long affective video clips), and there were obtained seven different felt emotions (neutral, happiness, sadness, surprise, fear, anger, and disgust) [75]. The EPEDD dataset, an acted emotional speech repository developed at Universidade Católica do Porto, involves 8 actors performing 9 emotions (anger, apathy, disgust, fear, interest, joy, sadness, surprise, and neutral) through 5 long sentences, 5 short sentences, and two single words, providing comprehensive recordings along with actor-related details [89]. The EMOUERJ database, originating from the State University of Rio de Janeiro, includes audio recordings featuring ten sentences for each emotion - happiness, anger, sadness, and neutral — each recorded by eight actors [85]. Some other

information about the datasets used are described in Table 4.1.

Table 4.1: Information about the datasets used in this work.

Name	Availability	Modalities	Labels	Stimuli	Nº Samples	References
ICANS	Private	EDA ECG Voice	5 Categorical Emotions [1]	Recall [2] Visual Auditory Audiovisual Acting	316	
YAAD	Public	EDA ECG	2D Valence-Arousal Quadrants [3]	Audiovisual	252	[88]
AMIGOS	Public	EDA ECG	7 Categorical Emotions [4]	Audiovisual	800	[75]
EPEDD	Public	Voice	9 Categorical Emotions [5]	Acting	826	[89]
EMOUERJ	Public	Voice	4 Categorical Emotions [6]	Acting	377	[85]

[1] Happiness, Stress, Sadness, Relaxation;
[2] Memory Recall;
[3] HVHA, LVHA, LVLA, HVLA;
[4] Happiness, Surprise, Anger, Fear, Disgust, Sadness, Neutral;
[5] Interest, Joy, Surprise, Anger, Disgust, Fear, Apathy, Sadness, Neutral; Happiness, Anger, Sadness, Neutral

4.1.1 Data Acquisition

Before the data acquisition process, participants completed demographic forms, providing age range, region of origin, and gender information.

The data acquisition protocol aimed to evoke emotions using various stimuli, including memory recall, visual and auditory cues, audiovisual presentations, and acting performances. Stimuli were chosen to elicit specific emotions, ensuring equal exposure to various emotional states. Visual stimuli were adjusted if a participant had phobias. Participants were also screened for phobias related to enclosed spaces or darkness.

The collected signals included audio, text, and biosignals (EDA and ECG). After each stimulus, participants rated their emotions (neutral, sadness, happiness, or relaxation), each one corresponding to a different quadrant of the 2D Valence-Arousal model, and their intensity on a scale from 1 to 5. Baseline status was recorded before data collection, with rest periods provided after each task. In Figure 4.1 is a illustration of the data acquisition protocol.

Two independent annotators evaluated each acquisition, assessing the individual's emotion and intensity. Both annotators provided their assessments based on the signal and reported emotions. The final emotion and intensity values resulted from a balanced combination of the subject's evaluation (weighted at 0.4) and each annotator's assessment (weighted at 0.3), ensuring a comprehensive assessment. Additionally, in Figure A.2 it's an illustration of the data acquisition setup.

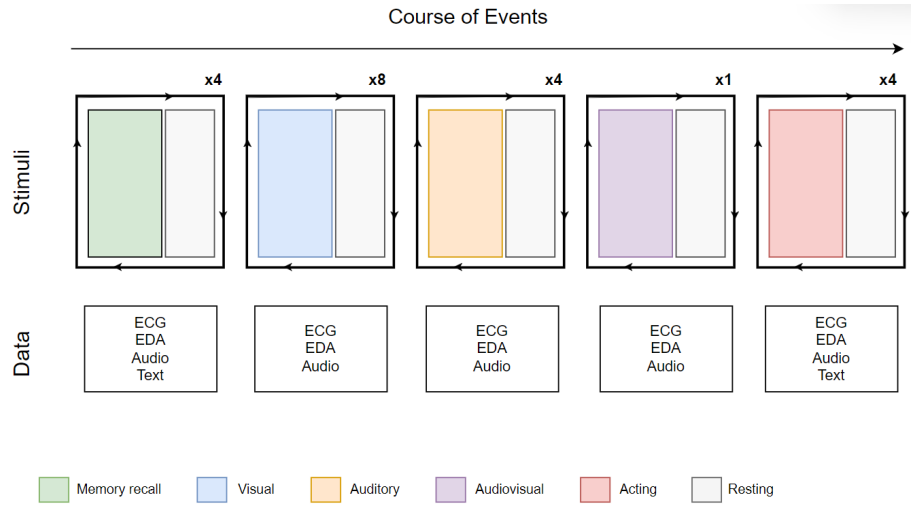


Figure 4.1: Squematic of the acquisition protocol. It consists of four memory recall, eight visual, four auditory, one audiovisual, and 4 acting stimuli. Data collected from each stimulus includes EDA, ECG, and Voice signals. Resting periods follow each stimulus.

4.2 Programming Environment

In order to develop this project it is necessary to set up a programming environment, being also needful to use some libraries as a way to correctly performance every task needed.

All the experiments performed in this work have been developed in Python language, on Visual Studio Code Integrated Development Environment (IDE). Multiple Python frameworks have been employed for signal pre-processing and feature extraction (Scipy [90], Numpy [91], Librosa [92], Neurokit2 [93], Heartpy [94], Biosppy [95]), feature manipulation (Pandas [96], UMAP [50]), data augmentation (Tsaug [97], Audiomentations [98]), data visualization (Seaborn [99], Matplotlib [100]), as well as ML (Scikit-learn [101], Lightgmb [102]) and DL (Tensorflow [103], keras [104]) algorithms implementation.

4.3 Data Processing

This subsection covers detailed data processing for EDA, ECG, and Voice signals. It starts with signal processing, then it's discussed feature extraction, feature normalizations are addressed, and data augmentation and dimensionality reduction techniques are explained.

4.3.1 Signal Processing

Signal processing is the initial step in the pipeline, and each signal type undergoes specific procedures:

- **EDA and ECG Signals:** resampled to 100 Hz, after which butterworth filters were computed to filter out undesired frequency bands. For ECG, a 5th order high-pass

filter (0.5 Hz cut-off) and a notch (50 Hz cut-off) were applied. For EDA, a 4th order low-pass filter (5 Hz cut-off) cleaned the signal. Both signals were framed into consecutive 5-second windows (50% overlap);

- **Voice Signal:** resampled to 16kHz. Long silences are removed based on an intensity-based threshold., followed by a pre-emphasis and a 4th order butterworth low-pass filter (6 kHz cut-off). The signal was framed into 60ms consecutive windows (50% overlap).

Normalization may or not be applied. For Voice, it ranges from -1 to 1. EDA and ECG signals normalize between 0 and 1, reflecting the original signal's minimum and maximum values. EDA and ECG normalization occurs after resampling, while Voice signal normalization follows beep and silence removal. In Figure A.3 it's a visual representation of the signals before and after the pre-processing steps described above.

4.3.2 Feature Extraction

Features are then extracted, either frame-based or signal-based, depending on the feature type. Frame-based features are aggregated through statistical descriptors (e.g., mean, maximum, minimum, kurtosis) along with the energy of the feature. In the ECG signal, additional statistical metrics such as the Root Mean Square (RMS) and Mean Absolute Deviation (MAD), which are commonly used in this modality, are included in their features. A summary of these extracted features per signal is provided in Table 4.2. In Figure A.4, a visual representation of the preprocessing steps applied to the EDA and ECG signals before feature extraction is presented.

Examining the details in Table 4.2, it's observed that the Voice signal contains features from all modalities, while the EDA signal covers time, amplitude, and spectral domains. In contrast, the ECG signal consists solely of features from the time and spectral domains. To access a comprehensive list of the features extracted from each signal and understand their composition, please refer to Annex I.

4.3.3 Feature Normalization

When it comes to feature normalization, various methods are employed, as described in subsection 2.2.3. Additionally, the normalization process can vary depending on specific conditions or groups, including:

- Full Column Normalization (FCN): scales the entire feature columns uniformly;
- Subject-Based Normalization (SBN): scales feature columns per subject;
- Baseline Normalization (BN): scales feature columns per each subject's baseline period (only applies to the ICANS dataset);

Table 4.2: Features extracted from each modality.

Data	Feature			
	Time	**Amplitude**	**Spectral**	**Cepstral**
EDA	SCR Risetime SCR Recovery Time SCR Intervals Number of SCRs	SCR Signal SCR Pulse SCL Signal	Rolloff Bandwidth Centroid Entropy Maximum Minimum Mean	-
ECG	HRV Measures Breathing Rate	-	Rolloff Bandwidth Centroid Entropy Maximum Minimum Mean VLF Power LF Power HF Power	-
Voice	Jitter ZCR	Shimmer	Rolloff Bandwidth Centroid Entropy Maximum Minimum Mean Contrast	MFCC CPP LPCC

- Rest Normalization (RN): identical approach to BN where features are scaled through the resting periods following each task (only applied to ICANS dataset, namely, EDA and ECG signals).

In total, there are 26 feature sets representing different combinations of normalization approaches: 4 types of feature normalization (FCN, SBN, BN, RN) × 3 methods of normalization (M, S, R) × signal normalization (yes or no). Additionally, there is a file dedicated solely to signal normalization or not, with no feature normalization applied.

The equation that describes this whole feature normalization process is:

$$X_t = S(X_i) \tag{4.1}$$

where X_i the initial set of values of a feature, and X_t the transformed ones, $S \in \{M, S, R\}$ and $t \in \{FCN, SBN, RN, BN\}$.

4.3.4 Feature Selection

Certain features within the dataset might lack relevance for the model's predictive capabilities. Consequently, feature selection techniques are employed to enhance the model's effectiveness and to reduce its dimensionality as much as possible. The number of selected features is optimized during the grid search process that is more described in subsection 4.4.1.

Besides that, it was first identified the best ten features from each feature set across the best-performing models in each dataset. These features were then scored on a 10-point

scale, with the best feature receiving 10 points and so forth. It was subsequently computed the cumulative scores for each feature across the best ten models. The ten features with the highest scores were chosen. Additionally, their relative importance was assessed by dividing the 'points' of each feature by the points awarded to the best feature. It is important to note that some features had multiple aggregations, and as a result, they might appear in the best ten multiple times. So, to avoid that, it was only focused primarily on their inclusion within the best ten features.

4.3.5 Data Augmentation

Due to thr large number of parameters in DL algorithms, a large number of training examples must be used to prevent overfitting issues. To overcome the lack of data, augmentation techniques applied before feeding data to train DL models.

Regarding ECG data, data augmentation did not apply since 5-second frames were used, resulting in thousands of examples, which seemed reasonable. For EDA signal, (a) time warp, (b) data drift, and (c) gaussian noise addition helped generate synthetic but morphologically identical signals. Voice signals' transformation was conducted via (d) pitch shifting, (e) clipping distortion, and (f) gaussian noise addition. Too see an example of an EDA and Voice signal before and after augmentation, refer to Figure A.5.

4.3.6 Dimensionality Reduction

Dimensionality reduction was applied to the Voice datasets with the purpose of data visualization and measure distance between datasets to investigate the intensity of stimuli or emotional responses in distinct contexts. To do this, UMAP technique was employed.

4.4 Model Development

After completing all data processing steps (including signal pre-processing and feature extraction), data is ready to be fed into classification models for training. In this stage, both feature harmonization and ML/DL optimization strategies are covered in detail. In Figure 4.2 it's an illustration of the pipeline of this work.

4.4.1 Machine Learning

For an effective classification ML pipeline, a sequence of key steps is essential. The process encompasses signal processing, feature extraction (and subsequent feature normalization). These processed features are then fed into ML algorithms, ultimately delivering the desired classifications. Initial tests with various classifiers identified overfitting concerns in some models, necessitating parameter adjustments. Then, it was conducted a grid search, and their parameters are described in detail in Table 4.3.

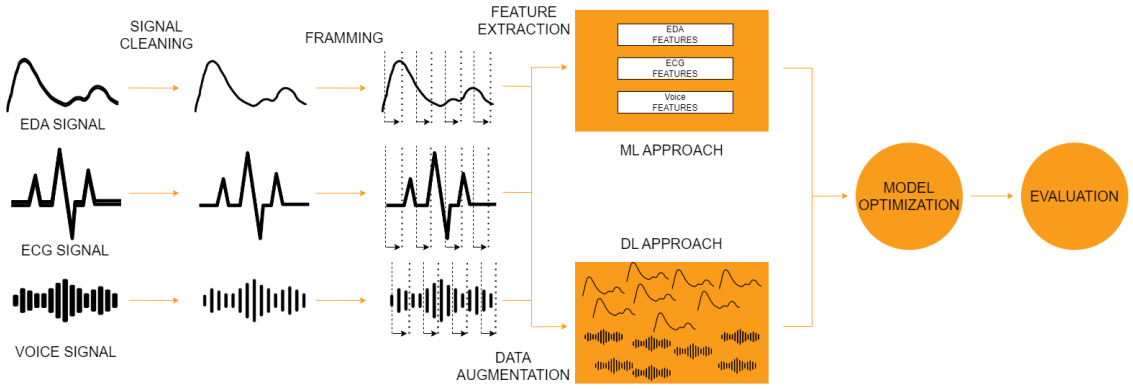


Figure 4.2: Model development of this work. It starts by the signal cleaning then it's framming. Then, in the ML approach, it occurs the feature extraction, whereas in the DL algorithm the data augmentation. Then, it's done the model optimization, and in the last step, the evaluation of the results obtained.

Table 4.3: Hyperparameter search ranges for diverse ML algorithms.

ML Algorithm	Hyperparameter	Search Range
DT	max_depth	[2, 8]
	min_samples_split	[0.05, 0.5]
	min_samples_leaf	[0.05, 0.5]
	max_leaf_nodes	[6, 14]
	max_features	[0.4, 0.9]
SVM	kernel	[rbf, poly]
	C	[0.01, 100]
KNN	n_neighbors	[5, 20]
	leaf_size	[10, 40]
RF	n_estimators	[4, 20]
	max_depth	[2, 8]
	min_samples_split	[0.05, 0.5]
	min_samples_leaf	[0.05, 0.5]
	max_leaf_nodes	[6, 14]
	max_features	[0.4, 0.9]
Light GBM	n_estimators	[4, 20]
	max_depth	[2, 8]
	num_leaves	[6, 14]
Ada Boost	n_estimators	[4, 20]

4.4.2 Deep Learning

The DL approach followed a sequential process encompassing, signal processing, data augmentation, and the implementation of dedicated deep learning algorithms to produce results. Simple CNN architectures were employed for each modality.

A notable difference across the modalities was the architecture of the convolutional layers. The EDA and ECG signals utilized 1D convolutional layers, while the Voice signal employed 2D convolutional layers. This architectural difference stemmed from the input data: the Voice signal used MFCC as input, whereas EDA and ECG signals used their raw waveform data. This architectural divergence accommodated the unique data representations and processing needs of each modality. In Table A.1, it's the training parameters for the DL algorithms. Conversely, Table A.2 contains the architecture details for the CNN used with all the signals.

4.4.3 Training Strategy

In the training strategy employed, it was utilized a stratified group k fold approach, where groups where subjects. This method was chosen to counteract potential subject bias, ensuring an unbiased distribution of data between the training and testing sets. This involved partitioning the data into three folds and separating the samples to achieve optimal balance among the classes.

For a comprehensive assessment of model performance, two main metrics (Balanced Accuracy (BA) and F1-macro), bolstered by an STD due to the folds, played a significant role in interpreting the grid search results. These metrics prove particularly valuable when dealing with imbalanced datasets. The issue of class imbalance holds significant importance in ML, as it can hinder the model's learning process and lead to overconfidence towards the majority class. In this context, the recall macro (BA) metric played a vital role by assessing recall for each individual class and then aggregating these values, thereby ensuring a fair and impartial evaluation across all classes, including those with limited instances.

4.5 Emotion Recognition

This section encompasses the exposition of the emotion recognition task through both unimodal and multimodal approaches, integrating all the themes discussed throughout this chapter.

4.5.1 Unimodal

Following the development of models and data processing, emotion recognition within each modality was meticulously evaluated, with model performance assessed using F1

and BA scores across various models described in both ML and DL sections (see subsection 4.4.1 and subsection 4.4.2 respectively). This encompassed public and private datasets, with a focus on selecting the best-performing models. Optimal features were extracted with different normalization techniques. Additionally, performance metrics were extracted for the best models in the private dataset, and validation extended across diverse dataset characteristics. In ML, a cross-validation strategy has been applied. In DL, it was exclusively employed the best models within their respective datasets, excluding cross-testing due to minor parameter variations that did not significantly affect outcomes.

4.5.2 Multimodal

Concerning multimodal emotion recognition involving data fusion techniques, this can be categorized into two primary techniques: late fusion and early fusion. In Figure 4.3 there's an illustration of the multimodal, emotion recognition approach.

4.5.2.1 Late Fusion

After performing grid search, the three best performing models and parameters per modality were chosen for late fusion, involving techniques as voting, averaging, weighted averaging, and max voting. Common samples from these modalities were included in the test group, and predictions from each modality were aggregated to leverage their strengths for better results. Initially, fusion was applied across all modalities, followed by combinations of two. This was exclusive to datasets with multiple modalities and the best-performing model (whether ML or DL).

4.5.2.2 Early Fusion

As for the early fusion technique, it involved the amalgamation of features from all modalities at an initial stage. Subsequently, the same methodology described in subsection 4.5.1 was applied.

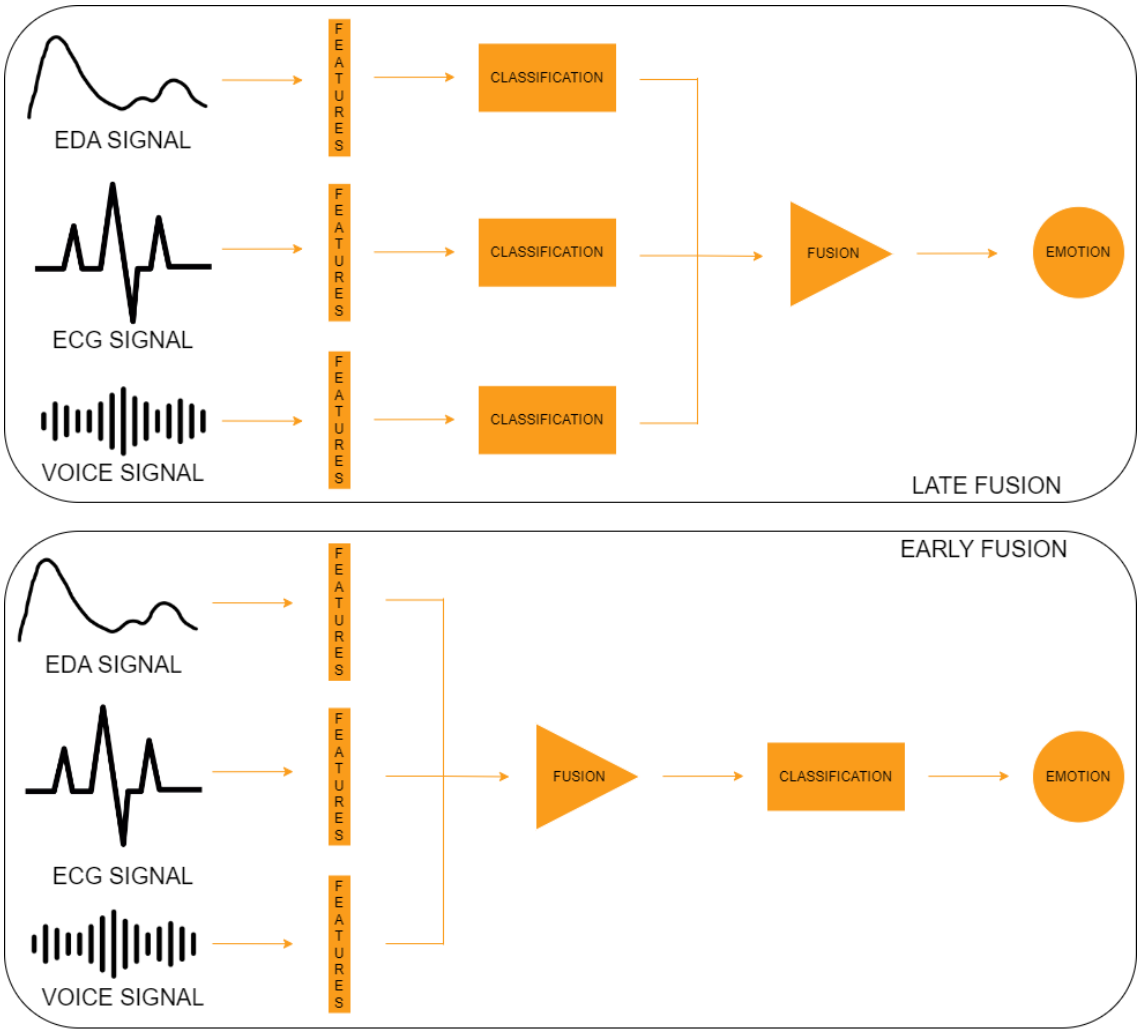


Figure 4.3: Representation of different strategies to combine the three different data modalities. Late (top) and Early (bottom) fusion techniques are shown.

5

Results and Discussion

This chapter presents the main results and takeaways under the proposed emotion recognition approach, validated in public and private datasets, alongside additional performance-based analysis.

5.1 Low Intensity Scenario

One of the primary and compelling conclusions that emerge from this study underscores the stark contrast in the intensity (expressiveness) of Voice recordings from the private dataset (ICANS) compared to the other two publicly available sources (EPEDD and EMOUERJ). This notable distinction is vividly captured in Figure 5.1, serving as a visual testament to the divergence in intensity levels across these datasets.

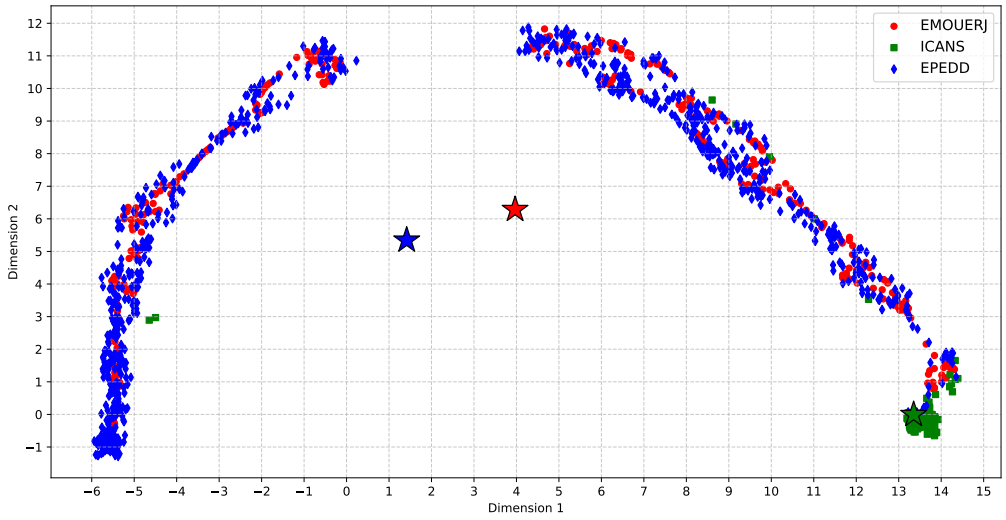


Figure 5.1: Dimensionality reduction of voice datasets. Star symbols represent the centroid of each representation, namely, for EMOUERJ (red), EPEDD (blue) and ICANS (green).

This visual representation was built by projecting the same set of vocal features from these datasets into a lower-dimensional space (via UMAP technique). By visual inspecting Figure 5.1, it clearly seems that the private dataset occupies a significantly smaller area on the graph in comparison to the other two datasets utilized in this study. This observation hints at a low-intensity stimuli environment on the ICANS dataset side, with subtle emotions being triggered instead of simulated (public datasets).

5.2 Unimodal Emotion Recognition

Within this section, it's delved into the results and discussion of the unimodal approach across the three signals: EDA, ECG, and Voice. For each modality, it's showcased the averaged performance results of the best models in both public and private datasets, considering both ML and DL approaches. Additionally, it's highlighted the optimal features across all datasets.

5.2.1 Electrodermal Activity

As detailed in subsection 4.5.1, the ML approach involved the execution of a grid search, where the choice of the best models relied on F1-macro and BA scores, whereas the DL techniques were evaluated through a sequential feed forward CNN architecture.

A comprehensive summary of the best results obtained for each ML and DL approach is depicted in Table 5.1. Models have been evaluated on the three validation datasets (ICANS, YAAD, AMIGOS) using the same set of scores (F1-macro, BA, and AUC).

Table 5.1: Summary of the best results from the EDA signal. All metrics are macro averaged. In the DL approach, a single train-validation-test fold is used. The best result is highlighted.

Dataset	**Model**	**Metrics**		
		F1	BA	AUC
ICANS	DT (SBN)	0.40 ± 0.01	0.44 ± 0.01	0.67 ± 0.07
	CNN	0.23 ± 0.00	0.16 ± 0.00	0.47 ± 0.23
YAAD	RF (SBN)	0.39 ± 0.02	0.45 ± 0.03	0.63 ± 0.05
	CNN	**0.45 ± 0.00**	0.37 ± 0.00	**0.72 ± 0.07**
AMIGOS	RF	**0.45 ± 0.00**	**0.49 ± 0.01**	0.70 ± 0.05
	CNN	0.28 ± 0.00	0.23 ± 0.00	0.48 ± 0.09

SBN - Subject Based Normalization.

By observing Table 5.1, it is evident that the emotion recognition performing assessment reached similar outcomes for all datasets, with slightly better performances obtained in the ML approach of the AMIGOS dataset (49% BA) and in the DL approach of the YAAD dataset (72% AUC), handling a 4-class classification problem. Figure B.1 and Figure B.2 depicts the ROC curves, across the three datasets, of the best models from ML and DL approaches, respectively. Additional experiments were carried out to also assess the generalizability of such models at evaluating samples from external datasets.

As expected, performance drops were consistently denoted no matter the dataset where the model was trained with. To view the results of these experiments, Figure B.7 provides a summary of the F1-score and BA across all datasets, including scenarios where models and parameters from one dataset are tested on others. Furthermore, in-depth results spanning from Table B.1 to Table B.9 encapsulate the comprehensive findings of these experiments.

Previous studies have already explored AMIGOS dataset for handling an emotion recognition task. Miranda-Correa et al. (2021) performed a binary classification (arousal, valence detection) using EDA signals [75]. When considering all the available data (from short and long video stimuli), these obtained F1-score of 54% for valence and arousal detection. The work done by Dessai, Virani (2023) achieved an accuracy of 98% in binary valence and arousal detection, though with the use of a very robust pre-trained type of CNN architecture [76]. Also, the experiences conducted by Rahim et al. (2019) lead to an accuracy of 68% in real-time 7-class classification [77]. Remarkably, in this work, employing a subject-independent ML approach, the results that were reached was of 73% and 60% of F1-score and 73% and 61% accuracy for valence and arousal recognition, respectively. Besides this, the best accuracy reached in this thesis, regarding a 4-class approach, was of 46%. No available results were found regarding the EDA signal analysis of the YAAD dataset.

After the performance-based analysis, feature ranking inspection was conducted to validate the consistency of the most often chosen features amongst datasets. The results are present in Table 5.2.

Table 5.2: Summary of the best ten features selected across the best models for the EDA modality.

Domain	**Features**	**Datasets**		
		ICANS	YAAD	AMIGOS
Amplitude	**SCL Amplitude**	✓	✓	✓
	SCR Amplitude	✓	-	✓
	SCR Pulse	-	-	✓
Spectral	Spectral Bandwidth	-	✓	-
	Spectral Maximum	-	-	✓
	Spectral Min	✓	✓	-
	Spectral Mean	✓	-	✓
	Spectral Rolloff	✓	-	✓
Time	Number SCRs	✓	✓	-
	SCR Intervals	✓	-	✓
	SCR Recovery Time	-	✓	✓
	SCR Rise Time	-	-	✓

Across both the public and private datasets, one feature consistently stands out in the best 10 rankings, which is SCL. SCL represents a vital component of an EDA signal, being usually reflected in emotional responses (arousal levels) [58]. However, from a broader perspective, more emphasis is placed on selecting features from the time and

amplitude domain rather than the spectral ones. Referencing Figure B.10, it can observed three graphs illustrating the top ten features of each dataset and their corresponding relative importance.

5.2.2 Electrocardiography

Recapping the research conducted on the ECG signal, as detailed in subsection 4.5.1, the ML aprroach involved the execution of a grid search. The selection of the best models relied on F1-macro and BA scores. In the DL approach, the performance of a sequential CNN architecture was also assessed.

A comprehensive summary of the best results obtained for each ML and DL approach is shown in Table 5.3. Models have been evaluated on the three validation datasets (ICANS, YAAD, AMIGOS) using the same set of scores (F1-macro, BA, and AUC).

Table 5.3: Summary of the best results from the ECG signal. All metrics are macro averaged. In the DL approach, a single train-validation-test fold is used. The best result is highlighted.

Dataset	**Model**	**Metrics**		
		F1	BA	AUC
ICANS	DT (FCN)	**0.51 ± 0.01**	**0.52 ± 0.01**	**0.70 ± 0.06**
	CNN	0.29 ± 0.00	0.27 ± 0.00	0.58 ± 0.09
YAAD	RF (SBN)	0.43 ± 0.12	0.49 ± 0.12	0.51 ± 0.15
	CNN	0.44 ± 0.00	0.41 ± 0.00	0.62 ± 0.10
AMIGOS	RF	0.43 ± 0.02	0.47 ± 0.03	**0.70 ± 0.04**
	CNN	0.25 ± 0.00	0.13 ± 0.00	0.52 ± 0.10

SBN - Subject Based Normalization | FCN - Full Collumn Normalization

Upon reviewing Table 5.3, it becomes clear that the evaluation of emotion recognition performance yielded consistent results across all datasets, with slightly improved performance observed in the ICANS dataset (51% BA, 70% AUC) when dealing with a 4-class classification problem. Figure B.3 and Figure B.4, depicts the ROC curves of the top-performing models obtained through ML and DL methodologies, respectively, across all three datasets. Further experiments were conducted to evaluate the models' ability to generalize by assessing samples from external datasets. As anticipated, performance consistently decreased regardless of the dataset used for model training. To examine the outcomes of these experiments, Figure B.8 presents a summary of the F1-score and BA across all datasets, encompassing situations where models and parameters from one dataset are assessed on others. Moreover, you can find comprehensive results in the range from Table B.10 to Table B.18, which cover the extensive findings of these experiments.

The work done by Najam, Dar and Rahim (2022) lead to an impressive 70% accuracy for ECG signal classification, using four different classes [88]. However, it was employed a complex LSTM neutral network architecture, besides having access to more extensive data, including signals from separate sessions. Another study done by Alam, Urooj and Ansari (2023) reported 99% accuracy in classifying 7 emotions, but this result might be

due to overfitting [105]. On the other hand, considering the four classes classes, this thesis, using a subject-independent approach, led to an accuracy of 44%.

For the AMIGOS dataset, a work done by Miranda-Correa et al. (2021) reported 55% F1-scores for valence and arousal [75]. He et. al (2021) achieved 71% and 72% accuracy for binary classification (arousal and valence) [81]. Sepúlveda et al. (2021) reached an accuracy of 89% in the valence dimension, 90% in arousal, (both using two classes) and 95% in a two-dimensional classification (four classes) [82]. In this work, 73% and 52% F1-score and 74% and 55% accuracy was obtained in detecting valence and arousal, respectively, and 45% of accuracy in a four class problem. Such variations can be explained by the use of simpler models or a subject-independent approach, which is not considered in the aforementioned studies.

Following the performance-based analysis, a feature ranking inspection was carried out to assess the consistency of the most frequently selected features across datasets. The findings are presented in Table 5.4.

Table 5.4: Summary of the best ten features selected across the best models for the ECG modality. All metrics are macro averaged. In the DL approach, a single train-validation-test fold is used. The best result is highlighted.

Domain	**Features**	**Datasets**		
		ICANS	**YAAD**	**AMIGOS**
Spectral	HF Power	✓	-	✓
	LF Power	✓	-	✓
	Spectral Entropy	✓	✓	-
	Spectral Max	✓	-	-
	Spectral Min	✓	✓	-
	Spectral Rolloff	✓	-	-
	Total Power	✓	-	✓
	VLF Power	-	-	✓
Time	**BPM**	✓	✓	✓
	HR MAD	✓	-	-
	HR Entropy	-	✓	✓
	IBI	✓	-	✓
	RMSSD	-	✓	✓
	RRPL	-	✓	-
	RRPW	-	-	✓
	STD of NN Intervals	-	✓	✓
	STDSD	-	✓	-

Examining Table 5.4, it becomes evident that the feature consistently chosen across all three datasets is the BPM. Furthermore, taking a broader view, it's noteworthy that metrics related to HRV occupy a prominent position in feature selection, especially when considering spectral features that are not signal-specific. This underscores the importance of HRV metrics in capturing essential physiological insights [106] across various datasets. Within Figure B.11, there are three graphs depicting the top ten features of each dataset, complete with their relative importance.

5.2.3 Voice Activity

Reviewing the research conducted on the Voice signal, as elaborated in subsection 4.5.1, the ML approach included the execution of a grid search. The selection of the optimal models relied on F1-macro and BA scores, while the DL techniques were evaluated through a sequential feed forward CNN architecture.

A comprehensive summary of the best results obtained for each ML and DL approach is depicted in Table 5.5. Models have been evaluated on the three validation datasets (ICANS, EPEDD, EMOUERJ) using the same set of scores (F1-macro, BA, and AUC).

Table 5.5: Summary of the best results from the Voice signal. All metrics are macro averaged. In the DL approach, a single train-validation-test fold is used. The best result is highlighted.

Dataset	**Model**	**Metrics**		
		F1	BA	AUC
ICANS	RF (SBN)	0.48 ± 0.05	0.51 ± 0.07	0.67 ± 0.06
	CNN	0.23 ± 0.00	0.16 ± 0.00	0.49 ± 0.09
EPEDD	SVM (SBN)	0.60 ± 0.03	0.61 ± 0.04	0.89 ± 0.06
	CNN	0.25 ± 0.00	0.17 ± 0.00	0.55 ± 0.14
EMOUERJ	SVM (SBN)	**0.77 ± 0.09**	**0.77 ± 0.09**	**0.93 ± 0.03**
	CNN	0.75 ± 0.00	0.73 ± 0.00	0.90 ± 0.10

SBN - Subject Based Normalization. All metrics are macro averaged.

After examining Table 5.5, it is evident that the assessment of emotion recognition performance produced better results in the public datasets, with the best being noted in the EMOUERJ dataset (77% BA, 93% AUC) when addressing a 4-class classification challenge. Figure B.5 and Figure B.6, depicts the ROC curves of the best-performing models from ML and DL approaches, respectively, are presented across all three datasets. Additional experiments were carried out to assess the models' generalization capabilities when evaluating samples from external datasets. Upon scrutinizing this preceding graph, the deductions highly contrast with the findings in the other two modalities. Notably, the outcomes diverge between the private and public datasets, with considerably higher values observed in the latter. Moreover, when transferring the best models and parameters from the private dataset to assess their performance on the public dataset, notable performance improvements are evident. Conversely, reversing this process leads to a drop in performance. Furthermore, in addition to these findings, the best results in the EMOUERJ dataset are observed when it utilizes the training pipeline of the EPEDD, and vice versa. This can be attributed to the similarity between these two datasets, as illustrated by their 2D projection in Figure 5.1. To review the results of these experiments, Figure B.9 provides a summary of the F1-score and BA across all datasets, including scenarios where models and parameters from one dataset are tested on others. Furthermore, a detailed breakdown of the results can be found in the tables from Table B.19 to Table B.27, providing a comprehensive overview of the experiment outcomes.

Comparing the results of this work with the findings from other studies utilizing the

[85] dataset, Luthman (2023) achieved an accuracy of 86% in a 4-class classification task [86]. However, it's worth noting that the approach involved using pre-trained DL algorithms, whereas this work primarily relies on straightforward CNN networks, with the best-performing model adopting an ML approach. Additionally, Duret, Parcollet and Estève (2023) reported an impressive accuracy of 97%, using four labels, when training on various languages and testing in Brazilian Portuguese [87] against an 80% accuracy score obtained in this work. No available results were found regarding the EPEDD dataset.

Following the performance-based analysis, we proceeded to inspect the ranking of features to confirm the consistency of the most frequently selected features across datasets (see Table 5.6).

Table 5.6: Summary of the top-ten features selected across the best models for the Voice modality.

Domain	**Features**	**Datasets**		
		ICANS	EPEDD	EMOUERJ
Cepstral	LPCC	✓	-	-
	MFCC	-	✓	✓
Spectral	Pitch	-	✓	✓
	Spectral Contrast	✓	✓	✓
	Spectral Min	-	✓	-
	Spectral Rolloff	-	-	✓

The consistent utilization of Spectral Contrast across all three datasets underscores its paramount importance in voice signal analysis [107]. Spectral Contrast excels at capturing variations in spectral content, a critical aspect of voice speech, enabling the discrimination of subtle acoustic patterns and nuances tied to emotions. A more comprehensive examination reveals that cepstral and spectral features stand out as the most crucial components in emotion recognition tasks. In Figure B.12, you'll find three graphs showcasing the top ten features of each dataset, along with their respective relative importance.

5.3 Multimodal Emotion Recognition

The multimodal approach involved selecting the top three models with their respective parameters for each modality across all datasets, prioritizing the ML approach for its superior performance. Various data fusion techniques were then assessed.

A distinction should be noted between the individual results shown in Table 5.7 and those presented in the previous sections. The former evaluates samples from the individual testing sets of each data modality in separate, while the latter encompasses only the common samples amongst them. As a result, voice signal results are expected to remain relatively stable, given the fewer samples due to certain tasks not extracting this signal, as explained in subsection 4.1.1. Furthermore, early fusion consolidates all features across each sample. Thus, Table 5.7 showcases the best-performing early and late fusion techniques for all possible modality combinations (using pairs or all three), along with the

respective F1 macro and BA for each modality within each fusion scenario.

Table 5.7: Performance outcomes for the finest early and late fusion techniques employed within each modality group across all fusion scenarios, in the ICANS dataset. The best result for each fusion experiment is highlighted.

Fusion Type		**Metrics**	**Fusion Experiments**			
			1	2	3	4
Unimodal	EDA	F1	0.39 ± 0.05	0.42 ± 0.05	0.38 ± 0.04	-
		BA	0.43 ± 0.05	0.43 ± 0.05	0.39 ± 0.06	-
	ECG	F1	0.37 ± 0.02	0.44 ± 0.01	-	0.47 ± 0.04
		BA	0.43 ± 0.01	0.50 ± 0.02	-	0.49 ± 0.07
	Voice	F1	0.48 ± 0.03	-	0.47 ± 0.03	0.48 ± 0.07
		BA	0.50 ± 0.04	-	0.48 ± 0.04	0.49 ± 0.07
Multimodal	Late	F1	**0.50 ± 0.03**	**0.50 ± 0.00**	**0.48 ± 0.05**	**0.55 ± 0.00**
		BA	**0.55 ± 0.02**	**0.55 ± 0.00**	**0.49 ± 0.07**	**0.61 ± 0.00**
	Early	F1	0.41 ± 0.05	0.40 ± 0.02	0.38 ± 0.04	0.37 ± 0.00
		BA	0.42 ± 0.06	0.41 ± 0.01	0.39 ± 0.05	0.37 ± 0.01

All metrics are macro averaged; (1) - EDA + ECG + VOICE | (2) - EDA + ECG | (3) - EDA + VOICE | (4) - ECG + VOICE

Among the fusion scenarios involving the three modalities and the EDA-Voice fusion, weighted average emerged as the most effective, while max voting proved optimal for other fusions. Late fusion consistently outperformed individual modalities, with the most significant enhancement observed in ECG-Voice fusion achieving a 61% BA. However, the EDA-Voice fusion showed a more modest improvement in performance metrics, possibly due to their distinct responses to emotional cues, resulting in class assignment discrepancies. In early fusion, results consistently exhibited lower F1 scores and BA compared to late fusion, with early fusion almost never outperforming individual modalities. This discrepancy arises from distinct models, parameters, and training sample sizes used by each modality, while early fusion imposes constraints on shared training samples and uniformity in model parameters. In Table C.1, it can be observed the best results for each data fusion technique, encompassing all possible combinations of the three modalities.

The accuracy outcomes for each late fusion modality are presented in Table C.2, with the table also displaying the percentage of correctly identified samples within the subset of correct samples of each modality. Values approaching one indicate close alignment between well classified samples samples of a given modality and its corresponding one. For instance, in the ECG-Voice late fusion, the voice-based model classified well 63% of samples correctly identified by the ECG model, and vice-versa with a 58% accuracy, indicating a better fusion effectiveness than any other combination. However, in cases involving two modalities with EDA, values are comparatively lower, meaning a lower raise in performance.

By observing Table 5.8, it is possible to see the discrepancy in ECG modality results in the YAAD dataset. This variation can be attributed to a notable STD among the best

Table 5.8: Performance outcomes for the finest early and late fusion techniques employed within the fusion of EDA and ECG signals, in the YAAD and AMIGOS datasets. The best result for each fusion is highlighted.

Fusion Type		Metrics	Datasets	
			YAAD	AMIGOS
Unimodal	EDA	F1	0.38 ± 0.02	0.45 ± 0.04
		BA	0.43 ± 0.05	0.49 ± 0.02
	ECG	F1	0.35 ± 0.01	0.44 ± 0.03
		BA	0.35 ± 0.02	0.47 ± 0.04
Multimodal	Late	F1	0.39 ± 0.00	**0.48 ± 0.00**
		BA	**0.59 ± 0.00**	**0.54 ± 0.00**
	Early	F1	**0.43 ± 0.07**	0.45 ± 0.04
		BA	0.46 ± 0.08	0.41 ± 0.01

three results, suggesting sensitivity to test sample arrangements, especially when combined with the EDA signal. Across both modalities and datasets, fusion techniques consistently match or enhance individual modality results. Late fusion leads to a better result in both datasets, achieving an BA of 59% in the YAAD dataset and of 54% in AMIGOS dataset, both utilizing majority voting. Comparing these strategies with those in the ICANS dataset, it's clear that the ICANS dataset consistently yields superior results, particularly with late fusion techniques. The optimal outcomes for each late fusion technique, applied to the combined EDA and ECG signals of both the YAAD and AMIGOS datasets, are displayed in Table C.3.

5.4 Emotion Recognition and Additional Metadata

Besides the main task that motivated this work, numerous supplementary experiments were conducted to gain a comprehensive understanding of the ML challenge and shed light on results from unimodal and multimodal emotion recognition. These experiments explored diverse signal and feature normalization techniques, detailed in subsection 4.3.3. The analysis focused on normalization approaches in the top ten results for each modality across all datasets, including signal normalization, feature normalization domain (or absence), and specific types of feature normalization. Refer to Table 5.9 for further details.

Voice signal often lacks normalization, while the ECG signal in ICANS and YAAD datasets predominantly leans towards full normalization. Notably, EDA signal normalization shows variability across models, possibly due to its versatility. However, in the AMIGOS dataset, normalization trends for both EDA and ECG signals differ, indicating a notable degree of signal similarity among subjects.

Additionally, experiments were conducted to assess the impact of the agreement between annotators in the models' performance in ICANS dataset. The analysis considered different types of agreement: between the subject and two external annotators (total agreement), between annotators themselves (2 annotators), and between an annotator and

Table 5.9: Summary of normalizations used in the best ten results across the three modalities, using all datasets. The proportion of times (within [0,1]) a normalization type has been chosen on each experiment of every dataset is reported per each data modality.

Modalities	Datasets	Signal Amplitude		Features Type				Features Domain				
		Yes	No	M	R	S	No	FCN	SBN	RN	BN	No
EDA	ICANS	0.5	0.5	0.7	0.0	0.3	0.0	0.2	0.3	0.3	0.2	0.0
	YAAD	0.4	0.6	0.3	0.4	0.3	0.0	0.5	0.5	-	-	0.0
	AMIGOS	0.5	0.5	0.1	0.0	0.0	0.9	0.1	0.0	-	-	0.9
ECG	ICANS	0.6	0.4	0.2	0.3	0.4	0.1	0.6	0.0	0.3	0.0	0.1
	YAAD	0.3	0.7	0.4	0.2	0.3	0.1	0.4	0.6	-	-	0.1
	AMIGOS	0.5	0.5	0.0	0.2	0.1	0.7	0.1	0.2	-	-	0.7
Voice	ICANS	0.1	0.9	0.3	0.2	0.5	0.0	0.0	1.0	-	0.0	0.0
	EPEDD	0.2	0.8	0.1	0.3	0.6	0.0	0.0	1.0	-	-	0.0
	EMOUERJ	0.1	0.9	0.4	0.0	0.5	0.1	0.3	0.6	-	-	0.1

M - MinMax Scaling | R - Robust Scaling | S - Standardization | FCN - Full Column Normalization | SBN - Subject Based Normalization | RN - Rest Normalization | BN - Base Normalization

the subject (annotator + subject). Figure 5.2 illustrates the performance changes across these agreement types.

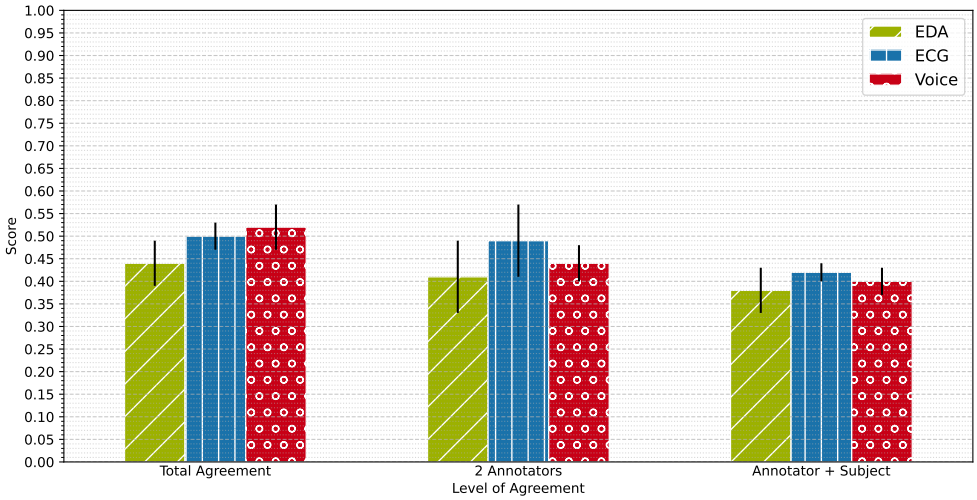


Figure 5.2: Accuracy based on the label agreement in the ICANS dataset. Total agreement occurs when both the external annotators and the subject agrees in the emotion that classifies that task, and so on (with two annotators and annotator + subject). Accuracy score is reported taking into account the best ten models of each modality.

The analysis highlights higher accuracy in samples where there is complete consensus between the individual and both annotators. Subsequently, agreement between both annotators comes with slightly worse performance, followed by the agreement between one annotator and the individual itself. This pattern suggests that a combined target label may potentially be more consistent and robust to mitigate isolate label outliers, which can be crucial in multiple tasks such as this one, where uncertainty on target emotions

can hinder the model learning process and lead to unreliable and non-optimal results.

Another experiment involved assessing the accuracies of various modalities within each task of the acquisition protocol. This aimed to identify the most effective modalities used in similar studies and their performance rankings. Additional efforts were made to establish associations between the best model performance and the type of stimuli/task applied during the acquisition protocol. Figure 5.3 provides a visual representation of such associations.

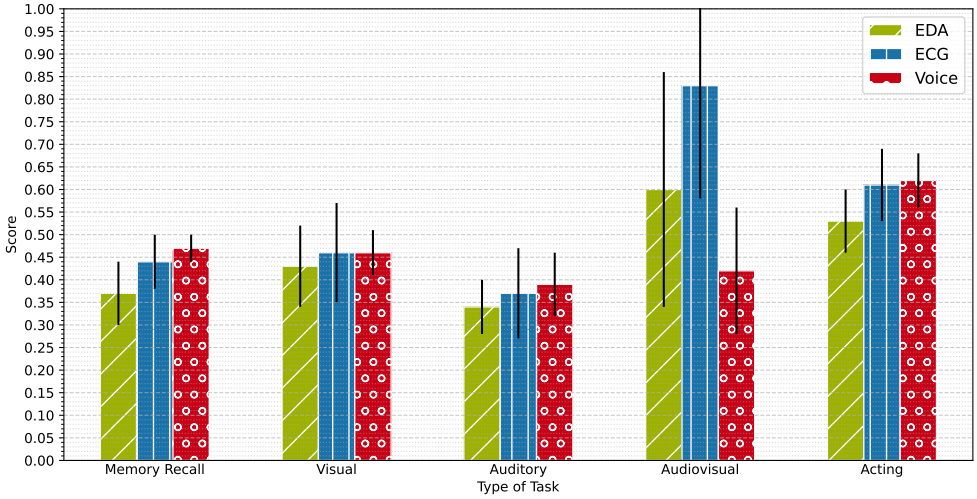


Figure 5.3: Graphical representation of the performance obtained for each type of stimuli/task, in the ICANS dataset. Accuracy score is reported taking into account the best ten models of each modality.

The audiovisual stimulus reached the highest accuracy, possibly due to its higher high triggering effect and dynamic nature, a common choice in many acquisition protocols [75, 88]. In contrast, the voice modality consistently had the lowest accuracy, possibly because participants anticipated the auditory stimuli, having indicated music choices in the acquisition form. After the audiovisual mode, the acting stimulus achieved the highest accuracy, especially in the voice signal, as it involves more pronounced expressions. Interestingly, physiological signals closely resembled the nature of "acting". Recall and visual tasks performed in the middle range. Although both are passive, the interactive nature of the visual stimulus differentiates it, as it doesn't have the same level of anticipation. Besides this, in Table D.1 it's the distribution of the labels in each type of task, that may influence these previous discussed outcomes.

Further experiments were conducted, correlating model performance with various additional metadata. These associations can be observed in figures from Figure D.1 to Figure D.6, and detailed explanations for each can be found in the respective figure captions.

6

Conclusions

Within this chapter, a concise summary of this study is provided, encapsulating its key findings and contributions. Additionally, potential avenues for future research are outlined, offering insights into possible directions for further exploration of this topic.

6.1 Main Findings

In pursuit of the established research objectives (section 1.2), beginning with the development of tools for emotion recognition, a comprehensive protocol was designed to capture emotional patterns in multiple distinct scenarios. Data collection encompassed EDA, ECG and Voice activity (audio) recordings. Diverse stimuli types were thoughtfully incorporated to generate emotion labels spanning the four quadrants of the 2D Valence-Arousal spectrum. These signals underwent meticulous preprocessing steps with state of the art methodologies, resulting in the extraction of meaningful features. Both signal and feature normalization techniques were thoroughly explored. To illustrate the effectiveness of simple ML algorithms, extensive optimization was conducted, involving both ML and initial DL experiments. The evaluation of multimodal fusion approaches, including both early and late fusion techniques, was conducted within the contexts of both (1) internally collected, where three modalities were present, and (2) external datasets, where only one or two of them were available. A four-class emotion classification problem was tackled in such experiments.

Concerning the investigation of a low-intensity scenario, it was determined that, when using Voice signals to simulate emotions, the models exhibited superior performance. In fact, the models achieved an average BA score of 69% (public) and 51% (private). In contrast, physiological signals demonstrated relatively similar performance levels. EDA yielded averaged 47% (public) and 44% (private) in BA, while ECG achieved of 48% BA(public) and 52% (private). Concerning the study of the complementary nature of these data, late fusion consistently outperformed these results, achieving 57% (public) and 60% (private) BA in a four-class classification problem for the private dataset. Notably, the 60% BA achieved in the private dataset resulted from the combination of ECG

and voice information, surpassing the performance of any other fusion involving EDA, ECG, and Voice data modalities in emotion recognition. Further experiments underscored the substantial influence of agreement between both annotators and the subject itself on subsequent emotion recognition performance. This agreement enhancement resulted in an impressive nearly 9% increase in accuracy compared to situations where agreement existed only between the subject and one annotator. Models trained with more consistent labels consistently exhibited superior performance when contrasted with those reliant on isolated labels. Moreover, an examination of stimulus types shed light on optimal model performance, indicating that participants excelled during (a) audiovisual and (b) acting tasks, suggesting the emotionally provoking nature of these stimuli. The audiovisual stimuli exhibited notable performance, achieving accuracies of 60% and 83% in the EDA and ECG signals, respectively. Conversely, the acting task yielded a commendable accuracy of 62% in the Voice modality.

In summary, this study unveils the potency of data fusion techniques, underscores the influence of audiovisual stimuli, and emphasizes the role of multiple annotators in enhancing emotion recognition. These findings collectively advance our comprehension of emotion recognition across diverse time series data.

6.2 Future Work

In the pursuit of more robust emotion recognition using EDA, ECG, and Voice signals, future research directions emerge. While promising, it must be recognized that the complexities of emotion recognition, marked by uncertainties in capturing validated emotional responses, remain. Combining different modalities enhances model performance, but more reliable assessment tools are needed to be adaptable to diverse healthcare scenarios.

Alternative approaches DL techniques (e.g., LSTM) should be explored, including pre-trained models. Enhancing feature extraction across all modalities is another avenue for improvement. Expanding analyses to encompass stimuli other than state of the art established types is an important topic that may help better understand subtle emotional patterns in real-world scenarios. Another avenue for future work is exploring early and late fusion techniques alongside DL algorithms. Finally, hybrid fusion techniques, combining early and late fusion, can be another avenue of work.

When the results become sufficiently robust, they can be applied in real biomedical applications, aiding in stress reduction, improving mental health, managing pain, and enabling early diagnosis of mental disorders.

Bibliography

[1] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. URL: https://github.com/joaomlourenco/novathesis/raw/main/template.pdf.

[2] S. Koelstra et al. "DEAP: A database for emotion analysis; Using physiological signals". In: *IEEE Transactions on Affective Computing* 3 (1 2012-01), pp. 18–31. ISSN: 19493045. DOI: 10.1109/T-AFFC.2011.15.

[3] S. Brave and C. Nass. "Emotion in Human–Computer Interaction". In: *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications* (2002-01). DOI: 10.1201/b10368-6.

[4] H. Bansal and R. Khan. "A Review Paper on Human Computer Interaction". In: *International Journal of Advanced Research in Computer Science and Software Engineering* 8 (4 2018-04), p. 53. ISSN: 22776451. DOI: 10.23956/ijarcsse.v8i4.630.

[5] S. D. Mamdiwar et al. "Recent Advances on IoT-Assisted Wearable Sensor Systems for Healthcare Monitoring". In: *Biosensors* 11.10 (2021). ISSN: 2079-6374. DOI: 10.3390/bios11100372.

[6] K. R. Scherer, A. Schorr, and T. Johnstone. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press, 2001. URL: https://psycnet.apa.org/record/2001-06810-000.

[7] P. Rashinkar and V. S. Krushnasamy. "An overview of data fusion techniques". In: *2017 International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*. 2017, pp. 694–697. DOI: 10.1109/ICIMIA.2017.7975553.

[8] H. F. Posada-Quintero and K. H. Chon. "Innovations in Electrodermal Activity Data Collection and Signal Processing: A Systematic Review". In: *Sensors* 20.2 (2020-01), p. 479. ISSN: 1424-8220. DOI: 10.3390/s20020479.

[9] A. Gautam et al. "A Data Driven Empirical Iterative Algorithm for GSR Signal Pre-Processing". In: *2018 26th European Signal Processing Conference (EUSIPCO)*. 2018, pp. 1162–1166. DOI: 10.23919/EUSIPCO.2018.8553191.

[10] G. Geršak and J. Drnovšek. "Electrodermal activity patient simulator". In: *PLOS ONE* 15 (2 2020-02), e0228949. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0228949.

[11] S. A. Hossein Aqajari et al. "pyEDA: An Open-Source Python Toolkit for Preprocessing and Feature Extraction of Electrodermal Activity". In: *Procedia Computer Science* 184 (2021), pp. 99–106. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2021.03.021.

[12] S. A. Hossein Aqajari et al. "pyEDA: An Open-Source Python Toolkit for Preprocessing and Feature Extraction of Electrodermal Activity". In: *Procedia Computer Science* 184 (2021), pp. 99–106. ISSN: 1877-0509. DOI: https://doi.org/10.1016/j.procs.2021.03.021.

[13] A. Greco et al. "cvxEDA: A Convex Optimization Approach to Electrodermal Activity Processing". In: *IEEE Transactions on Biomedical Engineering* 63.4 (2016-04), pp. 797–804. DOI: 10.1109/TBME.2015.2474131.

[14] M. Benedek and C. Kaernbach. "A continuous measure of phasic electrodermal activity". In: *J. Neurosci. Methods* 190 (2010), pp. 80–91. ISSN: 0165-0270. DOI: 10.1016/j.jneumeth.2010.06.001.

[15] M. A. Ludmer and S. Rasoul. "Cardiac Function and Dysfunction". In: *StatPearls*. StatPearls Publishing, 2021. URL: https://www.ncbi.nlm.nih.gov/books/NBK560651/.

[16] D. A. Kass et al. "Assessment of left ventricular contractile function by the pressure-volume relation in humans". In: *Circulation* 84.2 (1992), pp. 550–561. DOI: 10.1161/01.CIR.84.2.550.

[17] E. A. Ashley and J. Niebauer. *Cardiology Explained*. Remedica, 2004. URL: https://www.ncbi.nlm.nih.gov/books/NBK2204/.

[18] B. Halder, S. Mitra, and M. Mitra. "Detection and identification of ECG waves by histogram approach". In: *2016 2nd International Conference on Control, Instrumentation, Energy Communication (CIEC)*. 2016, pp. 168–172. DOI: 10.1109/CIEC.2016.7513749.

[19] T.-N. Nguyen et al. "A Deep Learning Framework for Inter-Patient ECG Classification". In: 19 (2020-03), pp. 74–84. URL: https://www.researchgate.net/publication/339884307_A_Deep_Learning_Framework_for_Inter-Patient_ECG_Classification.

[20] L. Tereshchenko and M. Josephson. "Frequency content and characteristics of ventricular conduction". In: *J Electrocardiol* 48.6 (2015-11), pp. 933–937. DOI: https://doi.org/10.1016/j.jelectrocard.2015.08.034.

[21] S. Kaplan Berkaya et al. "A survey on ECG analysis". In: *Biomedical Signal Processing and Control* 43 (2018), pp. 216–235. ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2018.03.003.

[22] A. Albarado-Ibañez et al. "Metabolic Syndrome Remodels Electrical Activity of the Sinoatrial Node and Produces Arrhythmias in Rats". In: *PLoS One* 8.11 (2013), e76534. DOI: 10.1371/journal.pone.0076534.

[23] S. A. Shufni and M. Y. Mashor. "ECG signals classification based on discrete wavelet transform, time domain and frequency domain features". In: *2015 2nd International Conference on Biomedical Engineering (ICoBE)*. 2015, pp. 1–6. DOI: 10.1109/ICoBE.2015.7235914.

[24] A. Palumbo et al. "A Novel Portable Device for Laryngeal Pathologies Analysis and Classification". In: vol. 55. 2010-01, pp. 335–352. ISBN: 978-3-642-05166-1. DOI: 10.1007/978-3-642-05167-8_19.

[25] J. S. Sobolewski. *Encyclopedia of Physical Science and Technology*. Third. Academic Press, 2001. ISBN: 978-0-12-227410-7.

[26] R. Ranjan and A. Thakur. "Analysis of Feature Extraction Techniques for Speech Recognition System". In: *International Journal of Innovative Technology and Exploring Engineering* (2019-01). ISSN: 2278-3075. URL: https://www.ijitee.org/wp-content/uploads/papers/v8i7c2/G10460587C219.pdf.

[27] P. Matias et al. "Clinically Relevant Sound-Based Features in COVID-19 Identification: Robustness Assessment With a Data-Centric Machine Learning Pipeline". In: *IEEE Access* 10 (2022), pp. 105149–105168. DOI: 10.1109/ACCESS.2022.3211295.

[28] A. Jaiswal and J. Laroche. "An Overview of Audio Signal Processing for Music Informatics". In: *The Oxford Handbook of Music and Virtuality*. Oxford University Press, 2019, pp. 131–152. DOI: 10.1093/oxfordhb/9780190468977.013.6.

[29] M. Barandas et al. "TSFEL: Time Series Feature Extraction Library". In: *SoftwareX* 11 (2020), p. 100456. ISSN: 2352-7110. DOI: https://doi.org/10.1016/j.softx.2020.100456.

[30] C. Brown et al. "Exploring Automatic Diagnosis of COVID-19 from Crowdsourced Respiratory Sound Data". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Association for Computing Machinery, 2020, pp. 3474–3484. ISBN: 9781450379984. DOI: https://doi.org/10.1145/3394486.3412865.

[31] J. L. Flanagan. *Speech Analysis, Synthesis and Perception*. 1st ed. Communication and Cybernetics. Springer Berlin, Heidelberg, 1965, pp. VIII, 317. DOI: https://doi.org/10.1007/978-3-662-00849-2.

[32] J. Delgado-Hernández et al. "Cepstral analysis of normal and pathological voice in Spanish adults. Smoothed Cepstral peak prominence in sustained vowels versus connected speech". In: *Acta Otorrinolaringologica Espanola* 69.3 (2018), pp. 134–140. DOI: 10.1016/j.otorri.2017.05.006.

[33] J. P. Teixeira, C. Oliveira, and C. Lopes. "Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters". In: *Procedia Technology* 9 (2013), pp. 1112–1122. ISSN: 2212-0173. DOI: https://doi.org/10.1016/j.protcy.2013.12.124.

[34] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. 4th. Pearson, 2021. URL: https://aima.cs.berkeley.edu/.

[35] M. I. Jordan and T. M. Mitchell. "Machine learning: Trends, perspectives, and prospects". In: *Science* 349.6245 (2015), pp. 255–260. DOI: 10.1126/science.aaa8415.

[36] L. Zhang, S. Wang, and B. Liu. "Deep Learning for Sentiment Analysis: A Survey". In: *WIREs Data Mining and Knowledge Discovery* 8 (2018), e1253. DOI: 10.1002/widm.1253.

[37] B. Mahesh. "Machine Learning Algorithms-A Review". In: *International Journal of Science and Research* (2018). ISSN: 2319-7064. DOI: 10.21275/ART20203995.

[38] K. Taunk et al. "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification". In: *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*. 2019, pp. 1255–1260. DOI: 10.1109/ICCS45141.2019.9065747.

[39] P. Xanthopoulos, P. M. Pardalos, and T. B. Trafalis. "Linear Discriminant Analysis". In: *Robust Data Mining*. Springer New York, 2013, pp. 27–33. ISBN: 978-1-4419-9878-1. DOI: 10.1007/978-1-4419-9878-1_4.

[40] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. Springer, 2009. URL: https://link.springer.com/book/10.1007/978-0-387-84858-7.

[41] Z.-H. Zhou. *Ensemble methods: Foundations and algorithms*. 1st. 2012. URL: https://tjzhifei.github.io/links/EMFA.pdf.

[42] L. Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32. ISSN: 1573-0565. DOI: https://doi.org/10.1023/A:1010933404324.

[43] G. Ke et al. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[44] Y. Freund and R. E. Schapire. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting". In: *Journal of Computer and System Sciences* 55.1 (1997), pp. 119–139. ISSN: 0022-0000. DOI: https://doi.org/10.1006/jcss.1997.1504. URL: https://www.sciencedirect.com/science/article/pii/S002200009791504X.

[45] T. Jebara. "Discriminative, generative and imitative learning". In: 2001. URL: https://api.semanticscholar.org/CorpusID:34343512.

[46] S. Pouyanfar et al. "A Survey on Deep Learning: Algorithms, Techniques, and Applications". In: *ACM Comput. Surv.* 51.5 (2018-09). ISSN: 0360-0300. DOI: 10.1145/3234150.

[47] S. Albelwi and A. Mahmood. "A Framework for Designing the Architectures of Deep Convolutional Neural Networks". In: *Entropy* 19.6 (2017), p. 242. DOI: 10.3390/e19060242.

[48] K. Bayoudh, R. Knani, F. Hamdaoui, et al. "A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets". In: *Visual Computing* 38 (2022), pp. 2939–2970. DOI: 10.1007/s00371-021-02166-7.

[49] L. van der Maaten, E. Postma, and H. Herik. "Dimensionality Reduction: A Comparative Review". In: *Journal of Machine Learning Research - JMLR* 10 (2007-01). URL: https://lvdmaaten.github.io/publications/papers/TR_Dimensionality_Reduction_Review_2009.pdf.

[50] L. McInnes, J. Healy, and J. Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: *arXiv preprint arXiv:1802.03426* (2020). URL: https://arxiv.org/abs/1802.03426.

[51] I. T. Jolliffe and J. Cadima. "Principal component analysis: a review and recent developments". In: *Phil. Trans. R. Soc. A* 374.2065 (2016), p. 20150202. DOI: 10.1098/rsta.2015.0202.

[52] L. van der Maaten and G. Hinton. "Visualizing data using t-SNE". In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605. URL: https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf?fbclid=IwAR0Bgg1eA5TFmqOZeCQXsIoL6PKrVXUFaskUKtg6yBhVXAFFvZA6yQiYx-M.

[53] M. Hancock. *What is the ROC curve?!* 2015-08. URL: http://notmatthancock.github.io/2015/08/18/what-is-the-roc-curve.html.

[54] R. W. Picard. *Affective computing*. The MIT Press, 2000. URL: https://mitpress.mit.edu/9780262661157/affective-computing/.

[55] R. W. Picard. "Does HAL cry digital tears? Emotions and computers". In: *Hal's Legacy: 2001's Computer as Dream and Reality*. The MIT Press, 1997, pp. 279–303. DOI: https://doi.org/10.7551/mitpress/3404.003.0015.

[56] "Multimodal Integration of Emotional Signals from Voice, Body, and Context: Effects of (In)Congruence on Emotion Recognition and Attitudes Towards Robots". In: *International Journal of Social Robotics* 11 (4 2019-08), pp. 555–573. ISSN: 18754805. DOI: 10.1007/s12369-019-00524-z.

[57] O. Mitruț et al. "Emotion Classification Based on Biophysical Signals and Machine Learning Techniques". In: *Symmetry* 12 (2019-12), p. 21. DOI: 10.3390/sym12010021.

[58] J. J. Braithwaite et al. *A Guide for Analysing Electrodermal Activity (EDA) Skin Conductance Responses (SCRs) for Psychological Experiments AcqKnowledge software*. 2013. URL: https://www.birmingham.ac.uk/documents/college-les/psych/saal/guide-electrodermal-activity.pdf.

[59] J. A. Domínguez-Jiménez et al. "A machine learning model for emotion recognition from physiological signals". In: *Biomedical Signal Processing and Control* 55 (2020-01). ISSN: 17468108. DOI: 10.1016/j.bspc.2019.101646.

[60] J. Vogt and E. André. "Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis". In: *EUROSPEECH 2001 Scandinavia, 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH Event*. 2001-09. DOI: 10.21437/Eurospeech.2001-34.

[61] M. Berking and P. Wupperman. "Emotion regulation and mental health: recent findings, current challenges, and future directions". In: *Curr Opin Psychiatry* 25.2 (2012), pp. 128–134. DOI: 10.1097/YCO.0b013e3283503669.

[62] E. Bal et al. "Emotion recognition in children with autism spectrum disorders: relations to eye gaze and autonomic state". In: *J Autism Dev Disord* 40.3 (2010), pp. 358–370. DOI: 10.1007/s10803-009-0884-3.

[63] L. A. Löffler et al. "Emotional dysfunctions in neurodegenerative diseases". In: *J Comp Neurol* 524.8 (2016), pp. 1727–1743. DOI: 10.1002/cne.23816. eprint: 2015Jun5.

[64] M. Dhuheir et al. "Emotion Recognition for Healthcare Surveillance Systems Using Neural Networks: A Survey". In: *2021 International Wireless Communications and Mobile Computing (IWCMC)*. 2021, pp. 681–687. DOI: 10.1109/IWCMC51323.2021.9498861.

[65] P. Adibi et al. "Emotion recognition support system: Where physicians and psychiatrists meet linguists and data engineers". In: *World J Psychiatry* 13.1 (2023), pp. 1–14. DOI: 10.5498/wjp.v13.i1.1.

[66] Y. Moriguchi and G. Komaki. "Neural mechanisms underlying Alexithymia". In: *Frontiers in Psychiatry* 10 (2019), p. 1026. DOI: 10.3389/fpsyt.2019.01026.

[67] S. Tivatansakul and M. Ohkura. "Healthcare System Focusing on Emotional Aspects Using Augmented Reality - Implementation of Breathing Control Application in Relaxation Service". In: *2013 International Conference on Biometrics and Kansei Engineering*. 2013, pp. 218–222. DOI: 10.1109/ICBAKE.2013.43.

[68] A. Zenonos et al. "HealthyOffice: Mood recognition at work using smartphones and wearable sensors". In: *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. 2016, pp. 1–6. DOI: 10.1109/PERCOMW.2016.7457166.

[69] A. S. Anusha et al. "Physiological Signal Based Work Stress Detection Using Unobtrusive Sensors". In: *Biomedical Physics & Engineering Express* 4.6 (2018-09), p. 065001. DOI: 10.1088/2057-1976/aadbd4.

[70] D. Tacconi et al. "Activity and emotion recognition to support early diagnosis of psychiatric diseases". In: *2008 Second International Conference on Pervasive Computing Technologies for Healthcare*. 2008, pp. 100–102. DOI: 10.1109/PCTHEALTH.2008.4571041.

[71] G. Pioggia et al. "Interreality: The use of advanced technologies in the assessment and treatment of psychological stress". In: *2010 10th International Conference on Intelligent Systems Design and Applications*. 2010, pp. 1047–1051. DOI: 10.1109/ISDA.2010.5687047.

[72] Y. Ma et al. "Daily Mood Assessment Based on Mobile Phone Sensing". In: *2012 Ninth International Conference on Wearable and Implantable Body Sensor Networks*. 2012, pp. 142–147. DOI: 10.1109/BSN.2012.3.

[73] N. A. A. Aziz et al. "Asian Affective and Emotional State (A2ES) Dataset of ECG and PPG for Affective Computing Research". In: *Algorithms* 16.3 (2023-02), p. 130. DOI: 10.3390/a16030130.

[74] E. Hristova, M. Grinberg, and E. Lalev. "Biosignal Based Emotion Analysis of Human-Agent Interactions". In: 2008-01, pp. 63–75. ISBN: 978-3-642-03319-3. DOI: 10.1007/978-3-642-03320-9_7.

[75] J. A. Miranda-Correa et al. "AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups". In: *IEEE Transactions on Affective Computing* 12.2 (2021), pp. 479–493. DOI: 10.1109/TAFFC.2018.2884461.

[76] A. Dessai and H. Virani. "Emotion Classification Based on CWT of ECG and GSR Signals Using Various CNN Models". In: *Electronics* 12.13 (2023). ISSN: 2079-9292. DOI: 10.3390/electronics12132795.

[77] A. Rahim et al. "Emotion Charting Using Real-time Monitoring of Physiological Signals". In: *2019 International Conference on Robotics and Automation in Industry (ICRAI)*. 2019, pp. 1–5. DOI: 10.1109/ICRAI47710.2019.8967398.

[78] M. Ali et al. "A Globally Generalized Emotion Recognition System Involving Different Physiological Signals". In: *Sensors* 18.6 (2018-06), p. 1905. DOI: 10.3390/s18061905.

[79] E. H. Jang et al. "Emotion classification based on bio-signals emotion recognition using machine learning algorithms". In: vol. 3. Institute of Electrical and Electronics Engineers Inc., 2014-11, pp. 1373–1376. ISBN: 9781479931965. DOI: 10.1109/InfoSEEE.2014.6946144.

[80] J. Xie, X. Xu, and L. Shu. "WT Feature Based Emotion Recognition from Multichannel Physiological Signals with Decision Fusion". In: *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)* (2018), pp. 1–6.

[81] W. He et al. "Online Cross-subject Emotion Recognition from ECG via Unsupervised Domain Adaptation". In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*. 2021, pp. 1001–1005. DOI: 10.1109/EMBC46164.2021.9630433.

[82] A. Sepúlveda et al. "Emotion Recognition from ECG Signals Using Wavelet Scattering and Machine Learning". In: *Applied Sciences* 11.11 (2021). ISSN: 2076-3417. DOI: 10.3390/app11114945.

[83] L. Christ et al. *Towards Multimodal Prediction of Spontaneous Humour: A Novel Dataset and First Results*. 2022-09. URL: https://arxiv.org/abs/2209.14272.

[84] B. Sun et al. "Combining Multimodal Features with Hierarchical Classifier Fusion for Emotion Recognition in the Wild". In: *Proceedings of the 16th International Conference on Multimodal Interaction*. ICMI '14. Istanbul, Turkey: Association for Computing Machinery, 2014, pp. 481–486. ISBN: 9781450328852. DOI: 10.1145/2663204.2666272.

[85] R. G. Bastos Germano et al. *emoUERJ: an emotional speech database in Portuguese (1.0.0)*. Version 1.0.0. Zenodo, 2021. DOI: 10.5281/zenodo.5427549.

[86] F. Luthman. "Multilingual Speech Emotion Recognition using pretrained models powered by Self-Supervised Learning". MA thesis. KTH, School of Electrical Engineering and Computer Science (EECS), 2022. URL: https://kth.diva-portal.org/smash/get/diva2:1650791/FULLTEXT01.pdf.

[87] J.-L. Duret, T. Parcollet, and Y. Estève. "Learning Multilingual Expressive Speech Representation for Prosody Prediction without Parallel Data". In: *arXiv preprint arXiv:2306.17199* (2023). URL: https://arxiv.org/abs/2306.17199.

[88] M. U. Akram, M. N. Dar, and A. Rahim. *Young Adult's Affective Data (YAAD) Using ECG and GSR Signals*. Mendeley Data, V4. 2022. DOI: 10.17632/g2p7vwxyn2.4.

[89] R. Ferro and P. Pestana. *European Portuguese Emotional Discourse Database*. Porto: Universidade Católica Portuguesa, Escola das Artes. 2017. URL: http://hdl.handle.net/10400.14/24156.

[90] P. Virtanen et al. "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python". In: *Nature Methods* 17 (2020), pp. 261–272. DOI: 10.1038/s41592-019-0686-2.

[91] C. R. Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (2020-09), pp. 357–362. DOI: 10.1038/s41586-020-2649-2.

[92] B. McFee et al. "librosa: Audio and music signal analysis in Python". In: *Proceedings of the 14th Python in Science Conference*. 2015, pp. 18–25. URL: https://conference.scipy.org/proceedings/scipy2015/pdfs/brian_mcfee.pdf.

[93] D. Makowski et al. "NeuroKit2: A Python toolbox for neurophysiological signal processing". In: *Behavior Research Methods* 53.4 (2021-02), pp. 1689–1696. DOI: 10.3758/s13428-020-01516-y.

[94] P. van Gent et al. "HeartPy: A novel heart rate algorithm for the analysis of noisy signals". In: *Transportation Research Part F: Traffic Psychology and Behaviour* 66 (2019), pp. 368–378. ISSN: 1369-8478. DOI: https://doi.org/10.1016/j.trf.2019.09.015.

[95] C. Carreiras et al. *BioSPPy: Biosignal Processing in Python*. [Online; accessed]. 2015–. URL: https://github.com/PIA-Group/BioSPPy/.

[96] T. pandas development team. *pandas-dev/pandas: Pandas*. Version latest. 2020-02. DOI: 10.5281/zenodo.3509134.

[97] Arundo Analytics, Inc. *tsaug*. https://github.com/arundo/tsaug. 2020.

[98] I. Jordal et al. "iver56/audiomentations: v0.26.0". Version v0.26.0. In: (2022-08). DOI: 10.5281/zenodo.7010042.

[99] M. L. Waskom. "seaborn: statistical data visualization". In: *Journal of Open Source Software* 6.60 (2021), p. 3021. DOI: 10.21105/joss.03021.

[100] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in Science & Engineering* 9.3 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55.

[101] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830. URL: https://scikit-learn.org/stable/.

[102] G. Ke et al. "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in neural information processing systems* 30 (2017), pp. 3146–3154. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[103] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015. URL: https://www.tensorflow.org/.

[104] F. Chollet et al. *Keras*. 2015. URL: https://github.com/fchollet/keras.

[105] A. Alam, S. Urooj, and A. Q. Ansari. "Human Emotion Recognition Models Using Machine Learning Techniques". In: *2023 International Conference on Recent Advances in Electrical, Electronics Digital Healthcare Technologies (REEDCON)*. 2023, pp. 329–334. DOI: 10.1109/REEDCON57544.2023.10151406.

[106] H.-W. Guo et al. "Heart Rate Variability Signal Features for Emotion Recognition by Using Principal Component Analysis and Support Vectors Machine". In: *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*. 2016, pp. 274–277. DOI: 10.1109/BIBE.2016.40.

[107] S. Kumar and S. Thiruvenkadam. "An Analysis of the Impact of Spectral Contrast Feature in Speech Emotion Recognition". In: *International Journal of Recent Contributions from Engineering, Science amp; IT (iJES)* 9.2 (2021-06), pp. 87–95. DOI: 10.3991/ijes.v9i2.22983.

A

Methodology

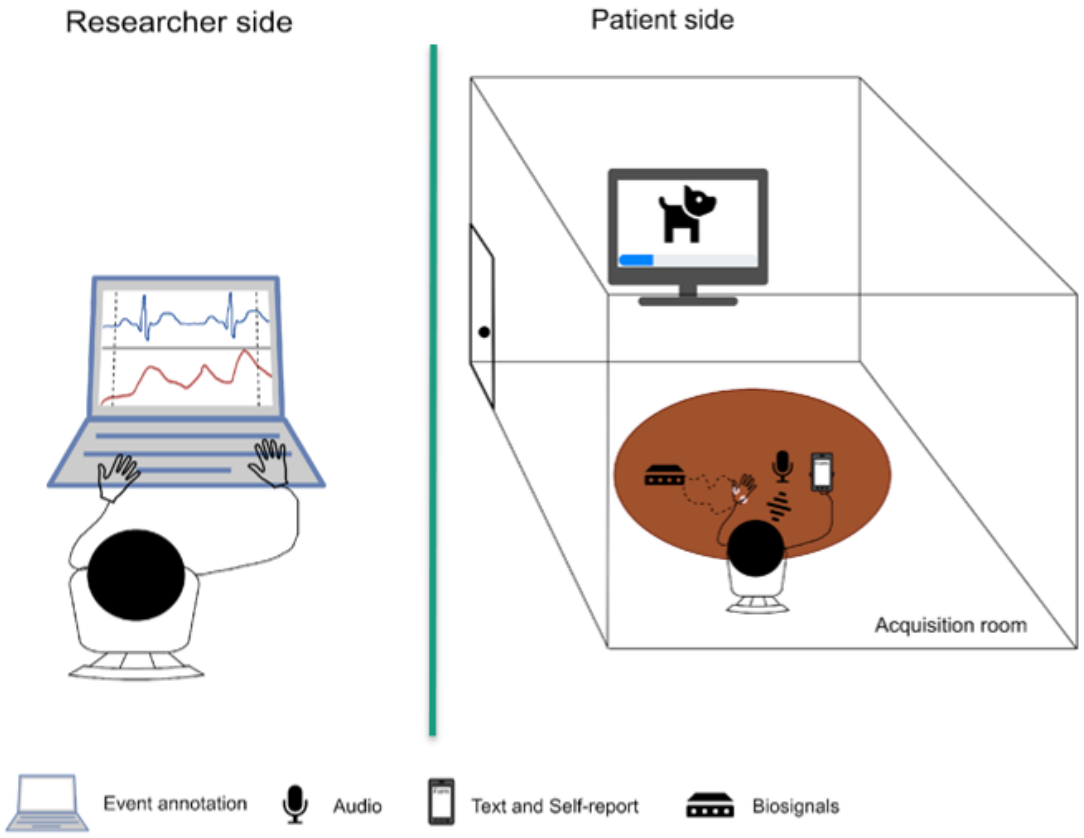
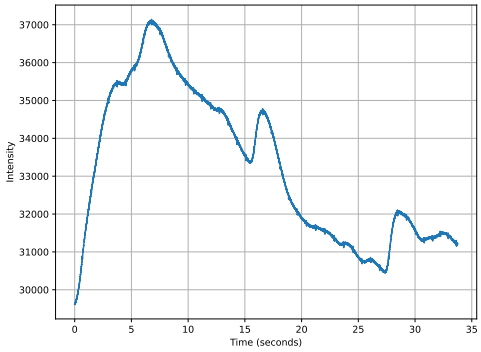


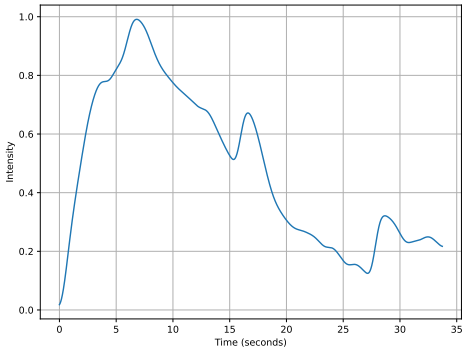
Figure A.2: Illustration of the data acquisition setup, which depicts the research environment, the patient's perspective, and the equipment employed.

Table A.1: Training parameters of the DL algorithms.

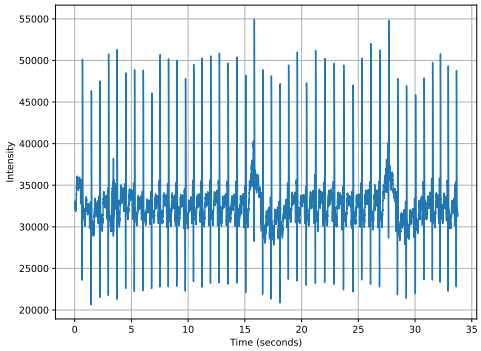
Loss Function	Optimizer			Epochs	Batch Size
	Type	Learning Rate	Weight Decay		
Categorical Cross Entropy	SGD	1e-4	1e-5	200	8 (EDA 32 (ECG and Voice)



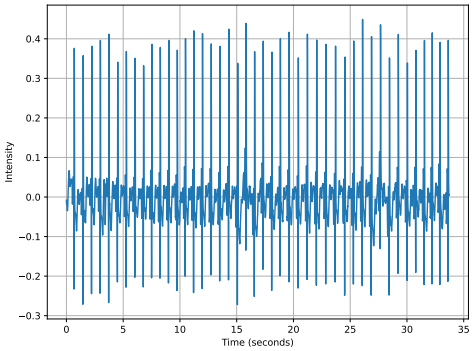
(a) EDA signal before.



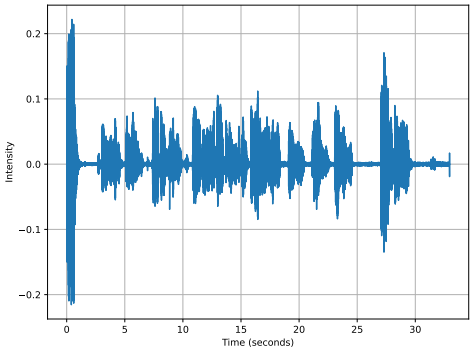
(b) EDA signal after.



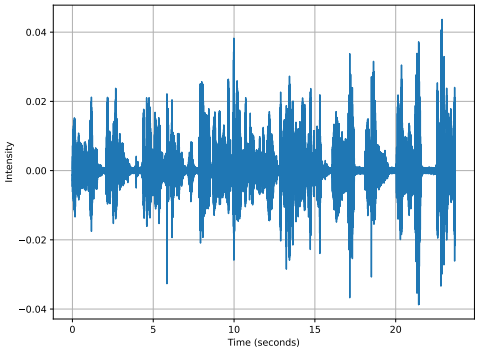
(c) ECG signal before.



(d) ECG signal after.



(e) Voice signal before.

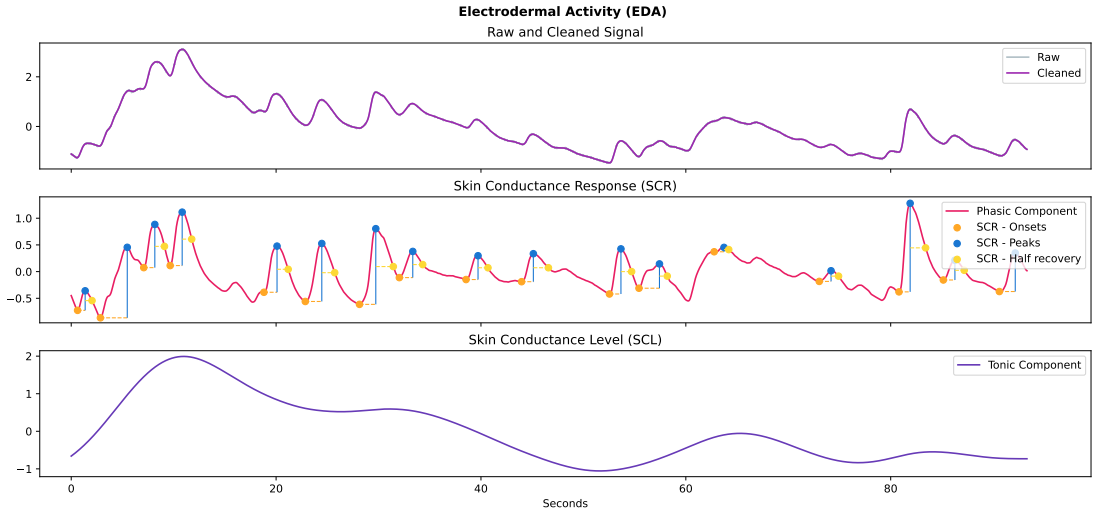


(f) Voice signal after.

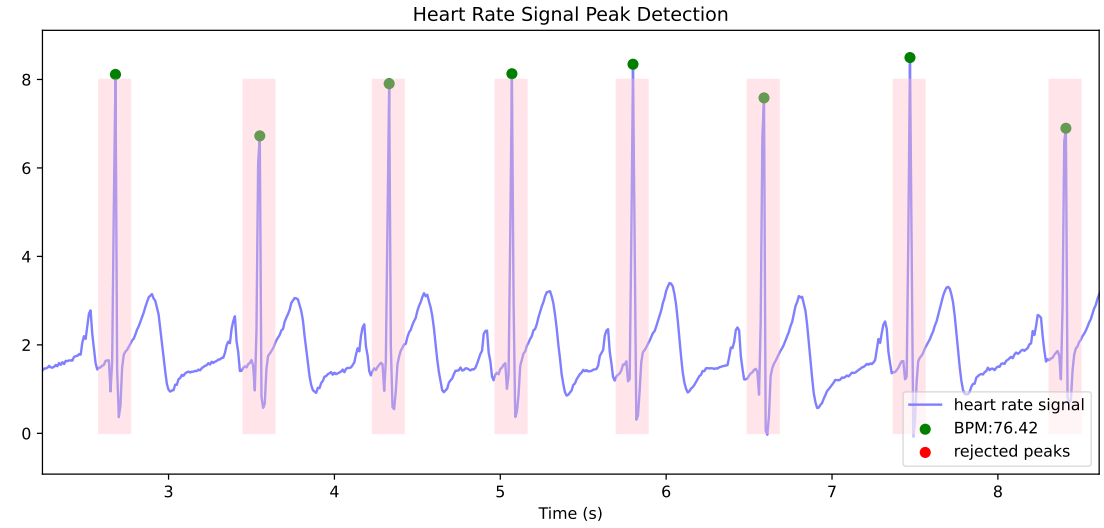
Figure A.3: Signals before and after pre-processing. (a) EDA signal before the pre-processing, (b) EDA signal after the pre-processing, (c) ECG signal before the pre-processing, (d) ECG signal after the pre-processing, (e) Voice signal before the pre-processing, (f) Voice signal after the pre-processing .

Table A.2: Architecture of the DL algorithms.

Modality	Nº Layers	Kernel Size	Nº Filters	Activation Function	Nº Parameters	Pooling Size
EDA and ECG	17	32, 16	128, 64, 32, 16	tanh, softmax	180372	4
Voice	17	8, 4	128, 64, 32, 16	tanh, softmax	220308	2

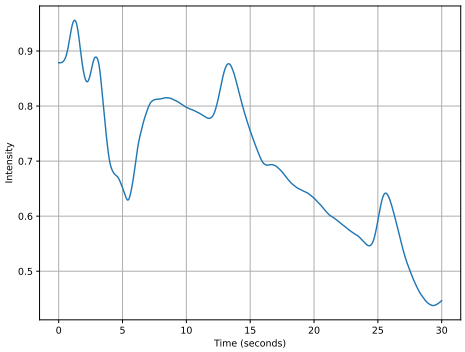


(a) EDA specific processing.

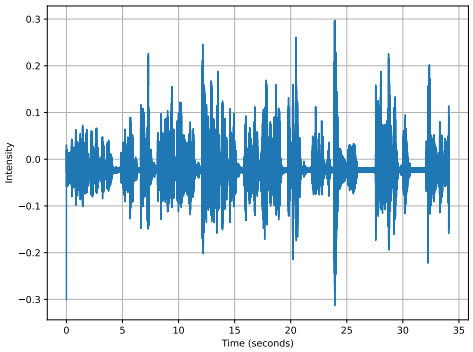


(b) ECG specific processing.

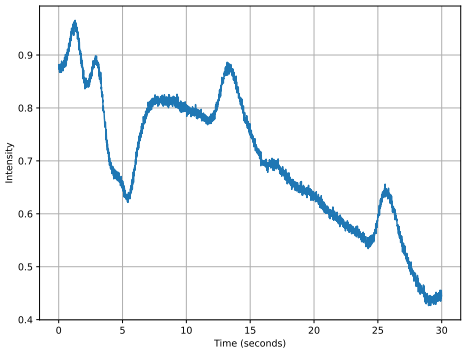
Figure A.4: Specific processing steps preceding feature extraction. Relevant signal regions are identified for (a) EDA signal, showcasing components like SCL and SCR with identified onsets, peaks, and half-recovery, and (b) the ECG signal, highlighting the QRS complex and R peak, alongside the computed BPM.



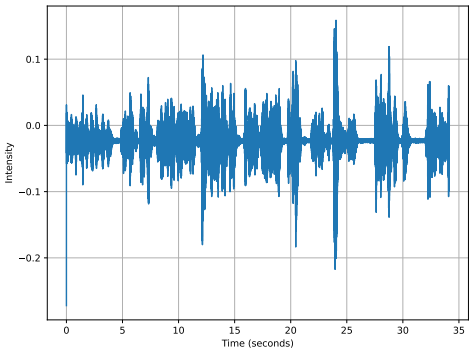
(a) EDA before.



(b) Voice before.



(c) EDA after.



(d) Voice after.

Figure A.5: Signals before and after augmentation, corresponding to (a) the EDA signal before augmentation, (b) the Voice signal before augmentation, (c) the EDA signal after the application of the AddNoise function from the tsaug package, with a scaling factor of 0.01, and (d) Voice signal after the application of PitchShift with min_semitones of -5 and max_semitones of 5 from the audiomentations package.

B

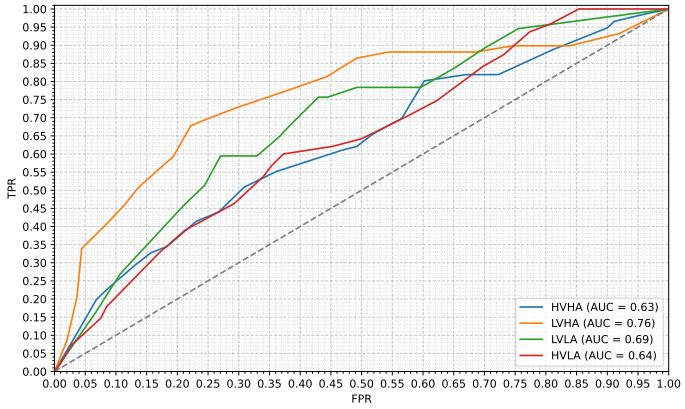
Unimodal Emotion Recognition

Table B.1: Summary of the results achieved by the best models and parameters on the ICANS dataset using the EDA signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

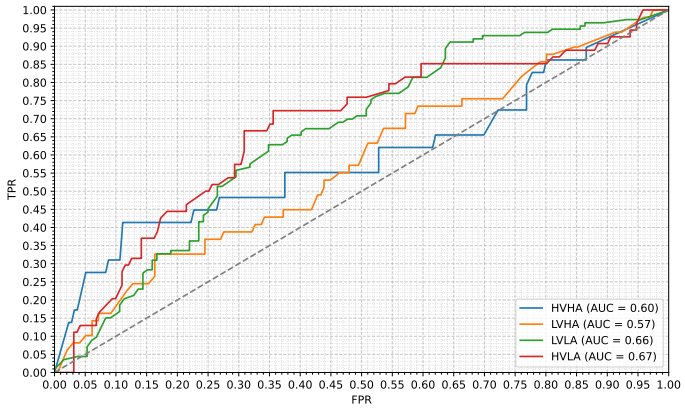
Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
DT	Yes	M	SBN	307	0.40 ± 0.01	**0.44 ± 0.01**	0.47 ± 0.02	0.51 ± 0.02
DT	No	S	FCN	307	**0.41 ± 0.04**	0.43 ± 0.04	0.47 ± 0.00	0.49 ± 0.01
DT	No	M	SBN	307	**0.41 ± 0.02**	0.42 ± 0.01	0.46 ± 0.02	0.49 ± 0.01
DT	No	M	FCN	307	0.40 ± 0.02	0.43 ± 0.01	0.55 ± 0.01	0.58 ± 0.01
DT	Yes	M	RN	297	**0.41 ± 0.03**	0.42 ± 0.03	0.47 ± 0.03	0.48 ± 0.02
DT	No	M	BN	300	0.40 ± 0.03	0.42 ± 0.03	0.46 ± 0.02	0.49 ± 0.04
RF	Yes	M	RN	297	0.39 ± 0.06	0.43 ± 0.06	0.74 ± 0.02	0.75 ± 0.02
LGBM	Yes	S	RN	299	0.38 ± 0.05	0.43 ± 0.05	0.43 ± 0.03	0.48 ± 0.02
DT	Yes	M	BN	300	0.39 ± 0.01	0.41 ± 0.01	0.55 ± 0.02	0.57 ± 0.01
DT	No	S	SBN	307	0.36 ± 0.02	0.41 ± 0.04	0.42 ± 0.01	0.48 ± 0.01

Table B.2: Summary of the results obtained on the YAAD dataset using the best models and parameters from the ICANS dataset with the EDA signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

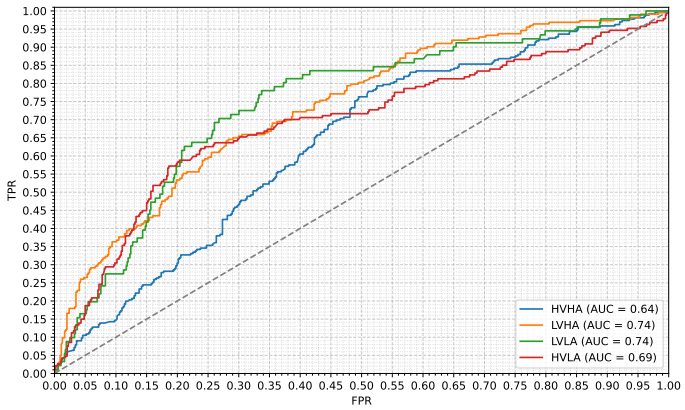
Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
DT	No	M	SBN	245	**0.32 ± 0.03**	**0.36 ± 0.03**	0.51 ± 0.06	0.58 ± 0.03
DT	Yes	M	SBN	245	0.31 ± 0.06	0.32 ± 0.06	0.54 ± 0.01	0.60 ± 0.02
DT	No	M	FCN	245	0.30 ± 0.02	0.33 ± 0.03	0.41 ± 0.02	0.49 ± 0.02
DT	No	S	FCN	245	0.23 ± 0.04	0.28 ± 0.05	0.37 ± 0.02	0.47 ± 0.02
DT	No	S	SBN	245	0.22 ± 0.06	0.28 ± 0.06	0.30 ± 0.09	0.43 ± 0.03



(a) ICANS dataset.

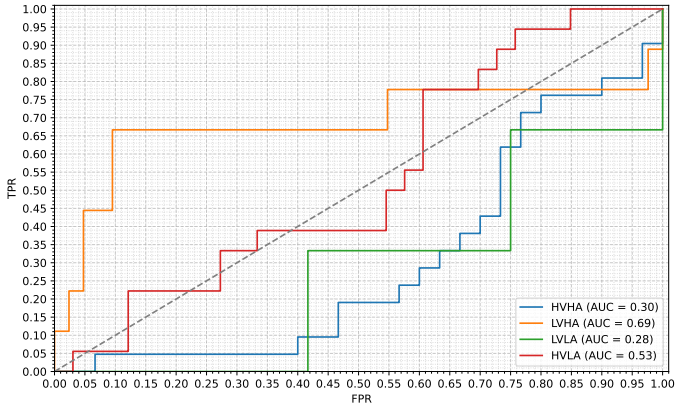


(b) YAAD dataset.

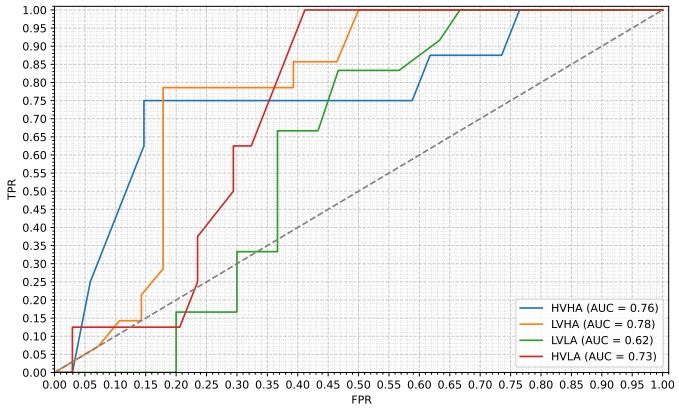


(c) AMIGOS dataset.

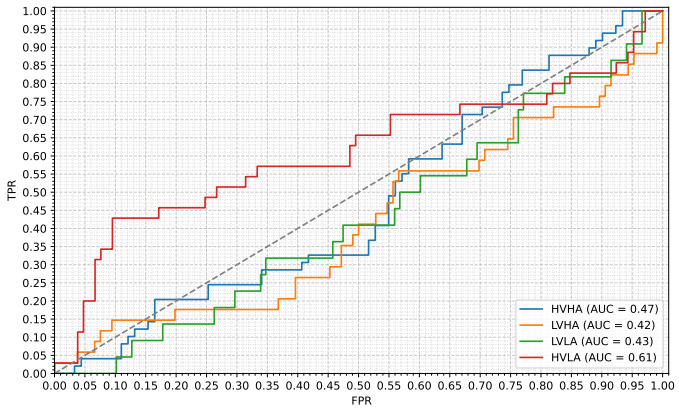
Figure B.1: ROC curves for the best models in each the ML approach, in all datasets, using the EDA signal. (a) ROC curve for the ICANS dataset, (b) ROC curve forthe YAAD dataset, (c) ROC curve for the AMIGOS dataset.



(a) ICANS dataset.

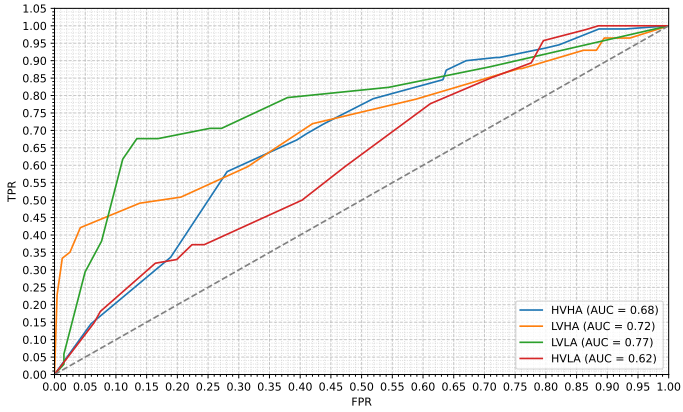


(b) YAAD dataset.

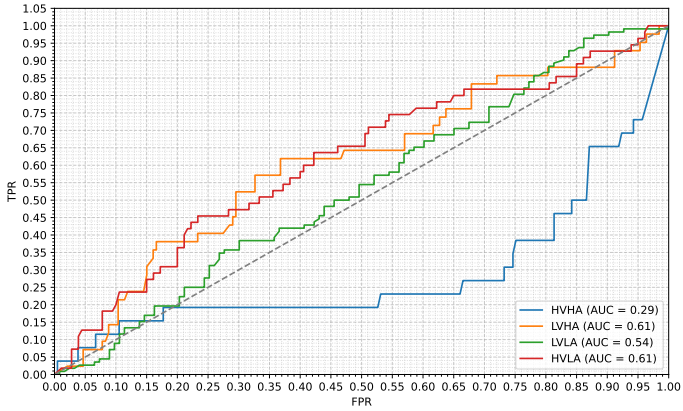


(c) AMIGOS dataset.

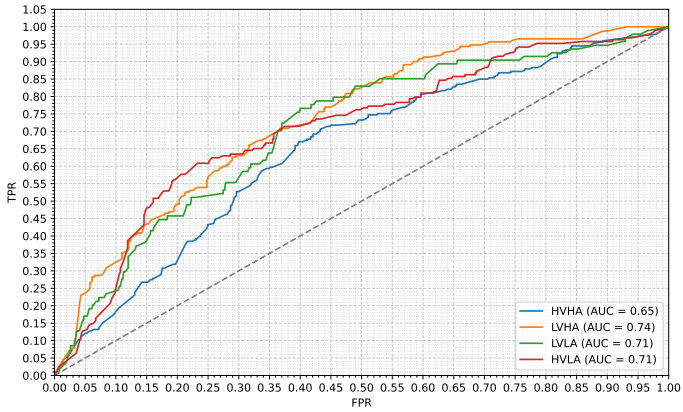
Figure B.2: ROC curves for the best models in each the DL approach, in all datasets, using the EDA signal. (a) ROC curve for the ICANS dataset, (b) ROC curve forthe YAAD dataset, (c) ROC curve for the AMIGOS dataset.



(a) ICANS dataset.

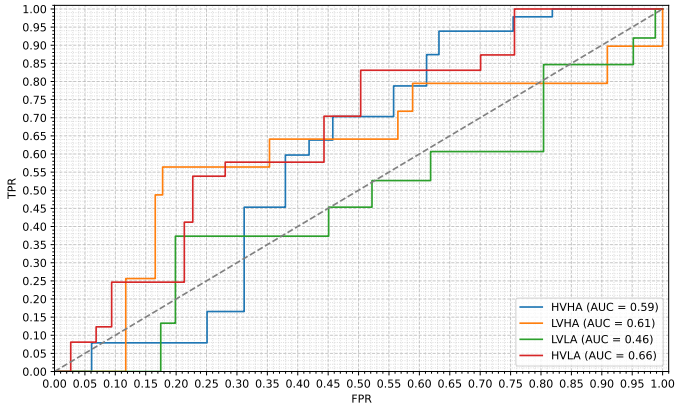


(b) YAAD dataset.

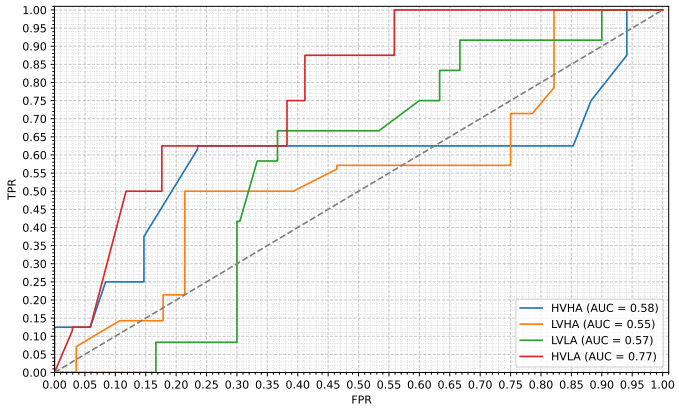


(c) AMIGOS dataset.

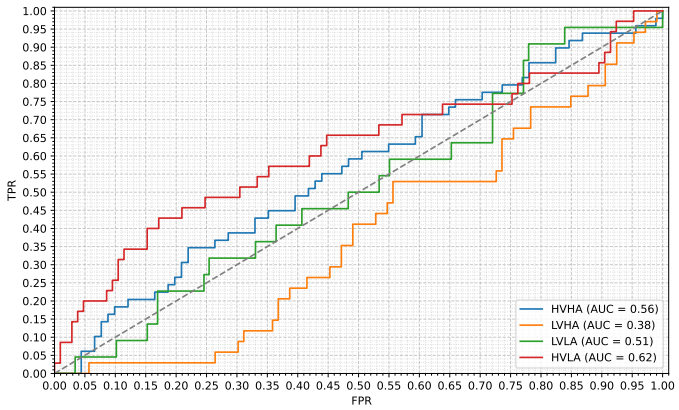
Figure B.3: ROC curves for the best models in each the ML approach, in all datasets, using the ECG signal. (a) ROC curve for the ICANS dataset, (b) ROC curve forthe YAAD dataset, (c) ROC curve for the AMIGOS dataset.



(a) ICANS dataset.

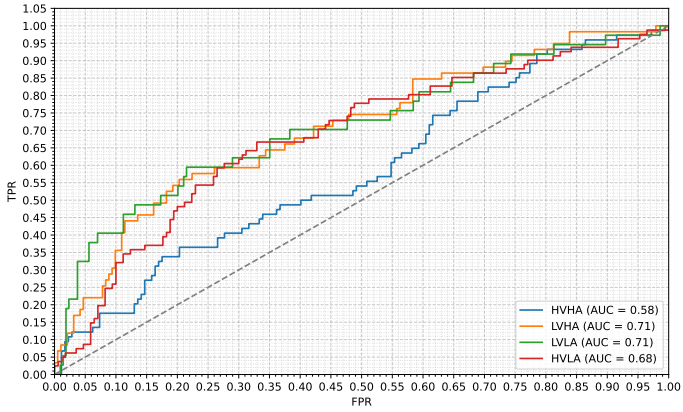


(b) YAAD dataset.

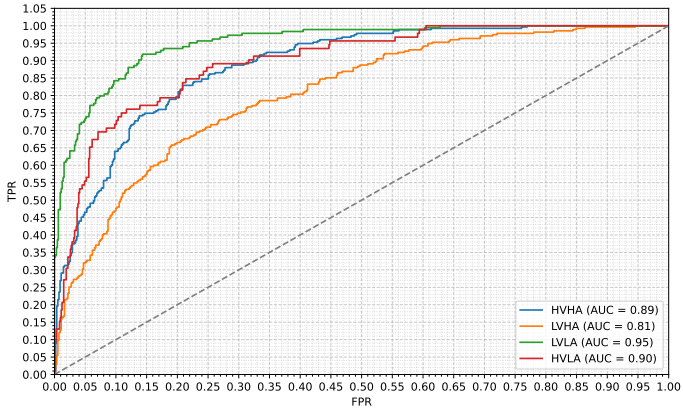


(c) AMIGOS dataset.

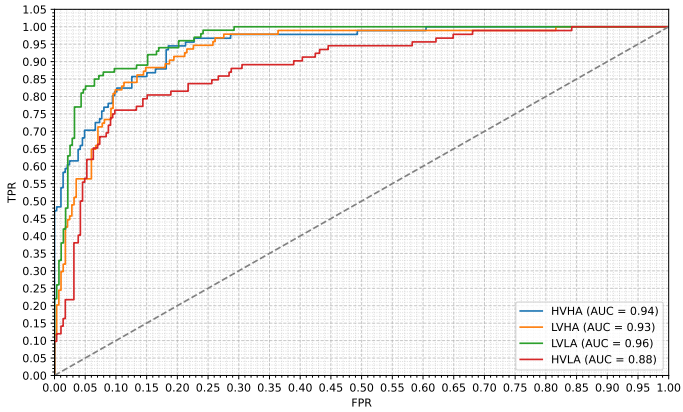
Figure B.4: ROC curves for the best models in each the DL approach, in all datasets, using the ECG signal. (a) ROC curve for the ICANS dataset, (b) ROC curve forthe YAAD dataset, (c) ROC curve for the AMIGOS dataset.



(a) ICANS dataset.

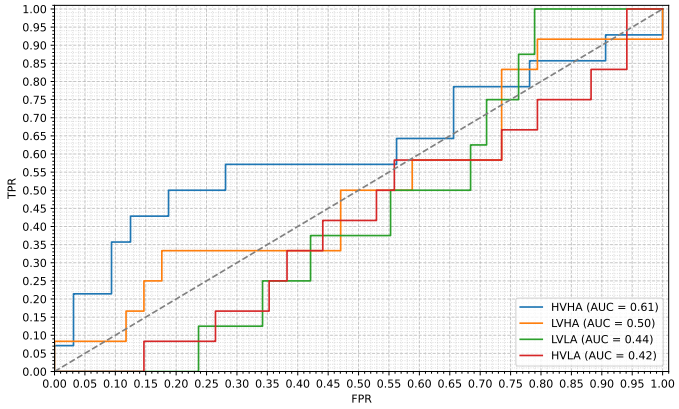


(b) EPEDD dataset.

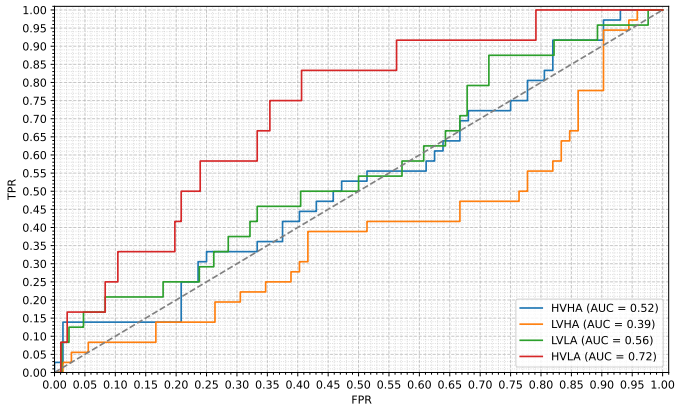


(c) EMOUERJ dataset.

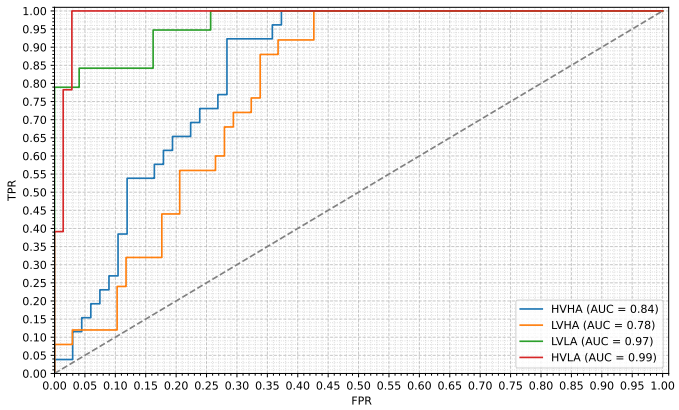
Figure B.5: ROC curves for the best models in each the ML approach, in all datasets, using the Voice signal. (a) ROC curve for the ICANS dataset, (b) EPEDD curve forthe YAAD dataset, (c) EMOUERJ curve for the AMIGOS dataset.



(a) ICANS dataset.



(b) EPEDD dataset.



(c) EMOUERJ dataset.

Figure B.6: ROC curves for the best models in each the DL approach, in all datasets, using the Voice signal. (a) ROC curve for the ICANS dataset, (b) EPEDD curve forthe YAAD dataset, (c) EMOUERJ curve for the AMIGOS dataset.

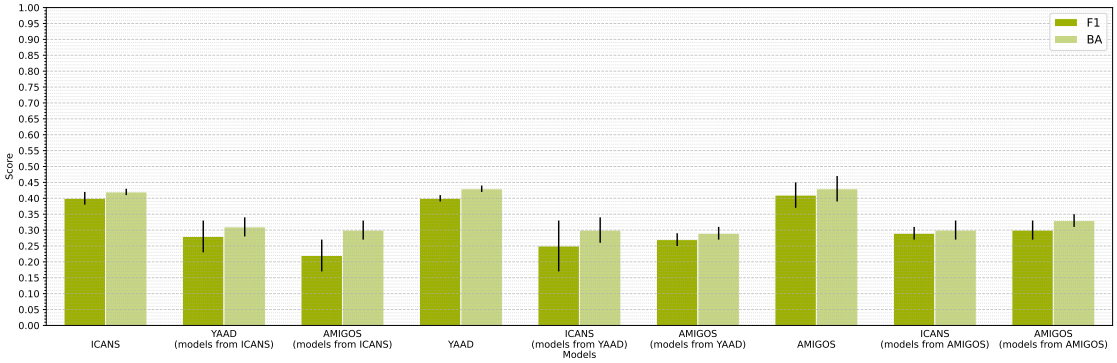


Figure B.7: Results in the models of public and private datasets, regarding the EDA signal. For each dataset, the results are shown sequentialy: (a) best performances in a dataset, (b) performance obtained in the remaining datasets using pipelines chosen during the train of the first dataset.

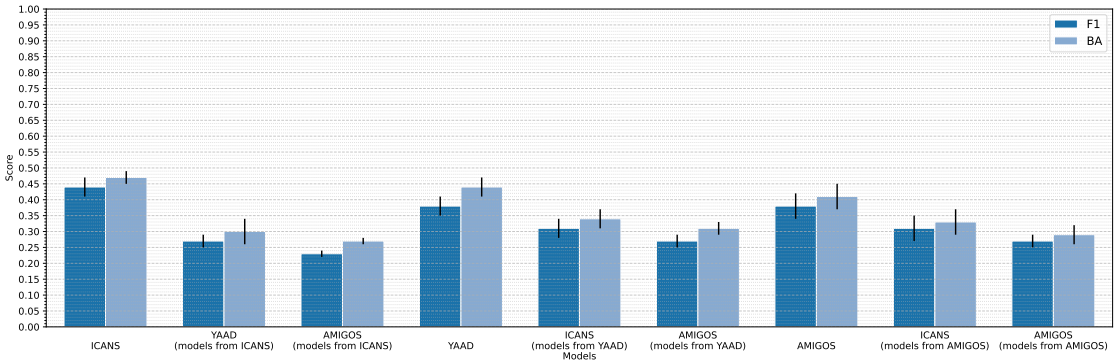


Figure B.8: Results in the models of public and private datasets, regarding the ECG signal. For each dataset, the results are shown sequentialy: (a) best performances in a dataset, (b) performance obtained in the remaining datasets using pipelines chosen during the train of the first dataset.

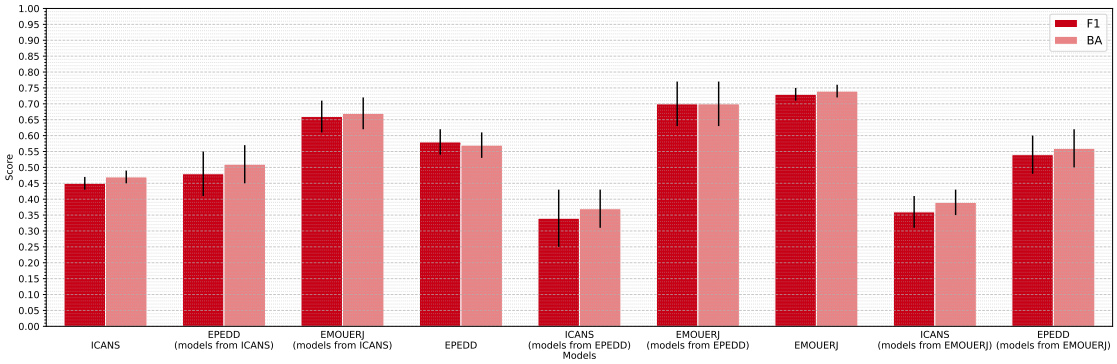


Figure B.9: Results in the models of public and private datasets, regarding the Voice signal. For each dataset, the results are shown sequentialy: (a) best performances in a dataset, (b) performance obtained in the remaining datasets using pipelines chosen during the train of the first dataset.

Table B.3: Summary of the results obtained on the AMIGOS dataset using the best models and parameters from the ICANS dataset with the EDA signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
DT	No	M	FCN	767	**0.29 ± 0.03**	**0.35 ± 0.05**	0.37 ± 0.02	0.55 ± 0.01
DT	Yes	M	SBN	767	0.24 ± 0.02	0.32 ± 0.03	0.34 ± 0.02	0.51 ± 0.03
DT	No	S	FCN	767	0.22 ± 0.02	0.27 ± 0.01	0.30 ± 0.04	0.43 ± 0.02
DT	No	M	SBN	767	0.19 ± 0.06	0.29 ± 0.07	0.28 ± 0.03	0.47 ± 0.03
DT	No	S	SBN	767	0.16 ± 0.01	0.27 ± 0.04	0.23 ± 0.01	0.43 ± 0.01

Table B.4: Summary of the results achieved by the best models and parameters on the YAAD dataset using the EDA signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
RF	No	R	SBN	245	0.39 ± 0.02	**0.45 ± 0.03**	0.57 ± 0.01	0.62 ± 0.02
RF	No	R	FCN	245	0.39 ± 0.03	**0.45 ± 0.04**	0.44 ± 0.04	0.49 ± 0.03
RF	No	M	SBN	245	**0.40 ± 0.08**	0.42 ± 0.11	0.39 ± 0.02	0.46 ± 0.03
RF	No	S	SBN	245	**0.40 ± 0.01**	0.42 ± 0.01	0.62 ± 0.02	0.65 ± 0.02
RF	Yes	M	FCN	245	**0.40 ± 0.06**	0.42 ± 0.09	0.45 ± 0.02	0.50 ± 0.03
RF	Yes	S	SBN	245	**0.40 ± 0.08**	0.42 ± 0.11	0.43 ± 0.07	0.50 ± 0.04
RF	Yes	R	FCN	245	**0.40 ± 0.06**	0.42 ± 0.09	0.45 ± 0.03	0.50 ± 0.03
RF	Yes	R	SBN	245	**0.40 ± 0.06**	0.42 ± 0.07	0.61 ± 0.02	0.63 ± 0.03
RF	No	M	FCN	245	0.39 ± 0.02	0.42 ± 0.07	0.42 ± 0.02	0.47 ± 0.01
RF	No	S	FCN	245	0.39 ± 0.03	0.42 ± 0.08	0.42 ± 0.03	0.47 ± 0.01

Table B.5: Summary of the results obtained on the ICANS dataset using the best models and parameters from the YAAD dataset with the EDA signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
RF	Yes	M	FCN	307	**0.32 ± 0.06**	0.34 ± 0.07	0.63 ± 0.02	0.68 ± 0.02
RF	No	M	FCN	307	**0.32 ± 0.04**	0.33 ± 0.05	0.56 ± 0.04	0.59 ± 0.02
RF	No	S	SBN	307	0.29 ± 0.02	0.33 ± 0.03	0.56 ± 0.01	0.59 ± 0.03
RF	No	S	FCN	307	0.31 ± 0.06	0.31 ± 0.06	0.53 ± 0.00	0.56 ± 0.00
RF	No	M	SBN	307	0.25 ± 0.03	**0.35 ± 0.01**	0.33 ± 0.04	0.44 ± 0.03
RF	Yes	R	FCN	307	0.29 ± 0.03	0.30 ± 0.04	0.60 ± 0.03	0.64 ± 0.03
RF	Yes	R	SBN	307	0.27 ± 0.02	0.29 ± 0.04	0.54 ± 0.03	0.58 ± 0.03
RF	Yes	S	SBN	307	0.23 ± 0.04	0.30 ± 0.03	0.32 ± 0.05	0.41 ± 0.03
RF	No	R	SBN	307	0.13 ± 0.01	0.25 ± 0.00	0.13 ± 0.01	0.25 ± 0.00
RF	No	R	FCN	307	0.10 ± 0.03	0.24 ± 0.01	0.14 ± 0.01	0.26 ± 0.01

Table B.6: Summary of the results obtained on the AMIGOS dataset using the best models and parameters from the YAAD dataset with the EDA signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
RF	Yes	M	FCN	767	**0.31 ± 0.01**	0.32 ± 0.01	0.47 ± 0.05	0.59 ± 0.02
RF	No	R	FCN	767	0.29 ± 0.02	**0.33 ± 0.02**	0.39 ± 0.00	0.51 ± 0.02
RF	Yes	R	SBN	767	0.29 ± 0.01	0.32 ± 0.04	0.41 ± 0.01	0.52 ± 0.01
RF	No	M	FCN	767	0.30 ± 0.02	0.31 ± 0.04	0.44 ± 0.03	0.55 ± 0.02
RF	Yes	S	SBN	767	0.28 ± 0.02	0.29 ± 0.03	0.44 ± 0.02	0.54 ± 0.03
RF	Yes	R	FCN	767	0.28 ± 0.01	0.29 ± 0.03	0.44 ± 0.02	0.56 ± 0.02
RF	No	S	FCN	767	0.27 ± 0.01	0.30 ± 0.02	0.36 ± 0.06	0.50 ± 0.01
RF	No	R	SBN	767	0.23 ± 0.02	0.27 ± 0.03	0.34 ± 0.02	0.44 ± 0.02
RF	No	M	SBN	767	0.22 ± 0.03	0.26 ± 0.01	0.31 ± 0.01	0.44 ± 0.03
RF	No	S	SBN	767	0.24 ± 0.02	0.23 ± 0.02	0.42 ± 0.02	0.53 ± 0.01

Table B.7: Summary of the results achieved by the best models and parameters on the AMIGOS dataset using the EDA signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
RF	Yes	No	No	767	**0.45 ± 0.00**	**0.49 ± 0.01**	0.53 ± 0.01	0.56 ± 0.01
DT	No	No	No	767	0.44 ± 0.02	0.48 ± 0.01	0.49 ± 0.01	0.51 ± 0.01
RF	Yes	No	No	767	0.44 ± 0.02	0.47 ± 0.02	0.51 ± 0.01	0.55 ± 0.01
DT	No	No	No	767	0.43 ± 0.01	0.45 ± 0.02	0.48 ± 0.01	0.50 ± 0.01
KNN	Yes	No	No	767	0.43 ± 0.01	0.43 ± 0.01	0.46 ± 0.02	0.46 ± 0.02
LGBM	Yes	No	No	767	0.42 ± 0.00	0.44 ± 0.01	0.69 ± 0.01	0.73 ± 0.01
LGBM	No	No	No	767	0.42 ± 0.02	0.43 ± 0.02	0.65 ± 0.02	0.68 ± 0.01
AB	No	No	No	767	0.37 ± 0.03	0.39 ± 0.04	0.54 ± 0.02	0.57 ± 0.02
AB	No	No	No	767	0.37 ± 0.03	0.38 ± 0.03	0.72 ± 0.02	0.72 ± 0.01
RF	No	M	FCN	767	0.35 ± 0.00	0.38 ± 0.00	0.55 ± 0.03	0.63 ± 0.03

Table B.8: Summary of the results obtained on the ICANS dataset using the best models and parameters from the AMIGOS dataset with the EDA signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
RF	Yes	No	No	307	**0.33 ± 0.01**	**0.34 ± 0.01**	0.62 ± 0.01	0.65 ± 0.02
KNN	Yes	No	No	307	0.32 ± 0.01	**0.34 ± 0.02**	0.39 ± 0.02	0.40 ± 0.01
AB	No	No	No	307	0.30 ± 0.02	0.32 ± 0.01	0.57 ± 0.02	0.57 ± 0.02
DT	No	No	No	307	0.29 ± 0.01	0.32 ± 0.01	0.46 ± 0.04	0.51 ± 0.01
LGBM	No	No	No	307	0.29 ± 0.01	0.30 ± 0.01	0.68 ± 0.01	0.73 ± 0.01
RF	No	M	FCN	307	0.29 ± 0.03	0.30 ± 0.03	0.67 ± 0.03	0.71 ± 0.03
RF	Yes	No	No	307	0.28 ± 0.03	0.29 ± 0.03	0.66 ± 0.01	0.70 ± 0.01
LGBM	Yes	No	No	307	0.28 ± 0.03	0.28 ± 0.03	0.68 ± 0.01	0.73 ± 0.01
DT	No	No	No	307	0.25 ± 0.03	0.29 ± 0.01	0.46 ± 0.05	0.53 ± 0.02
AB	No	No	No	307	0.26 ± 0.06	0.26 ± 0.05	0.76 ± 0.01	0.76 ± 0.03

Table B.9: Summary of the results obtained on the YAAD dataset using the best models and parameters from the AMIGOS dataset with the EDA signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
RF	Yes	No	No	245	**0.34 ± 0.03**	**0.36 ± 0.05**	0.46 ± 0.01	0.51 ± 0.02
DT	No	No	No	245	0.33 ± 0.05	0.35 ± 0.06	0.44 ± 0.02	0.50 ± 0.03
LGBM	Yes	No	No	245	0.32 ± 0.01	0.34 ± 0.05	0.44 ± 0.00	0.51 ± 0.02
RF	Yes	No	No	245	0.32 ± 0.03	0.33 ± 0.04	0.47 ± 0.00	0.50 ± 0.02
LGBM	No	No	No	245	0.30 ± 0.03	0.33 ± 0.03	0.45 ± 0.03	0.51 ± 0.02
RF	No	M	FCN	245	0.31 ± 0.05	0.32 ± 0.04	0.46 ± 0.01	0.51 ± 0.02
AB	No	No	No	245	0.29 ± 0.01	0.32 ± 0.05	0.40 ± 0.01	0.49 ± 0.02
AB	No	No	No	245	0.29 ± 0.03	0.32 ± 0.05	0.43 ± 0.01	0.49 ± 0.03
DT	No	No	No	245	0.28 ± 0.04	0.32 ± 0.03	0.41 ± 0.02	0.49 ± 0.02
KNN	Yes	No	No	245	0.24 ± 0.03	0.28 ± 0.03	0.34 ± 0.03	0.39 ± 0.02

Table B.10: Summary of the results achieved by the best models and parameters on the ICANS dataset using the ECG signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
DT	Yes	S	FCN	295	**0.51 ± 0.01**	**0.52 ± 0.01**	0.51 ± 0.03	0.53 ± 0.03
DT	Yes	S	RN	297	0.44 ± 0.01	0.49 ± 0.01	0.52 ± 0.01	0.57 ± 0.01
DT	No	S	RN	296	0.44 ± 0.02	0.49 ± 0.02	0.48 ± 0.04	0.53 ± 0.02
DT	Yes	M	FCN	295	0.46 ± 0.06	0.46 ± 0.06	0.50 ± 0.04	0.52 ± 0.03
RF	Yes	S	RN	297	0.42 ± 0.03	0.46 ± 0.04	0.62 ± 0.03	0.67 ± 0.03
DT	No	No	No	295	0.43 ± 0.03	0.45 ± 0.03	0.61 ± 0.03	0.65 ± 0.01
DT	Yes	R	FCN	295	0.43 ± 0.06	0.45 ± 0.06	0.51 ± 0.02	0.51 ± 0.03
DT	Yes	M	FCN	295	0.43 ± 0.06	0.45 ± 0.06	0.51 ± 0.02	0.51 ± 0.03
DT	No	R	FCN	295	0.43 ± 0.06	0.45 ± 0.06	0.51 ± 0.02	0.51 ± 0.03
DT	No	R	FCN	295	0.43 ± 0.06	0.45 ± 0.06	0.51 ± 0.02	0.51 ± 0.03

Table B.11: Summary of the results obtained on the YAAD dataset using the best models and parameters from the ICANS dataset with the ECG signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
DT	No	No	No	235	**0.29 ± 0.03**	**0.37 ± 0.13**	0.59 ± 0.02	0.64 ± 0.02
DT	Yes	M	FCN	235	**0.29 ± 0.06**	0.33 ± 0.05	0.48 ± 0.05	0.58 ± 0.03
DT	No	R	FCN	235	0.23 ± 0.02	0.33 ± 0.08	0.44 ± 0.03	0.65 ± 0.01
DT	Yes	S	FCN	235	0.25 ± 0.08	0.29 ± 0.07	0.47 ± 0.06	0.58 ± 0.04
DT	No	R	FCN	235	0.27 ± 0.03	0.26 ± 0.02	0.52 ± 0.04	0.59 ± 0.04
DT	Yes	M	FCN	235	0.27 ± 0.03	0.26 ± 0.02	0.52 ± 0.04	0.59 ± 0.04
DT	Yes	R	FCN	235	0.27 ± 0.03	0.26 ± 0.02	0.52 ± 0.04	0.59 ± 0.04

Table B.12: Summary of the results obtained on the AMIGOS dataset using the best models and parameters from the ICANS dataset with the ECG signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
DT	No	No	No	789	**0.26 ± 0.02**	0.27 ± 0.02	0.48 ± 0.02	0.52 ± 0.01
DT	Yes	S	FCN	789	0.23 ± 0.03	**0.28 ± 0.02**	0.33 ± 0.04	0.48 ± 0.01
DT	No	R	FCN	789	0.23 ± 0.02	**0.28 ± 0.01**	0.33 ± 0.02	0.51 ± 0.01
DT	No	R	FCN	789	0.24 ± 0.00	0.26 ± 0.02	0.33 ± 0.02	0.50 ± 0.02
DT	Yes	M	FCN	786	0.22 ± 0.04	0.27 ± 0.03	0.33 ± 0.04	0.49 ± 0.02
DT	Yes	M	FCN	786	0.22 ± 0.05	0.26 ± 0.03	0.30 ± 0.04	0.48 ± 0.01
DT	Yes	R	FCN	786	0.22 ± 0.05	0.26 ± 0.03	0.30 ± 0.04	0.48 ± 0.01

Table B.13: Summary of the results achieved by the best models and parameters on the YAAD dataset using the ECG signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
RF	No	S	SBN	235	**0.43 ± 0.12**	**0.49 ± 0.12**	0.77 ± 0.04	0.82 ± 0.01
DT	No	S	SBN	235	0.35 ± 0.03	0.48 ± 0.13	0.56 ± 0.04	0.65 ± 0.01
RF	No	No	No	235	0.36 ± 0.05	0.45 ± 0.02	0.55 ± 0.02	0.64 ± 0.02
DT	Yes	M	SBN	235	0.36 ± 0.07	0.45 ± 0.11	0.55 ± 0.03	0.65 ± 0.03
RF	Yes	M	SBN	235	0.36 ± 0.04	0.45 ± 0.09	0.73 ± 0.02	0.79 ± 0.02
RF	No	M	FCN	235	0.39 ± 0.11	0.41 ± 0.11	0.57 ± 0.01	0.63 ± 0.05
RF	No	M	SBN	235	0.39 ± 0.11	0.41 ± 0.11	0.57 ± 0.01	0.63 ± 0.05
RF	No	R	FCN	235	0.38 ± 0.17	0.42 ± 0.20	0.50 ± 0.03	0.57 ± 0.05
RF	No	S	FCN	235	0.35 ± 0.06	0.44 ± 0.12	0.55 ± 0.02	0.61 ± 0.03
RF	Yes	R	FCN	235	0.37 ± 0.13	0.42 ± 0.18	0.59 ± 0.04	0.66 ± 0.04

Table B.14: Summary of the results obtained on the ICANS dataset using the best models and parameters from the YAAD dataset with the ECG signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
RF	No	S	SBN	295	**0.36 ± 0.04**	**0.37 ± 0.03**	0.75 ± 0.02	0.79 ± 0.02
RF	Yes	R	SBN	295	0.34 ± 0.03	**0.37 ± 0.01**	0.61 ± 0.02	0.66 ± 0.02
RF	No	R	FCN	295	0.31 ± 0.05	**0.37 ± 0.06**	0.54 ± 0.02	0.57 ± 0.02
RF	No	M	SBN	295	0.33 ± 0.02	0.34 ± 0.02	0.61 ± 0.02	0.64 ± 0.02
RF	Yes	M	SBN	295	0.33 ± 0.03	0.34 ± 0.03	0.74 ± 0.03	0.77 ± 0.04
RF	No	S	FCN	295	0.33 ± 0.09	0.34 ± 0.10	0.53 ± 0.04	0.60 ± 0.01
RF	No	No	No	295	0.30 ± 0.02	0.33 ± 0.03	0.56 ± 0.01	0.61 ± 0.02
RF	No	M	FCN	295	0.30 ± 0.05	0.32 ± 0.03	0.61 ± 0.04	0.65 ± 0.04
DT	No	S	SBN	295	0.27 ± 0.07	0.31 ± 0.06	0.46 ± 0.03	0.56 ± 0.03
DT	Yes	M	SBN	295	0.27 ± 0.02	0.30 ± 0.02	0.50 ± 0.03	0.55 ± 0.02

Table B.15: Summary of the results obtained on the AMIGOS dataset using the best models and parameters from the YAAD dataset with the ECG signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
RF	No	M	FCN	789	**0.30 ± 0.03**	**0.34 ± 0.04**	0.46 ± 0.03	0.57 ± 0.01
RF	No	R	FCN	789	0.29 ± 0.02	0.34 ± 0.01	0.36 ± 0.04	0.50 ± 0.01
RF	No	S	SBN	789	0.29 ± 0.02	0.30 ± 0.02	0.63 ± 0.05	0.57 ± 0.02
RF	No	No	No	789	0.28 ± 0.01	0.30 ± 0.01	0.44 ± 0.03	0.48 ± 0.01
RF	Yes	M	SBN	786	0.28 ± 0.03	0.30 ± 0.05	0.54 ± 0.02	0.66 ± 0.00
DT	No	S	SBN	789	0.25 ± 0.02	0.32 ± 0.03	0.35 ± 0.04	0.54 ± 0.02
RF	No	M	SBN	789	0.27 ± 0.01	0.30 ± 0.02	0.46 ± 0.03	0.57 ± 0.01
RF	Yes	R	FCN	786	0.27 ± 0.05	0.30 ± 0.04	0.45 ± 0.00	0.59 ± 0.00
RF	No	S	FCN	789	0.26 ± 0.02	0.30 ± 0.07	0.38 ± 0.03	0.54 ± 0.02
DT	Yes	M	SBN	786	0.22 ± 0.04	0.26 ± 0.05	0.34 ± 0.51	0.55 ± 0.02

Table B.16: Summary of the results achieved by the best models and parameters on the AMIGOS dataset using the ECG signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
RF	No	No	No	789	**0.43 ± 0.02**	**0.47 ± 0.03**	0.52 ± 0.01	0.55 ± 0.01
DT	Yes	No	No	789	**0.43 ± 0.02**	0.46 ± 0.04	0.44 ± 0.02	0.49 ± 0.01
RF	Yes	No	No	789	0.42 ± 0.01	0.45 ± 0.02	0.50 ± 0.00	0.53 ± 0.01
LGBM	Yes	No	No	789	0.39 ± 0.02	0.42 ± 0.02	0.69 ± 0.01	0.73 ± 0.01
LGBM	No	No	No	789	0.39 ± 0.02	0.41 ± 0.05	0.58 ± 0.01	0.63 ± 0.01
DT	Yes	No	No	789	0.37 ± 0.06	0.39 ± 0.07	0.45 ± 0.01	0.48 ± 0.02
AB	No	No	No	789	0.35 ± 0.03	0.37 ± 0.04	0.75 ± 0.01	0.76 ± 0.00
RF	No	S	SBN	789	0.34 ± 0.01	0.37 ± 0.03	0.60 ± 0.04	0.69 ± 0.03
RF	No	S	FCN	789	0.33 ± 0.01	0.38 ± 0.01	0.46 ± 0.01	0.59 ± 0.01
RF	No	R	SBN	786	0.34 ± 0.02	0.37 ± 0.03	0.53 ± 0.01	0.63 ± 0.02

Table B.17: Summary of the results obtained on the ICANS dataset using the best models and parameters from the AMIGOS dataset with the ECG signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
DT	Yes	No	No	295	**0.38 ± 0.03**	**0.41 ± 0.04**	0.51 ± 0.02	0.52 ± 0.02
DT	Yes	No	No	295	0.34 ± 0.08	0.38 ± 0.07	0.52 ± 0.02	0.54 ± 0.02
RF	No	S	FCN	295	0.35 ± 0.04	0.37 ± 0.02	0.62 ± 0.01	0.67 ± 0.01
RF	No	No	No	295	0.33 ± 0.03	0.35 ± 0.01	0.66 ± 0.02	0.70 ± 0.01
RF	No	S	SBN	295	0.31 ± 0.02	0.32 ± 0.03	0.79 ± 0.02	0.82 ± 0.02
RF	No	RN	SBN	295	0.31 ± 0.02	0.32 ± 0.01	0.72 ± 0.02	0.75 ± 0.01
LGBM	No	No	No	295	0.29 ± 0.05	0.32 ± 0.05	0.65 ± 0.06	0.71 ± 0.04
RF	Yes	No	No	295	0.28 ± 0.04	0.30 ± 0.04	0.56 ± 0.04	0.60 ± 0.04
AB	No	No	No	295	0.27 ± 0.04	0.28 ± 0.05	0.84 ± 0.01	0.84 ± 0.01
LGBM	Yes	No	No	295	0.25 ± 0.06	0.28 ± 0.04	0.56 ± 0.01	0.61 ± 0.00

Table B.18: Summary of the results obtained on the YAAD dataset using the best models and parameters from the AMIGOS dataset with the ECG signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
LGBM	Yes	No	No	235	0.29 ± 0.02	**0.36 ± 0.03**	0.50 ± 0.04	0.60 ± 0.02
RF	No	RN	SBN	235	**0.30 ± 0.04**	0.31 ± 0.03	0.58 ± 0.02	0.65 ± 0.02
AB	No	No	No	235	0.27 ± 0.03	0.30 ± 0.01	0.68 ± 0.04	0.72 ± 0.01
RF	Yes	No	No	235	0.28 ± 0.01	0.28 ± 0.02	0.65 ± 0.02	0.70 ± 0.03
DT	Yes	No	No	235	0.27 ± 0.06	0.29 ± 0.04	0.54 ± 0.02	0.52 ± 0.04
LGBM	No	No	No	235	0.26 ± 0.03	0.29 ± 0.04	0.51 ± 0.02	0.61 ± 0.02
RF	No	S	SBN	235	0.26 ± 0.02	0.29 ± 0.02	0.78 ± 0.06	0.83 ± 0.03
RF	No	S	FCN	235	0.27 ± 0.03	0.28 ± 0.00	0.71 ± 0.06	0.76 ± 0.02
DT	Yes	No	No	235	0.24 ± 0.07	0.25 ± 0.04	0.53 ± 0.03	0.59 ± 0.04
RF	No	No	No	235	0.24 ± 0.03	0.24 ± 0.02	0.56 ± 0.02	0.60 ± 0.05

Table B.19: Summary of the results achieved by the best models and parameters on the ICANS dataset using the Voice signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
RF	No	S	SBN	251	**0.48 ± 0.05**	**0.51 ± 0.07**	0.69 ± 0.02	0.72 ± 0.00
RF	No	R	SBN	251	**0.48 ± 0.03**	0.49 ± 0.03	0.75 ± 0.02	0.76 ± 0.02
RF	No	S	SBN	251	0.47 ± 0.05	0.49 ± 0.05	0.85 ± 0.02	0.86 ± 0.02
DT	No	S	SBN	251	0.45 ± 0.03	0.49 ± 0.01	0.66 ± 0.01	0.69 ± 0.01
RF	No	R	SBN	251	0.46 ± 0.02	0.48 ± 0.04	0.80 ± 0.01	0.82 ± 0.01
RF	No	M	SBN	251	0.44 ± 0.03	0.47 ± 0.01	0.67 ± 0.02	0.70 ± 0.02
RF	No	M	SBN	251	0.45 ± 0.03	0.45 ± 0.05	0.81 ± 0.01	0.82 ± 0.02
LGBM	No	S	SBN	251	0.44 ± 0.01	0.46 ± 0.02	0.60 ± 0.03	0.62 ± 0.03
SVM	No	S	SBN	251	0.44 ± 0.03	0.45 ± 0.03	0.83 ± 0.03	0.85 ± 0.01
DT	Yes	M	SBN	250	0.43 ± 0.03	0.44 ± 0.02	0.57 ± 0.03	0.61 ± 0.01

Table B.20: Summary of the results obtained on the EPEDD dataset using the best models and parameters from the ICANS dataset with the Voice signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
SVM	No	S	SBN	826	**0.58 ± 0.04**	**0.60 ± 0.04**	0.91 ± 0.02	0.93 ± 0.01
RF	No	R	SBN	826	0.54 ± 0.03	0.56 ± 0.04	0.71 ± 0.02	0.75 ± 0.01
RF	No	R	SBN	826	0.53 ± 0.04	0.56 ± 0.05	0.70 ± 0.02	0.73 ± 0.02
LGBM	No	S	SBN	826	0.53 ± 0.04	0.56 ± 0.04	0.77 ± 0.04	0.81 ± 0.03
RF	No	S	SBN	826	0.47 ± 0.09	0.51 ± 0.06	0.74 ± 0.03	0.77 ± 0.03
DT	No	S	SBN	826	0.47 ± 0.07	0.50 ± 0.08	0.64 ± 0.02	0.69 ± 0.01
RF	No	M	SBN	826	0.45 ± 0.09	0.48 ± 0.08	0.71 ± 0.02	0.74 ± 0.02
RF	No	M	SBN	826	0.44 ± 0.10	0.48 ± 0.10	0.63 ± 0.04	0.67 ± 0.03
RF	No	S	SBN	826	0.43 ± 0.09	0.48 ± 0.07	0.65 ± 0.04	0.70 ± 0.04
DT	Yes	M	SBN	824	0.32 ± 0.02	0.38 ± 0.02	0.47 ± 0.02	0.55 ± 0.02

Table B.21: Summary of the results obtained on the EMOUERJ dataset using the best models and parameters from the ICANS dataset with the Voice signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
SVM	No	S	SBN	377	**0.76 ± 0.03**	**0.77 ± 0.03**	0.99 ± 0.00	0.99 ± 0.00
LGBM	No	S	SBN	377	0.71 ± 0.03	0.71 ± 0.03	0.92 ± 0.03	0.92 ± 0.03
RF	No	S	SBN	377	0.68 ± 0.10	0.69 ± 0.10	0.90 ± 0.01	0.91 ± 0.01
RF	No	R	SBN	377	0.68 ± 0.06	0.68 ± 0.06	0.89 ± 0.03	0.89 ± 0.03
RF	No	R	SBN	377	0.67 ± 0.04	0.67 ± 0.03	0.90 ± 0.02	0.90 ± 0.02
RF	No	M	SBN	377	0.67 ± 0.07	0.67 ± 0.07	0.91 ± 0.01	0.91 ± 0.01
RF	No	S	SBN	377	0.66 ± 0.08	0.67 ± 0.08	0.84 ± 0.03	0.84 ± 0.03
DT	No	S	SBN	377	0.62 ± 0.06	0.62 ± 0.06	0.84 ± 0.03	0.84 ± 0.03
RF	No	M	SBN	377	0.60 ± 0.06	0.61 ± 0.06	0.83 ± 0.02	0.83 ± 0.02
DT	Yes	M	SBN	376	0.58 ± 0.03	0.60 ± 0.04	0.79 ± 0.02	0.79 ± 0.02

Table B.22: Summary of the results achieved by the best models and parameters on the EPEDD dataset using the Voice signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
SVM	No	S	SBN	826	**0.60 ± 0.03**	**0.61 ± 0.04**	0.89 ± 0.02	0.91 ± 0.02
RF	No	R	SBN	826	0.58 ± 0.05	**0.61 ± 0.06**	0.73 ± 0.03	0.76 ± 0.03
SVM	No	R	SBN	826	0.56 ± 0.03	0.58 ± 0.04	0.78 ± 0.09	0.82 ± 0.08
RF	No	S	SBN	826	0.55 ± 0.05	0.58 ± 0.04	0.71 ± 0.03	0.75 ± 0.03
DT	No	S	SBN	826	0.54 ± 0.07	0.57 ± 0.06	0.64 ± 0.02	0.67 ± 0.03
DT	No	R	SBN	826	0.54 ± 0.06	0.56 ± 0.05	0.60 ± 0.02	0.64 ± 0.02
RF	Yes	S	SBN	824	0.53 ± 0.04	0.57 ± 0.04	0.64 ± 0.01	0.70 ± 0.01
KNN	No	S	SBN	826	0.53 ± 0.04	0.54 ± 0.05	0.65 ± 0.04	0.65 ± 0.03
SVM	No	M	SBN	826	0.51 ± 0.07	0.54 ± 0.07	0.73 ± 0.02	0.76 ± 0.01
SVM	Yes	S	SBN	824	0.52 ± 0.01	0.53 ± 0.00	0.86 ± 0.01	0.88 ± 0.01

Table B.23: Summary of the results obtained on the ICANS dataset using the best models and parameters from the EPEDD dataset with the Voice signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
SVM	No	S	SBN	251	**0.43 ± 0.02**	**0.45 ± 0.02**	0.78 ± 0.04	0.80 ± 0.02
SVM	Yes	S	SBN	250	**0.43 ± 0.04**	0.44 ± 0.05	0.74 ± 0.02	0.75 ± 0.02
RF	No	S	SBN	251	0.38 ± 0.06	0.40 ± 0.07	0.78 ± 0.02	0.79 ± 0.03
DT	No	R	SBN	251	0.38 ± 0.05	0.39 ± 0.04	0.52 ± 0.01	0.53 ± 0.03
RF	Yes	S	SBN	250	0.37 ± 0.06	0.38 ± 0.06	0.75 ± 0.03	0.77 ± 0.03
RF	No	R	SBN	251	0.35 ± 0.02	0.36 ± 0.01	0.81 ± 0.03	0.82 ± 0.02
SVM	No	M	SBN	251	0.30 ± 0.04	0.35 ± 0.02	0.47 ± 0.08	0.54 ± 0.03
KNN	No	S	SBN	251	0.31 ± 0.02	0.33 ± 0.02	0.42 ± 0.02	0.42 ± 0.01
DT	No	S	SBN	251	0.30 ± 0.04	0.33 ± 0.04	0.62 ± 0.02	0.65 ± 0.01
SVM	No	R	SBN	251	0.11 ± 0.02	0.24 ± 0.01	0.14 ± 0.03	0.27 ± 0.00

Table B.24: Summary of the results obtained on the EMOUERJ dataset using the best models and parameters from the EPEDD dataset with the Voice signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

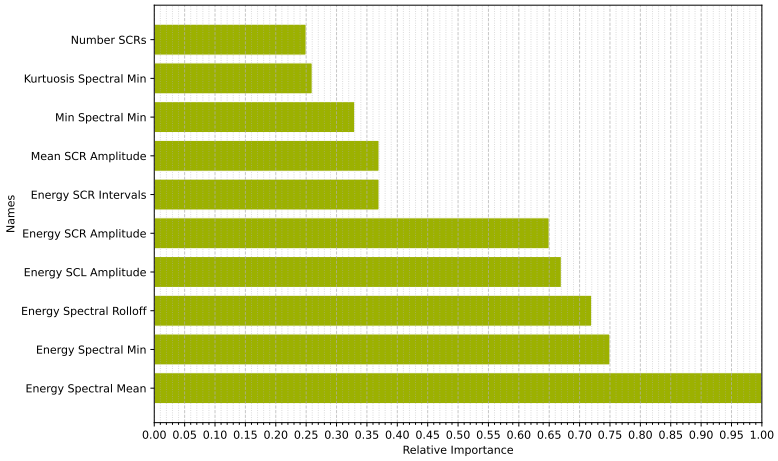
Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
SVM	No	M	SBN	377	**0.80 ± 0.04**	**0.80 ± 0.04**	0.98 ± 0.00	0.98 ± 0.00
SVM	No	S	SBN	377	0.76 ± 0.01	0.77 ± 0.01	0.91 ± 0.03	0.91 ± 0.03
KNN	No	S	SBN	377	0.75 ± 0.03	0.75 ± 0.03	1.00 ± 0.00	1.00 ± 0.00
SVM	No	RN	SBN	377	0.70 ± 0.03	0.70 ± 0.04	0.91 ± 0.02	0.91 ± 0.02
RF	No	S	SBN	377	0.70 ± 0.07	0.70 ± 0.07	0.82 ± 0.02	0.82 ± 0.02
RF	Yes	S	SBN	376	0.69 ± 0.06	0.69 ± 0.06	0.80 ± 0.03	0.80 ± 0.03
SVM	Yes	S	SBN	376	0.69 ± 0.03	0.69 ± 0.03	0.89 ± 0.04	0.89 ± 0.04
RF	No	R	SBN	377	0.68 ± 0.06	0.69 ± 0.06	0.82 ± 0.02	0.82 ± 0.02
DT	No	S	SBN	377	0.64 ± 0.09	0.64 ± 0.09	0.93 ± 0.01	0.93 ± 0.01
DT	No	R	SBN	377	0.54 ± 0.04	0.56 ± 0.02	0.96 ± 0.00	0.96 ± 0.00

Table B.25: Summary of the results achieved by the best models and parameters on the EMOUERJ dataset using the Voice signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

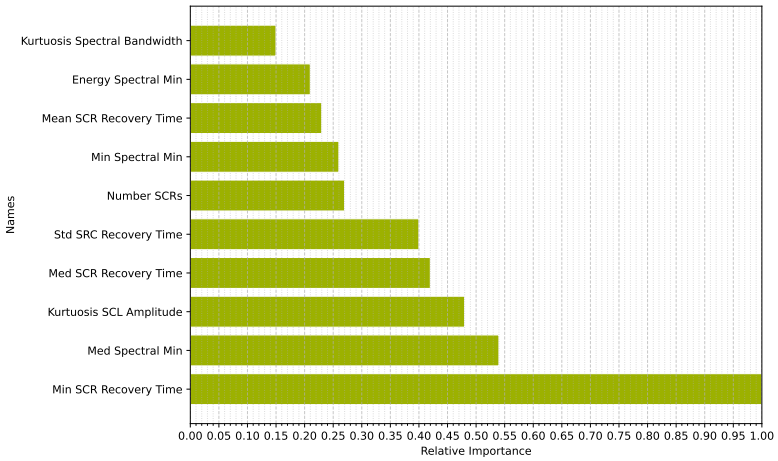
Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
SVM	No	M	SBN	377	**0.77 ± 0.09**	**0.77 ± 0.09**	1.00 ± 0.00	1.00 ± 0.00
SVM	No	M	FCN	377	0.75 ± 0.06	0.75 ± 0.06	0.94 ± 0.01	0.94 ± 0.01
RF	No	S	SBN	377	0.75 ± 0.03	0.75 ± 0.03	0.70 ± 0.04	0.71 ± 0.04
AB	No	M	SBN	377	0.75 ± 0.05	0.75 ± 0.05	1.00 ± 0.00	1.00 ± 0.00
SVM	No	S	FCN	377	0.74 ± 0.07	0.74 ± 0.07	0.97 ± 0.01	0.97 ± 0.01
KNN	No	M	SBN	377	0.73 ± 0.03	0.73 ± 0.03	0.85 ± 0.01	0.85 ± 0.01
KNN	No	S	SBN	377	0.73 ± 0.05	0.73 ± 0.05	0.85 ± 0.00	0.85 ± 0.01
RF	Yes	S	SBN	376	0.71 ± 0.06	0.71 ± 0.06	0.82 ± 0.02	0.82 ± 0.02
KNN	No	S	FCN	377	0.70 ± 0.04	0.71 ± 0.03	0.87 ± 0.03	0.87 ± 0.03
RF	No	No	No	377	0.70 ± 0.04	0.71 ± 0.04	0.89 ± 0.02	0.89 ± 0.02

Table B.26: Summary of the results obtained on the ICANS dataset using the best models and parameters from the EMOUERJ dataset with the Voice signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

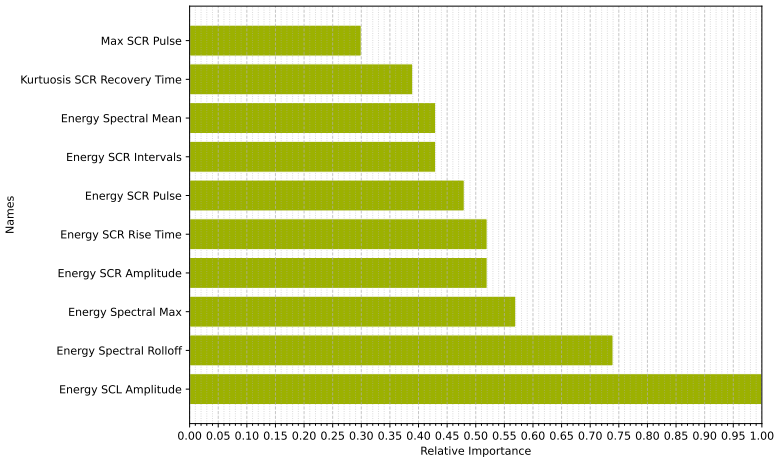
Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
SVM	No	M	FCN	251	**0.43 ± 0.02**	**0.45 ± 0.02**	0.78 ± 0.04	0.80 ± 0.03
SVM	No	S	FCN	251	**0.43 ± 0.02**	**0.45 ± 0.02**	0.78 ± 0.04	0.80 ± 0.03
SVM	No	M	SBN	251	0.40 ± 0.05	0.41 ± 0.06	1.00 ± 0.00	1.00 ± 0.00
RF	Yes	S	SBN	250	0.38 ± 0.04	0.41 ± 0.04	0.59 ± 0.01	0.62 ± 0.03
RF	No	S	SBN	251	0.36 ± 0.03	0.38 ± 0.03	0.68 ± 0.02	0.71 ± 0.03
RF	No	No	No	251	0.35 ± 0.04	0.38 ± 0.04	0.76 ± 0.03	0.77 ± 0.02
KNN	No	S	SBN	251	0.34 ± 0.02	0.36 ± 0.02	0.45 ± 0.06	0.45 ± 0.05
KNN	No	M	SBN	251	0.33 ± 0.01	0.35 ± 0.02	0.39 ± 0.02	0.40 ± 0.02
KNN	No	S	FCN	251	0.31 ± 0.05	0.34 ± 0.04	0.38 ± 0.02	0.40 ± 0.02
AB	No	M	SBN	251	0.30 ± 0.04	0.33 ± 0.03	0.99 ± 0.01	0.99 ± 0.01



(a) ICANS dataset.

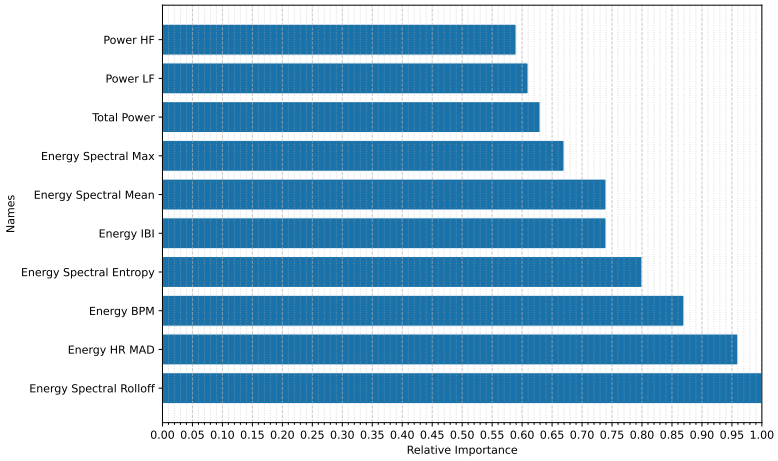


(b) YAAD dataset.

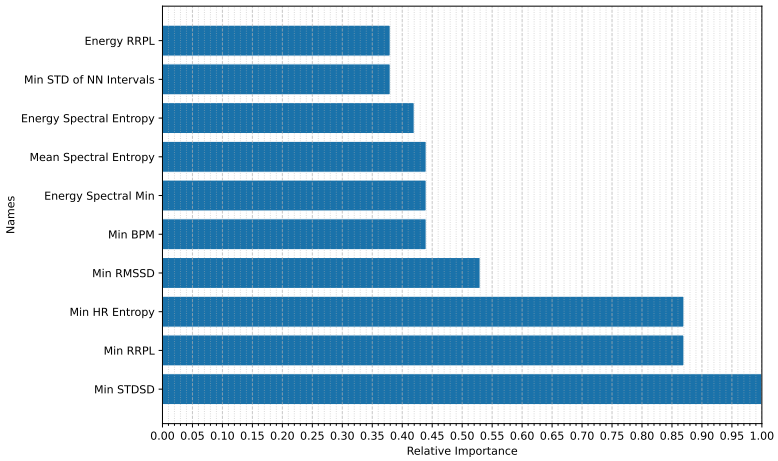


(c) AMIGOS dataset.

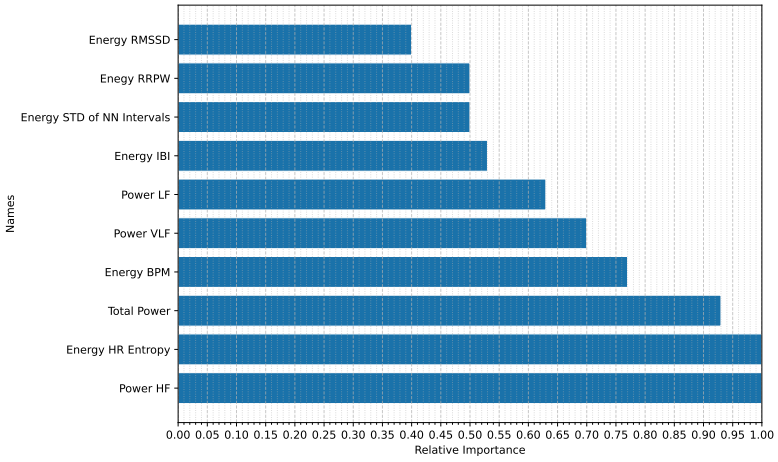
Figure B.10: Relative importance of the top ten features used in each dataset for the EDA signal, namely (a) ICANS dataset, (b) YAAD dataset, and (c) AMIGOS dataset. Max - Maximum; Med - Median; Min - Minimum; Std - Standard Deviation.



(a) ICANS dataset.

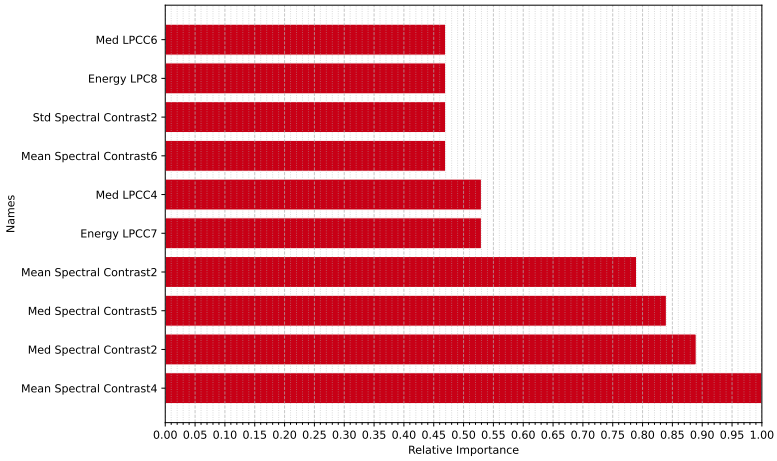


(b) YAAD dataset.

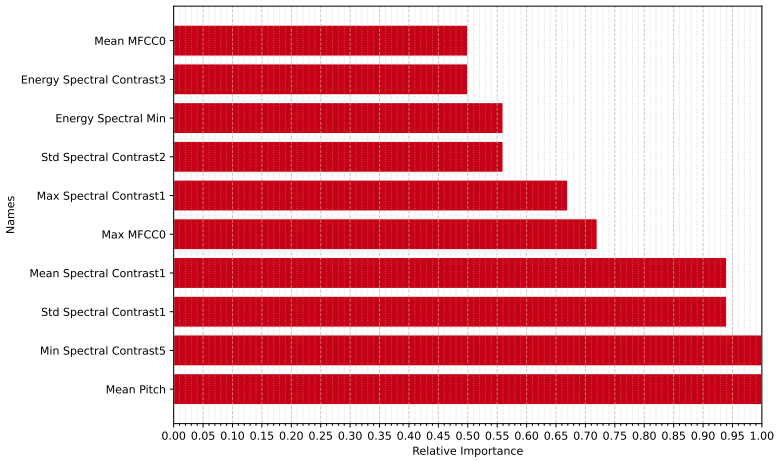


(c) AMIGOS dataset.

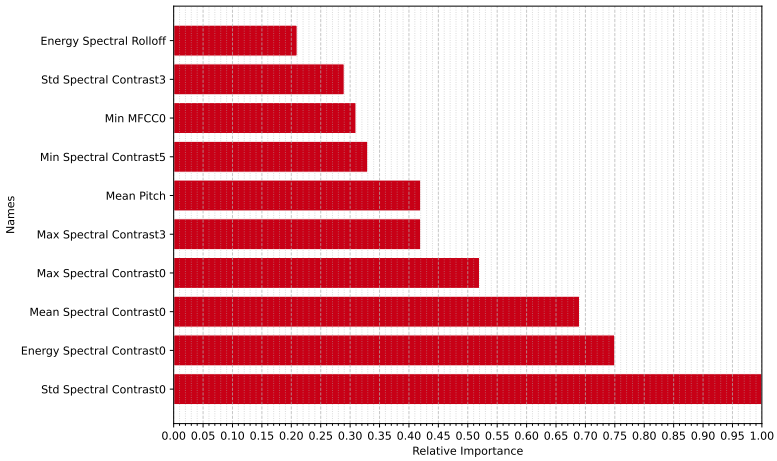
Figure B.11: Relative importance of the top ten features used in each dataset for the ECG signal, namely (a) ICANS dataset, (b) YAAD dataset, and (c) AMIGOS dataset. Min - Minimum.



(a) ICANS dataset.



(b) EPEDD dataset.



(c) EMOUERJ dataset.

Figure B.12: Relative importance of the top ten features used in each dataset for the Voice signal, namely (a) ICANS dataset, (b) EPEDD dataset, and (c) EMOUERJ dataset. Max - Maximum; Med - Median; Min - Minimum; Std - Standard Deviation.

Table B.27: Summary of the results obtained on the EPEDD dataset using the best models and parameters from the EMOUERJ dataset with the Voice signal. The best F1-score and BA are highlighted and the results are sorted in descending order.

Model	Normalization			Nº Samples	Test		Train	
	Signal	Feature						
	Amplitude	Type	Domain		F1	BA	F1	BA
SVM	No	M	SBN	826	**0.67 ± 0.01**	**0.66 ± 0.01**	1.00 ± 0.00	1.00 ± 0.00
SVM	No	S	FCN	826	0.59 ± 0.02	0.61 ± 0.03	0.85 ± 0.01	0.88 ± 0.01
RF	No	S	SBN	826	0.57 ± 0.01	0.61 ± 0.01	0.66 ± 0.04	0.70 ± 0.04
SVM	No	M	FCN	826	0.56 ± 0.02	0.59 ± 0.03	0.71 ± 0.01	0.75 ± 0.01
KNN	No	M	SBN	826	0.55 ± 0.03	0.56 ± 0.02	0.61 ± 0.02	0.61 ± 0.02
KNN	No	S	SBN	826	0.53 ± 0.03	0.54 ± 0.02	0.63 ± 0.02	0.62 ± 0.01
RF	No	No	No	826	0.51 ± 0.04	0.54 ± 0.04	0.61 ± 0.00	0.65 ± 0.00
RF	Yes	S	SBN	824	0.49 ± 0.01	0.54 ± 0.02	0.56 ± 0.01	0.63 ± 0.01
AB	No	M	SBN	826	0.48 ± 0.03	0.47 ± 0.03	0.82 ± 0.02	0.82 ± 0.02
KNN	No	S	FCN	826	0.47 ± 0.04	0.47 ± 0.03	0.58 ± 0.02	0.57 ± 0.01

C

Multimodal Emotion Recognition

Table C.1: Optimal outcomes for each late fusion technique across different fusion types in the ICANS dataset.

Fusion Type	Metrics	Fusion Experiments			
		1	2	3	4
Voting	F1	**0.51 ± 0.03**	0.43 ± 0.03	0.45 ± 0.07	0.47 ± 0.04
	BA	0.52 ± 0.04	0.49 ± 0.04	0.48 ± 0.08	0.50 ± 0.06
Averaging	F1	0.50 ± 0.03	0.49 ± 0.03	0.43 ± 0.04	0.49 ± 0.06
	BA	0.54 ± 0.02	0.53 ± 0.04	0.44 ± 0.05	0.52 ± 0.09
Max Voting	F1	0.47 ± 0.00	**0.50 ± 0.00**	**0.48 ± 0.00**	**0.55 ± 0.00**
	BA	0.51 ± 0.00	**0.55 ± 0.00**	0.45 ± 0.00	**0.61 ± 0.00**
Weighted Averaging	F1	0.50 ± 0.03	0.50 ± 0.01	**0.48 ± 0.05**	0.51 ± 0.02
	BA	**0.55 ± 0.02**	0.52 ± 0.02	**0.49 ± 0.07**	0.53 ± 0.05

All metrics are macro averaged; (1) - EDA + ECG + VOICE | (2) - EDA + ECG | (3) - EDA + VOICE | (4) - ECG + VOICE

Table C.2: Contigency matrix of the best late fusion results obtained in each group of modalities used. This representation is meant to keep track of which modalities might be hindering or benefiting from the fusion strategy. The best results are highlighted.

		Modalities			Accuracy
		EDA	ECG	Voice	
Modalities	EDA	1.00	0.55	0.61	0.41
		1.00	0.55	-	0.44
		1.00	-	0.45	0.39
	ECG	0.58	1.00	0.55	0.39
		0.55	1.00	-	0.44
		-	1.00	**0.63**	0.47
	Voice	0.51	0.43	1.00	0.49
		0.45	-	1.00	0.49
		-	**0.58**	1.00	0.50

Accuracy score is reported; (1) - EDA + ECG + VOICE; (2) - EDA + ECG; (3) - EDA + VOICE; (4) - ECG + VOICE.

Table C.3: Optimal outcomes for each late fusion technique across different fusion types in the YAAD and AMIGOS dataset.

Fusion Type	**Metrics**	**Datasets**	
		YAAD	AMIGOS
Voting	F1	0.38 ± 0.05	0.47 ± 0.03
	BA	0.47 ± 0.09	0.48 ± 0.04
Averaging	F1	0.38 ± 0.06	0.45 ± 0.04
	BA	0.38 ± 0.08	0.48 ± 0.03
Max Voting	F1	**0.39 ± 0.00**	**0.48 ± 0.00**
	BA	**0.59 ± 0.00**	**0.54 ± 0.00**
Weighted Averaging	F1	0.38 ± 0.00	0.46 ± 0.03
	BA	0.48 ± 0.08	0.50 ± 0.02

D

Additional Experiments

Table D.1: Number of emotions per type of task.

Task	HVHA	LVHA	LVLA	HVLA
Memory Recall	34	11	1	15
Visual	33	18	24	46
Auditory	15	14	5	26
Audiovisual	12	0	0	0
Acting	26	18	7	11

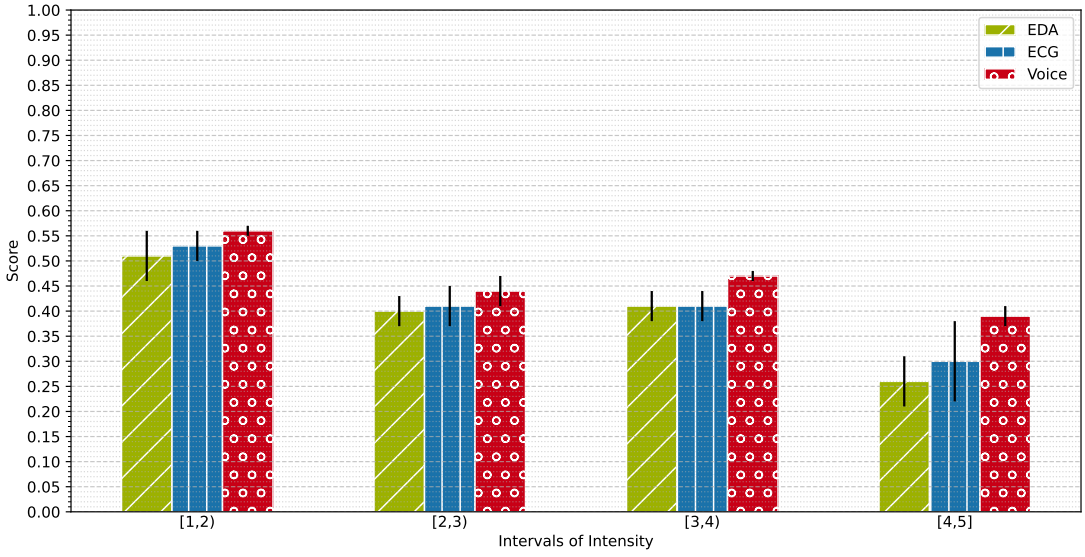


Figure D.1: Graphical representation of the performance obtained for each interval of intensity of an emotion, in the ICANS dataset. Accuracy score is reported taking into account the best ten models of each modality.

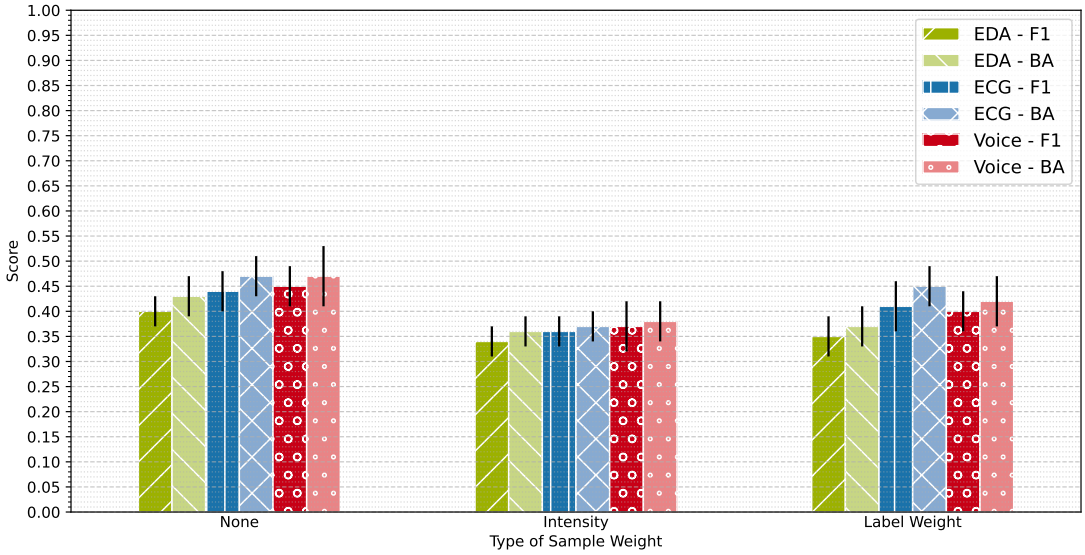


Figure D.2: Graphical representation of the performance obtained for each type of sample weight, in the ICANS dataset. Accuracy score is reported taking into account the best ten models of each modality.

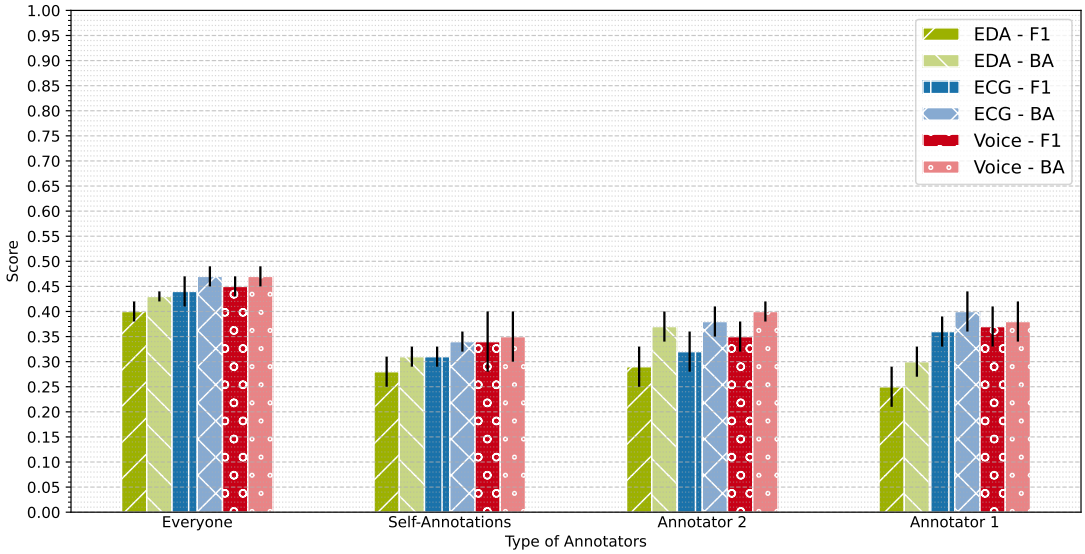


Figure D.3: Graphical representation of the performance obtained for each label from different annotator (or subject itself), in ICANS dataset. Accuracy score is reported taking into account the best ten models of each modality.

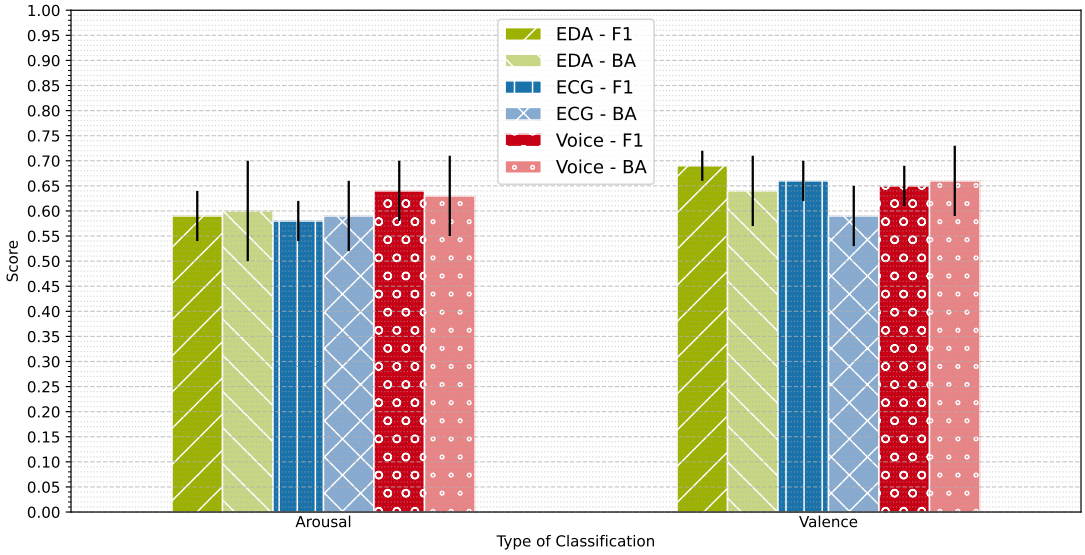


Figure D.4: Graphical representation of the performance obtained for arousal and valence (2-class problem in each case), in ICANS dataset. Accuracy score is reported taking into account the best ten models of each modality.

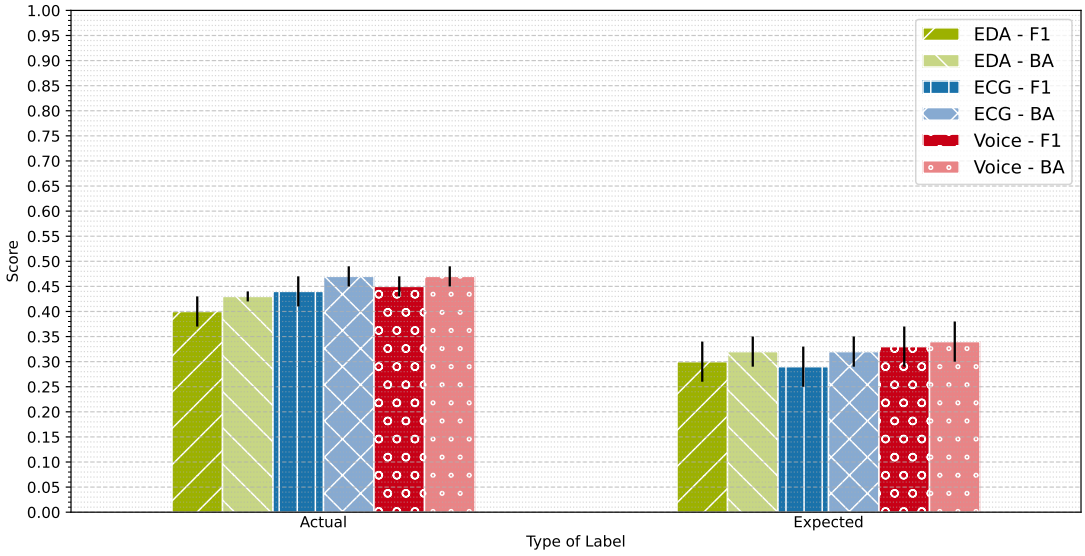


Figure D.5: Graphical representation of the performance obtained for the actual versus expected emotion, in ICANS dataset. Accuracy score is reported taking into account the best ten models of each modality.

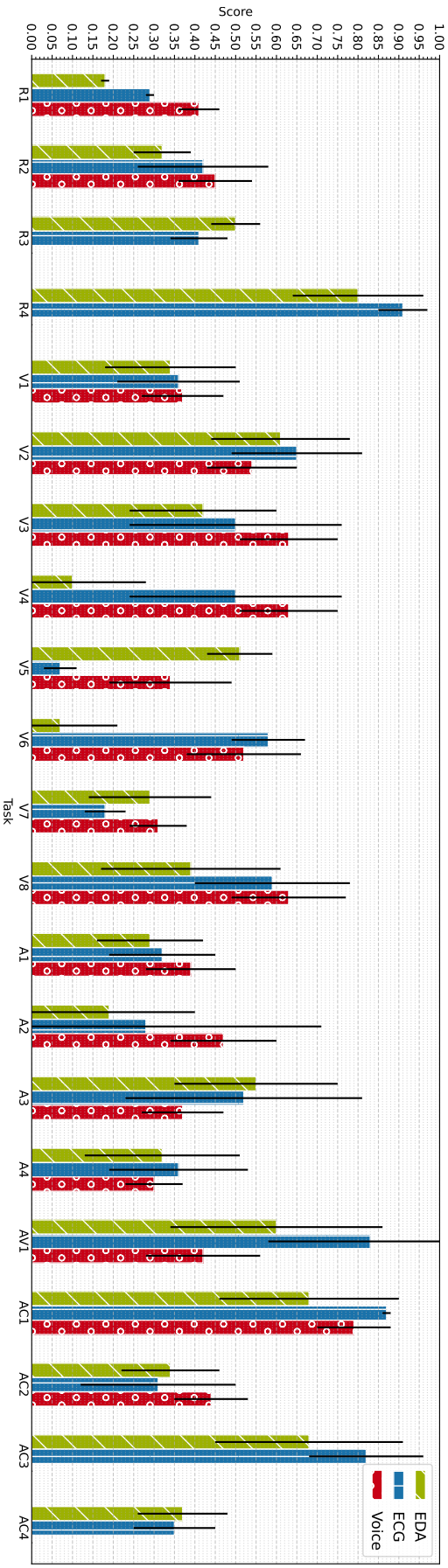


Figure D.6: Graphical representation of the performance obtained trought the acquisition protocol, in the ICANS dataset. Each bin of the X-axis represents a different stimuli. Accuracy score is reported taking into account the best ten models of each modality.

I

Features Extracted

Detailed descriptions of some features extracted on each data modality are showed in this section. Features extracted from the spectral domain of the three modalities are described:

- **Spectral Rolloff:** frequency below which a certain percentage of the total spectral energy of the signal is contained. For example, the spectral rolloff frequency for 90% of the energy represents the frequency below which 90% of the total energy is contained;
- **Spectral Bandwidth:** a measure of the width or spread of frequencies in a signal. It quantifies how wide or narrow the frequency content is in a given segment of the signal. A higher spectral bandwidth indicates a broader range of frequencies, while a lower spectral bandwidth suggests a narrower concentration of frequencies;
- **Spectral Centroid:** weighted average of the frequencies present in the signal, with each frequency weighted by its amplitude value;
- **Spectral Contrast:** the STD of the power spectrum around the spectral centroid;
- **Spectral Entropy:** measure of the disorder or randomness of the frequency distribution of the signal;
- **Spectral Maximum**: maximum amplitude value in the frequency spectrum of the signal;
- **Spectral Minimum:** minimum amplitude value in the frequency spectrum of the signal;
- **Spectral Mean:** arithmetic mean of the amplitude values in the frequency spectrum of the signal.

Features extracted from the whole EDA signal:

- **SCR Pulse:** the amplitudes values from a SCR pulse;

- **SCR Risetime:** the time between the stimuli and the instants where the SCR pulse reaches the peak;
- **SCR Recoverytime:** when the SCR pulse peak recovers (decline) to half of it's amplitude;
- **SCR Amplitude:** the phasic component of the EDA signal;
- **SCL Amplitude:** the tonic component of the EDA signal;
- **SCR Intervals:** the consecutive intervals between peaks of SCRs pulses;
- **Number of SCRs Pulses:** it quantifies the count of pulses corresponding to SCRs pulses within a signal.

Features extracted from each frame of an ECG signal:

- **BPM:** measures the HR, indicating the number of heartbeats occurring in one minute;
- **IBI:** represents the time interval between successive heartbeats, essentially measuring the duration between two heartbeats;
- **STD of NN INtervals:** measures the variability of HR, specifically the STD of the intervals between consecutive normal heartbeats (NN intervals);
- **STDSD:** quantifies the variability between consecutive HR differences, providing insight into short-term HR fluctuations;
- **RMSSD:** calculates the square root of the mean of the squared differences between successive HR intervals, primarily reflecting short-term HRV;
- **pnn20:** it's the percentage of adjacent RR intervals differing by more than 20 milliseconds, an index of HRV and cardiac autonomic function;
- **pnn50:** it's the percentage of adjacent RR intervals differing by more than 50 milliseconds an index of HRV and cardiac autonomic function;
- **HRMAD:** a measure of the dispersion of HR values around the mean;
- **RRPL:** STD of points perpendicular to the line of identity in a Poincaré plot, a graph that shows the relationship between successive NN intervals;
- **RRPW:** STD of points along the line of identity in a Poincaré plot;
- **HR Entropy:** short-term variability, a measure of the variability of HR over short time intervals;
- **RRPL/RRPW:** the ratio of the STD of points perpendicular to the line of identity to the STD of points along the line of identity in a Poincaré plot;

- **Breathing Rate:** the number of breaths per minute, estimated from the Respiratory Sinus Arrhythmia (RSA), the variation in heart rate that occurs with breathing.

Features extracted from the the full ECG signal:

- **VLF:** the power in the signal that is located in the frequency ranging within [0.0033, 0.04] Hz;
- **LF:** the power in the signal that is located in the frequency ranging within [0.04, 0.15] Hz;
- **HF:** the power in the signal that is located in the frequency ranging within [0.15, 0.4] Hz;
- **LF/HF:** the ratio of the power in the LF band to the power in the HF band;
- **Total Power:** the total power in the HR signal, which is the sum of the power across all frequency bands;
- **VLF Percentage:** the percentage of power in the VLF band relative to the total power in the HR signal;
- **LF Percentage:** the percentage of power in the LF band relative to the total power in the HR signal;
- **HF Percentage:** the percentage of power in the HF band relative to the total power in the HR signal;
- **LF Norm:** the normalized power in the LF band, which is the LF power divided by the total power minus the VLF power;
- **HF Norm:** the normalized power in the HF band, which is the HF power divided by the total power minus the VLF power.

Features extracted from the frames of a Voice signal:

- **MFCC:** a representation of the spectral envelope of a sound signal, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency;
- **CPP:** a measure of the prominence of the highest peak in the cepstral spectrum, which can indicate the clarity of the most prominent formant in a speech signal;
- **LPCC:** a set of coefficients obtained from the linear prediction analysis of a speech signal, used to model the spectral envelope of the signal;
- **ZCR:** it's the rate at which the signal changes sign, which can be used to estimate the frequency content and periodicity of the signal.

Features extracted from the full Voice signal:

- **Mean Pitch:** represents the average F0 of a voice signal and indicates the typical pitch level of a speaker's voice during a segment of speech;
- **STD of Pitch:** measures the variation in F0 values within a voice signal, reflecting the degree of pitch instability or variability during speech;
- **Harmonic to Noise Ratio (HNR):** a measure that quantifies the amount of additive noise in the voice signal;
- **Local Jitter:** this is the average absolute difference between consecutive periods, divided by the average period;
- **Local Absolute Jitter:** this is the average absolute difference between consecutive periods, in seconds;
- **Rap Jitter:** this is the Relative Average Perturbation, the average absolute difference between a period and the average of it and its two neighbours, divided by the average period;
- **PPQ5 Jitter:** this is the five-point Period Perturbation Quotient, the average absolute difference between a period and the average of it and its four closest neighbours, divided by the average period;
- **DDP Jitter:** this is the average absolute difference between consecutive differences between consecutive periods, divided by the average period;
- **Local Shimmer:** this is the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude;
- **Local dB Shimmer:** this is the average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20;
- **APQ3 Shimmer:** this is the three-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of its neighbours, divided by the average amplitude;
- **AQPQ5 Shimmer:** this is the five-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its four closest neighbours, divided by the average amplitude;
- **APQ11 Shimmer:** this is the 11-point Amplitude Perturbation Quotient, the average absolute difference between the amplitude of a period and the average of the amplitudes of it and its ten closest neighbours, divided by the average amplitude;
- **DDA Shimmer:** this is the average absolute difference between consecutive differences between the amplitudes of consecutive periods.