Feel My Heart: Emotion Recognition Using the Electrocardiogram

Inês Antunes de Magalhães

Mestrado Integrado em Bioengenharia

Supervisor: Jaime dos Santos Cardoso, Ph.D.

Co-Supervisor: João Ribeiro Pinto, M.Sc.

July 16, 2021

Feel My Heart: Emotion Recognition Using the Electrocardiogram

Inês Antunes de Magalhães

Mestrado Integrado em Bioengenharia

Approved in public examination by the Jury:

President: Ana Maria Rodrigues de Sousa Faria de Mendonça

Opponent: Pedro Miguel Martins Ferreira

Referee: Jaime dos Santos Cardoso

Referee: João Tiago Ribeiro Pinto

July 16, 2021

Resumo

Emoções são processos complexos, que envolvem sentimentos, movimento corporal e até mesmo reações cognitivas ou pensamentos. Através das emoções, o ser humano comunica para além das palavras, avaliando pontos-chave como as expressões faciais, a entoação do discurso e a linguagem corporal. Embora esta avaliação seja feita de uma forma quase inconsciente, torna-se interessante pensar na possibilidade de sistemas tecnológicos também serem capazes de distinguir estados afetivos e reconhecer as emoções do seu utilizador.

A Computação Afetiva estuda e desenvolve algoritmos que têm a capacidade de reconhecer e processar emoções. Tais sistemas podem revolucionar as interações humano-computador, desenvolvendo tecnologia mais inteligente e humanizada, e conseguindo assim uma intercomunicação eficiente e avançada.

Para obter este tipo de dispositivos, é necessário utilizar indicadores ou modalidades que sejam sensíveis às variações de emoção, tais como expressões faciais, fala ou mesmo sinais fisiológicos. Desta forma, esta dissertação visa realizar uma revisão completa dos conceitos fundamentais e dos métodos de estado-da-arte sobre o reconhecimento das emoções utilizando o electrocardiograma, com o objectivo final de desenvolver uma arquitectura fiável de deep learning, para além de avaliar algumas limitações relacionadas com o uso de dados pessoais, tais como a subject dependency. Posto isto, esta dissertação aborda diferentes métodos do estado da arte e explora a sua utilização de modo a desenvolver a arquitectura final a ser apresentada. Além disso, são consideradas diferentes divisões comparando o desempenho obtido quando utilizadas divisões mais realistas em vez de aleatórias entre os sets de treino e teste.

Embora tenham havido evidentes desafios durante as replicações dos métodos presentes na literatura, uma abordagem de self-supervised learning foi replicada com sucesso e utilizada como base para a arquitectura de deep learning desenvolvida. Foram ainda aplicadas melhorias, tais como a implementação de um pré-treino mais robusto e a adição de algumas arquitecturas de aprendizagem, tais como MLPs e LSTMs, de modo a explorar a informação temporal do sinal de ECG. Assim, esta abordagem baseada no método de self-supervised learning foi desenvolvida, e melhorada através da aplicação de técnicas MIL que levaram a precisões de reconhecimento de emoções de cerca de 78%, 82% e 86%, dependendo do método de agregação aplicado. No entanto, no que diz respeito à utilização de signal e subject-independent settings, bem como a experiências de cross-database, os resultados diminuíram acentuadamente, obtendo precisões em torno dos 50% e 60%. Estes desempenhos revelam uma elevada dependência da arquitetura em relação aos dados utilizados, que resulta da utilização bases de dados reduzidas e limitadas.

Em conclusão, através da utilização de divisões aleatórias dos dados, foi possível obter resultados de elevada precisão, tais como os reportados nos métodos da literatura. No entanto, considerando divisões e condições mais realistas, tornou-se claro que a área do reconhecimento das emoções ainda não se encontra tão desenvolvida quanto se poderia pensar, havendo bastante espaço para melhoria, assim que se verifique o aumento da quantidade de dados adquiridos e utilizados para treinar arquitecturas de deep learning.

Abstract

Emotions are complex processes, involving feelings, body movement and even cognitive reactions or thoughts. Through emotions, humans communicate beyond words, by assessing key points such as facial expressions, speech intonation, and body language. Although this deep assessment is done in an almost unconscious way, it becomes interesting to think about the possibility of machines also being able to distinguish affective states and recognize their user's emotions.

Affective Computing studies and develops algorithms that can recognize and process human affects. Such systems may revolutionize human-computer interactions, by developing more intelligent and human-like technology and thus, achieving efficient and advanced intercommunication.

In order to obtain this type of devices, there's the need to use emotion-sensitive indicators such as facial expressions, speech or even physiological signals. In this way, this dissertation aimed to perform a complete overview of the fundamentals and prior art methods on emotion recognition using the electrocardiogram, with the final goal of developing a reliable deep learning architecture able to recognize emotions and assessing some limitations related to personal data such as subject dependency. Having this said, this dissertation addresses different prior art approaches and explores their use and application so as to build the final architecture. Furthermore, different data divisions were considered so as to compare the performance obtained when more realistic settings were used instead of random data divisions between train and test sets.

Although there were some evident challenges during the literature approaches' replications, a self-supervised learning approach was successfully followed and used as the basis for the deep learning architecture developed during this dissertation. Improvements were further developed, such as the implementation of a stronger and more robust pre-training and the addition of some deep learning architectures such as MLPs and LSTMs, so as to explore the temporal information of the ECG signal. Thus, this self-supervised learning-based approach was developed and improved through the application of MIL techniques that led to emotion recognition accuracies of around 78%, 82% and 86%, depending on the aggregation method applied. However, concerning signal and subject-independent settings, as well as cross-database experiments, results decreased considerably, obtaining accuracies of about 50% and 60%. These performances reveal a high data dependency that results from the use of small and limited datasets.

In conclusion, through the use of random splits, it was possible to obtain high accuracy results, such as the ones present in the literature. However, through the assessment of more realistic conditions, concerning data divisions, it became clear that the emotion recognition field is still less developed than one could firstly notice, and there is still a lot of room for improvement once larger amount of data is acquired and used to train deep learning architectures.

Acknowledgements

Dizem que estes são os melhores anos da nossa vida, que nada supera os tempos de faculdade em que a responsabilidade é pouca e a liberdade muita. Talvez tenham razão porque, agora, no topo destes 5 anos, eu só queria poder repetir tudo de novo. É uma sensação muito estranha, esta de terminar um ciclo. Aquele misto de sensações entre o entusiasmo e o medo de dar um novo passo e sair da bolha. Mas, e porque começamos entre ditados e dizeres, há quem diga que nada dura para sempre, e que tudo o que é bom, acaba depressa.

Olho para trás e sinto que tudo passou demasiado rápido: a euforia de caloira, as amizades criadas em segundos, o desespero partilhado nas épocas de exame, a família que se criou naquela cidade chamada de Porto, num apartamento pequenino de todos nós, naquele 2ºN pronto a receber todo e qualquer um que por lá quisesse passar (mesmo um G chateado e resingão).

Por tudo isto e por muito mais, só tenho que agradecer. Aos amigos que se tornaram família e que tão bem me conhecem: À Catarina, que, mais que amiga, é companheira de casa, irmã de coração e a pessoa que me percebe mesmo sem eu dizer uma palavra. À Xana que toma conta de todos nós e do futuro de cada vez que nega uma palhinha de plástico, e que será sempre das pessoas mais especiais que conheci. À Mariana, o meu projeto vida, o meu porto de abrigo, e a pessoa que me enche o coração todos os dias, de amizade, de apoio e de amor. Aos meninos do OG, ao Henrique, que só queremos muitas vezes enfiar dentro de um jarro de água quente com vinho e enviá-lo para Itália, e ao Venâncio, FCPorto da cabeça aos pés e 100% bondade. Ao meu Miguel, namorado e melhor amigo, pessoa que me vê como sou e que, todos os dias, me ama por e para além disso. Só quero o Mundo para ti e, de preferência, comigo a teu lado.

A ERASMUS e ao nosso Zweringweg 11. Há quem diga que não fazemos amigos, só os reconhecemos, por isso, ao Pedro, que imediatamente reconheci quando naquela noite de Março chegou a Twente. À Cat Morgado, a pessoa do "Sul" que eu mais adoro e manager de redes, e a todos os momentos, do gym ao sofá, do trabalho às festinhas. Aos amigos que por lá fizemos e às histórias que nunca serão esquecidas. À Ritinha que também sabe ser Ritona, aos bolinhos e às trancinhas, aos mapas astrais e linguagem corporal.

À Peanur, que foi madrinha de 9 pequenos rebentos! A cada uma dessas pessoas, às noites que passamos juntos, de preto ou a cores, em modo festa ou solene. Em especial, à Rita Barros, porque sei que há uma ligação entre nós que o tempo não quebrará.

À ANEEB, essa associação que é tão mais que trabalho. A vocês que não são companheiros de direção, são amigos: aos momentos cringe do António, ao robe laranja da Tatiana, às cabrinhas da Lua, à Maria da Ana, à velocidade 2x do Pedro quando fala, à incríiiibel Glória. Às nossas chamadas longas, fotos de gatos, histórias partilhadas. Trabalhar com vocês não é trabalho. A EAS, esse departamento tão utópico, tão diferente, tão carinhosamente meu e de todos os que ainda lá ficarão. Obrigada, equipa, vou ter tantas saudades.

Ao João Pinto, que efetivamente merece um parágrafo exclusivo. Por ser mais que um orientador durante todo este percurso. Esta tese é tão minha quanto tua, e só desejo que todos os dias tenhas o reconhecimento que mereces pelo profissional e pessoa que és. Uma vez disseste-me que

o dia de um aluno de doutoramento tem o dobro das horas, mas algo me diz que isso não será verdade para todos os alunos. Por isso, o meu enorme e mais sincero obrigada! Ao Professor Jaime, por todos os conselhos e orientação durante esta dissertação, e ao grupo de Biometria que tão bem acolhe os seus alunos. À Telma, parceira da ANEEB e da tese. Obrigada pelo apoio que foste e que és. Só espero que sejas sempre muito feliz e que o Mundo nunca te roube essa bondade tão tua, tão genuína e tão única.

A Amarante, e tudo o que de bom me continua a dar. À Rita, artista apaixonada e cheia de cor, à Sara Alexandra, pessoa tão doce e tão amiga, à Sara Nunes, violinista de profissão e com uma alma cheia de música e à Natália, parceira no voleibol e em tão mais que isso. Ao Parsec e as suas maluqueiras, à Bárbara, enfermeira de todos os dias, ao grupo das cafezadas nas noites de Verão e naquelas festas de Junho que ninguém esquece. Por tudo o que foram e o que continuam a ser para mim, o meu enorme obrigada.

À família. A todos eles, que me permitiram ser quem sou. À mãe e ao pai, essas forças da natureza que me deram a vida, me ofereceram um lar e me ensinaram a ser mais e melhor. A essas duas pessoas que me ofereceram mais duas, dois irmãos para chatear e amar incondicionalmente. À madrinha e à tia Fatinha, por me inspirarem todos os dias, por serem casa e conforto, por estarem sempre presentes. À avó Mila, matriarca da família, mulher forte, rija e eterna. À avó Cisa, e à sua forma tão peculiar de cuidar, com a sopa da semana, as compotas e a marmelada, o enxoval que desde cedo me prepara. Ao meu gatinho Akira, bola de pelo laranja com tanto amor para dar como só um animal consegue, e a todas as suas sestas durante a produção desta tese.

Ao Porto, à FEUP, a todas as pessoas que cruzaram o meu caminho. O tempo para nós nunca irá acabar, teremos sempre os Clérigos para descer a correr, os Aliados e as suas serenatas, os trajes pretos e as jantaradas até de madrugada. E se agora é o momento da partida, que breve venham os de reencontro.

"Capa negra de saudade, no momento da partida. Segredos desta cidade, levo comigo para a vida."

Não podia ter sido melhor.

Inês Antunes

"You learn that it doesn't matter where you have reached,
But where you are going to.
But if you don't know where you are going to, anywhere will do"

William Shakespeare

Contents

List of Figures

List of Tables

Abbreviations

1D	Unidimensional
2D	Bidimensional
3D	Tridimensional
AC	Affective Computing
ACF_{coef}	Maximum Autocorrelation Coefficient
ACF_{freq}	Maximum Autocorrelation Lag Time
ADM	Affective Dimensional Models
AMIGOS	A Dataset for Multimodal Research of Affect, Personality Traits and Mood on Individuals and GrOupS
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
ANS	Autonomic Nervous System
ApEn	Approximate Entropy
ASCERTAIN	A Multimodal Database for Implicit Personality and Affect Recognition
AV	Atrioventricular
AVNN	Average of NN Intervals
AU	Action Unit
BP	Back Propagation
BDT	Binary Decision Tree
BEMD	Bidimensional Empirical Mode Decomposition
BiLSTM	Bidirectional Long Short-Term Memory Networks
BVP	Blood Volume Pulse
CCA	Canonical Correlation Analysis
CCC	Concordance Correlation Coefficient
CNN	Convolutional Neural Network
CNS	Central Nervous System
CVRR	Correlation of Variation of RR Intervals
CWT	Continuous Wavelet Transform
DEAP	A Database for Emotion Analysis using Physiological Signals
DBN	Deep Belief Network
DCNN	Deep Convolutional Neural Network
DECAF	MEG-based Multimodal Database for Decoding Affective Physiological Responses
DEM	Discrete Emotional Models
DREAMER	A Database for Emotion Recognition through EEG and ECG Signals from Wireless Low-cost Off-the-Shelf Devices
DFA	Detrended Fluctuation Analysis
DFT	Discrete Fourier Transform

DNN	Deep Neural Network
DPM	Deformable Part Models
DWT	Discrete Wavelet Transform
EDA	Electrodermal Activity
EDR	ECG-derived Respiration
ECG	Eletrocardiogram
EEG	Electroencephalogram
ELM	Extreme Learning Machine
EMD	Empirical Mode Decomposition
EMG	Electromyogram
EOG	Electrooculography
ET	Ensemble Tree
FACS	Facial Action Coding System
FCN	Fully Convolutional Layer
FER	Facial Emotion Recognition
FFT	Fast Fourier Transform
FkNN	Fuzzy k-Nearest Neighbour
FP	Fisher Projection
FVS	Finite Variance Scaling
GA	Genetic Algoirthm
GAN	Generative Adversarial Networks
GDA	Genetic Dsicriminant Analysis
GSR	Galvanic Skin Response
HCI	Human-Computer Interaction
HF	High Frequency
HHT	Hilbert-Huang Transform
HMM	Hidden Markov Model
HOG	Histogram of Oriented Gradient
HOS	High Order Statistics
HR	Heart Rate
HRV	Heart Rate Variability
HVHA	High Valence High Arousal
HVLA	High Valence Low Arousal
IADS	International Affective Digitized System
IAPS	International Affective Picture System
IBI	Interbeat Interval
ICA	Independent Component analysis
IMF	Intrinsic Mode Function
kNN	k-Nearest Neighbour
LDA	Linear Discriminant Analysis
LA	Left Arm
LBP	Local Binary Pattern
LF	Low Frequency
LL	Left Leg
LPC	Linear Predictive Coding
LPCC	Linear Preditive Cepstral Coefficient
LSTM	Long Short-Term Memory Networks
LS-SVM	Least-squares Support Vector Machine

LTP	Local Ternary Pattern
LVHA	Low Valence High Arousal
LVLA	Low Valence Low Arousal
MAD	Mean Absolute Deviation
MAHNOB-HCI	A Multimodal Database for Affect Recognition and Implicit Tagging
MCML	Maximally Collapsing Metric Learning
MESAE	Multiple-fusion-layer based Ensemble Classifier of Stacked Autoencoder
MFCC	Mel-Frequency Cepstral Coefficient
MIL	Multiple Instance Learning
MLP	Multilayer Perceptron
ML	Machine Learning
mRRI	Mean RR Intervals
MSE	Mean Squared Error
MSPCA	Multiscale Principal Component Analysis
NCA	Neighborhood Component Analysis
NB	Naive Bayes
PC	Principal Component
PCA	Principal Component Analysis
PNN	Probabilistic Neural Network
PNS	Peripherical Nervous System
PPG	Photoplethysmogram
PSD	Power Spectral Density
RA	Right Arm
RBS	Radial Basis Function
RECOLA	Remote Collaborative and Affective Interactions
RF	Random Forest
RL	Right Leg
RMSSD	Root Mean Square for Standard Deviation
RNN	Recurrent Neural Network
RRS	Rescaled Range Statistics
RSA	Respiratory Sinus Arrhythmia
RSP	Respiration
SA	Sinoatrial Node
SAE	Stacked Autoencoder
SampEn	Sample Entropy
SBS	Sequential Backward Selection
SC	Skin Conductance
SCR	Skin Conductance Response
SD_1	Standard Deviation of the Instantaneous beat-to-beat RR Internal Variability
SD_2	Standard Deviation of the Continuous Long-Term beat-to-beat RR Internal Variability
SDNN	Standard Deviation of Normal to Normal (NN) Intervals
SDSD	Standard Deviation of Successive Differences
SFFS	Sequential Floating Foward Selection
SFS	Sequential Foward Selection
SKT	Skin Temperature

SMI	Symphathetic Modulation Index
SNS	Somatic Nervous System
STD	Standard Deviation
SVI	Symphatovagal Balance Index
SVM	Support Vector Machine
SWELL-KW	SWELL Knowledge Work Dataset for Stress and User Modeling Research
VLF	Very Low Frequency
VMI	Vagal modulation index
WESAD	A Database for Wearable Stress and Affect Detection
WT	Wavelet Transform

Chapter 1

Introduction

1.1 Emotion Recognition

Affective computing is a growing field, involving disciplines like engineering, psychology, cognitive science and even sociology, that explores how the use of technology enables the recognition and interpretation of human emotions or affects. Emotions are considered the way to communicate beyond words. In this way, the automatic recognition of emotions may be an exciting opportunity for human-computer interaction or even the gaming industry. The exponential increase of smart technologies in society brings the desire to make them even more custom-made, by assessing the needs of their owner and choosing the most appropriate solution. In this way, machines with the capability of assessing human emotions may revolutionize the human-robot interaction [1].

However, other areas besides robotics [2; 3] may be interested in this automated emotion evaluation such as marketing, education or even the entertainment industry, as already mentioned. In marketing, advertisement can become more effective by taking into account the emotional state of customers [4]. As far as education is concerned, learning processes may be improved by analyzing emotional responses from students [5; 6]. Finally, regarding the entertainment industries, like gaming, more appropriate entertainment can be proposed by assessing the audience's response to the ones already available in the market [7].

Another example is call centres that could largely benefit from their customers' emotion recognition, allowing them to detect problems and maximise customer satisfaction. [8]. However, the design of such robust algorithms for automated recognition of human emotions is still a major challenge [9].

Emotions are complex processes, including feelings, body language, cognitive reactions and behaviour or thoughts [10]. Different models have been proposed for automatically recognise emotions, taking into account the way all of these processes may interact with each other. However, there's still no universally accepted formulation to model emotions. Nevertheless, to use engineering principles to the recognition of such a personal and non-exact parameter as the human emotion, affects need to be conceptualized in a clear and strict way. There are mainly two categories regarding the way emotional models are conceived: The first one considers *Discrete*

Emotional Models (DEM), focusing on the six most basic emotions (happiness, sadness, anger, fear, surprise and disgust). This method considers that, regardless the situation or culture, humans perceive the environment and react to it in a similar way and with a distinguishable emotion. In this way, the main goal is to identify and label standard emotional states. On the other hand, *Affective Dimensional Models (ADM)* characterize an emotion as a set of parameters, forming an n-dimensional emotional space, and with arousal, valence, and dominance being the most commonly used dimensions [9] [11].

Arousal can be described as the measure of emotional stimulation (activation level) and it can vary from low/calmness to high/excitement. *Valence* is a measure of pleasure, defined by a polarity of positive or negative feelings. Finally, *dominance* is related to the subject's feeling of control, indicating if the human feels without control or empowered (dominant *vs.* submissive) [12]. For example, fear has a high level of arousal, negative valence and is a submissive emotion [13]. Having this said, pattern recognition approaches must be applied for affect recognition, relying on the acquisition of data with different affective states from subjects experiencing a given situation. Different modalities have been used to collect this data, from facial expressions, peripherical physiological signals or even speech intonation. Regarding the emotional stimuli, it may include media, like video, audio or even music clips, as well as the creation of different kinds of environments (relaxed, stressful, among others) [14–18].

Nowadays, most experiments developed in order to recognize emotion focus on behavioural (visible or audible) modalities [19]. Human communication is composed of verbal and non-verbal components, that are able to carry emotional information. Daily, humans rely on their own interpretation of facial and speech tone to infer the emotional states of other people. Thus, this recognition relies on the techniques used by humans to understand each other [8; 20–22]. Although emotion recognition from facial expressions or voice tone has been improved, there's still some uncertainty in the use of those kinds of evaluation parameters, since they might be purposely faked and altered. Individuals can consciously regulate their own emotions or naturally suppress them, without even acknowledging there are doing it [23; 24].

On the other hand, physiological signals are acquired in a more unconscious mode, allowing for a more trustworthy data collection. Physiological signals should provide relevant insights on emotion, since they're associated with the autonomic nervous system's responses [25; 26]. Emotion is both a psychological and a physiological expression, associated with mood, personality and all the cognitive processes involved. Besides, these signals are recorded in a continuous way, which enables detecting emotion variations through time [27]. Different physiological signals have been used in order to efficiently detect emotions, like electroencephalogram (EEG) [28], electrocardiogram (ECG) [29; 30], galvanic skin response (GSR) [26], electromyogram (EMG) [31], among others. The electrocardiogram (ECG) is a powerful signal, being considered one of the most used diagnostic tools in medicine. Recent researches have proven to be a prospective technique for emotion recognition, allowing to measure signals that can be affected by changing emotional states.

1.2 Emotion Recognition Using Physiological Signals: Challenges and Opportunities

As already mentioned, one of the main benefits of detecting emotions using physiological signals is that they are involuntary and uncontrollable reactions of the body, and thus, difficult to hide or mask [32].

The human heart is known to be affected by what people feel. In a non-scientific way, people refer to the heart as what aches when something bad happens, or even what they give to someone they love. Expressions like “heartbreak”, “big heart”, “heart of stone” or “heavy heart” are just some examples of different ways of expressing a variety of emotions. However, in a more logical view, the heart is also associated with what humans feel. When people are nervous, the heart beats faster, when they’re relaxed, the pulsation goes down. So it becomes reasonable to think about the possibility of detecting variations in emotions and affect through the heart’s changing behaviour [23].

The electrocardiogram is a physiological signal obtained by the cyclic contraction and recovery of the heart[33]. In this way, the electrocardiogram can be seen as a code, conveying all the changes correlated to the humans’ emotional states, especially by variations in the heart rate and other indicators. Various studies show that the use of ECG signal provides relevant information that can be correlated with emotion [34].

However, there are still several drawbacks regarding the use of physiological signals and specifically the electrocardiogram. Although it is accepted that physiological signals are affected by emotions, the effects on the waveform patterns are still unknown and not well defined. For this reason, many researchers are still trying to find the most emotion-related features, that can provide clearer information.

Besides, the experimental protocols are far more complex than those from behavioural emotion research. This problem is related to the need of obtaining high-quality physiological data, which is difficult, and the requirement of obtaining genuine emotions, which depends largely on the emotion elicitation material. The signal acquisition is also a challenge, since it is a more invasive process, with the sensors in contact with the human body during the recording session. Thus, proper methods and equipment should be chosen, having into account all the challenges mentioned, and trying to diminish these problems. Nonetheless, all the limitations mentioned above lead to commonly small and restrained databases, with a reduced amount of data, which can be highly problematic for deep learning approaches to be developed.

Another drawback is related to the labelling process, since physiological signals are subjective and it is difficult to establish their ground truth [23; 32]. Moreover, most databases and experimental protocols don’t have a large number of subjects nor a relevant number of signals per subject. Due to this limited amount of data, the performance of the classifiers applied is compromised. To solve this problem, more subjects can be included in the experimental protocols, which should also become longer so that the number of data increases, or more samples can be used (for example, by considering multiple segments of the same physiological signal). The use of segments is

highly recurrent in order to obtain a "fake" data enlargement, however, this multiple segmentation of the same signal implies that the same label is considered to all segments, which can sometimes be misleading. Furthermore, since random splits are usually applied, segments from the same signal are found in train and test sets, which can result in misleading performances presented in the literature, that must be evaluated under more controlled and realistic settings.

Having this said, signal and subject-independent settings can be one of the main problems regarding physiological signals, mainly due to the lack of large amount of data. Furthermore, concerning the ECG signal that, in normal conditions, presents the same behaviour and deflections for all subjects, it is considered to have a high degree of variability. This variability can be intra-subject, consisting of variations within the consecutive heartbeats of the same subject, or inter-subject, corresponding to variation between heartbeats of different subjects [33].

Especially the inter-subject variations can become a problem when the major goal is to detect specific and equal patterns to equal states of emotion. Having this said, the major problem nowadays is associated with this pattern detection when data shows a wide range of behaviour to the same affect felt by different people. When in biometric research, these inter-subject features can be highly positive, allowing for a more efficient recognition of each individual. However, when working to correlate an emotion to a given pattern, it can be challenging.

1.3 Disseration Goals and Contributions

As it can be understood, there is a variety of opportunities and challenges in using physiological signals, more specifically ECG, for emotion recognition. The work done during this dissertation aimed to explore the use of ECG and all of its good features to allow emotion recognition, trying to minimize the challenges presented. Nonetheless, more strict data settings, such as signal and subject-independent were evaluated to conclude about the extent of dependency associated with the approaches available nowadays.

Having into account the problem recurrently found in the use of ECG signals to emotion recognition, it becomes important to develop techniques that may be able to overtake this kind of issues, such as considering the enlargement of the databases available for emotion recognition. Thereby, the main goal of this work has been the development of robust deep learning methodologies for emotion recognition using ECG, assessing its behavior and performance in more demanding settings such as signal and subject-independent ones.

With the work developed in this dissertation, we contributed to the field of ECG-based emotion recognition by identifying the limitations of the state-of-the-art, especially regarding the data availability, diversity, and the realism of evaluation setups. Moreover, we build upon a promising literature methodology to surpass these limitations, bringing the field closer to real applications of ECG-based emotion recognition.

1.4 Dissertation Outline

To enable achieving the aforementioned goals, this dissertation includes, beyond this introduction, a review of the fundamentals, concerning both the electrocardiographic signal and the emotion recognition theory and principles, in Chapter 2. Regarding the electrocardiographic signal, its variability and main acquisition methodologies are analysed and discussed. As far as emotion recognition is concerned, the existing emotion models are presented and detailed, as well as the different modalities used for emotion recognition.

In Chapter 3, a review of the most relevant databases available is done, describing and comparing them. After that, Chapter 4 presents a state of the art analysis, where some of the most relevant approaches used in emotion recognition nowadays, through machine and deep learning techniques, are described and further discussed.

In Chapter 5, the replicated approaches are presented, being thoroughly described. In addition, the results obtained are illustrated and discussed, to take some important conclusions.

Chapter 6, focus on the improvement of the self-supervised learning approach, previously presented in Chapter 5. Furthermore, different data settings, such as signal and subject-independent are considered and used to evaluate the model developed.

In Chapter 7, other deep learning structures are taken into account and further added to the network in development, to obtain the final deep learning architecture of this dissertation. Once more, after presenting the results of these additions and improvements, different data divisions are considered, analysing their impact on the model performance.

Finally, in Chapter 8, final remarks are considered, such as an overall discussion and analysis of this dissertation's work and the current position of the emotion recognition field, analysing the achievements and its evolution state, as well as pointing out its limitations. In this way, future work is also discussed, indicating some paths to be explored in further investigations.

Chapter 2

Fundamental Concepts

2.1 Affective Computing and Emotion Recognition

Emotion can be described as a "strong feeling deriving from one's circumstances, mood, or relationships with others" or even as an "instinctive or intuitive feeling as distinguished from reasoning or knowledge" [35]. It is common sense that emotions affect both human physiological and psychological status, playing an important role in a person's daily life and how someone may deal with his own happiness, victories, or even losses or disappointments. However, it is also true that emotions are still believed to be inherently non-scientific sensations, far from rational thought or common logic, being marginalized from science.

However, this public perception is not completely correct. Although emotions arise from a human's ability to feel and process what happens around him, they also have a large impact as far as essential cognitive processes are concerned. Lisetti [36] highlights results from neurological literature indicating how emotions are not only associated with human creativity and intelligence, but also with basic rational thinking and decision making.

Furthermore, Lin Shu *et al.* [32] defines emotion as a "mental state, that arises spontaneously rather than through conscious effort and is often accompanied by physical and physiological changes that are relevant to the human organs and tissues such as the brain, heart, blood flow, muscle, facial expressions, voice, etc", pointing out the close link between emotions and rather scientific parameters like physiological signals and basic biology functioning.

With the exponential growth of human-computer interactions (HCI), it becomes reasonable to consider that the interpretation of emotions is an important and fundamental asset to enable accurate and efficient intercommunication [34; 36]. From this need, Affective Computing (AC) emerges as a new and exciting field that can enable smarter and more humane technology.

AC tries to assign to computers the ability to perceive, observe and even generate affect responses, increasing human-computer communication quality. This field is currently one of the most emerging and active topics in research, due to its large and promising spectrum of applications, in different areas. Affective Computing involves multidisciplinary knowledge like psychology, physiology and computer sciences [37]. Its great potential translates into a wide range

of applications in several environments, like robotics, marketing or even education. The development of recommendation systems also becomes possible since Affective Computing will allow understanding customers preferences and opinions.

Within Affective Computing, two distinct research areas can be identified: sentimental analysis and emotion recognition [38]. While Sentiment Analysis consists of a simple classification task between a positive/neutral/negative state, emotion detection involves a more detailed and thorough methodology that aims to distinguish between a set of emotions. In this way, product reviewers can use Sentiment Analysis to assess how the product was perceived by its customers (liked or not). On the other hand, Emotion Recognition is able to differentiate if someone is angry, sad or bored. As it can be understood, these two areas can overlap, since a happy customer translates into a positive reaction. However, Emotion Recognition is a finer discrimination and assessment of a general reaction to a given stimulus. This scientific analysis of emotions and the capability of recognizing them depends on a correct emotion modelling, followed by proper data collection and the application of specific methodologies, which will be further explained in the following sections.

2.1.1 Emotion Modelling

Emotions have been studied for centuries and philosophical studies around them can date back to the ancient Greeks and Romans. The desire of defining emotions as quantitative parameters, in order to evaluate and assess them in a more scientific and logical way is as old as it can be, and started with Cicero, when he tried to organized emotions into four basic categories: *metus* (fear), *aegritudo* (pain), *libido* (lust) and *laetitia* (pleasure). In the late 19^{th} century even Darwin focused some of his attention on emotions, developing a theory that stated that emotions evolved via natural selection [39]. However, since emotions are complex processes, involving body language, cognitive reactions, feelings and thoughts, modelling them is a very challenging task. Besides these first attempts, psychologists have been trying to develop a logical and universally accepted emotion model for decades, however, this goal hasn't yet been completely achieved. Nevertheless, the main scientific models developed and used in Emotion Recognition are the Discrete Emotional Models (DEM) and the Affective Dimensional Models (ADM).

2.1.1.1 Discrete Emotional Models

In the early 1970s, Ekman [40] conceived emotions as discrete and measurable categories, being psychological states that fit specific criteria. Six basic emotions were defined: happiness, sadness, anger, fear, surprise and disgust. All other emotions were considered to be combinations or reactions to these.

Discrete models suggest that affective states are universal among people, regardless of their differences in gender, origin, or even moral beliefs. Also, when in a similar situation, people would react in the same way, showing analogous physiological expressions and patterns [11].

However, in 1980, Plutchnik [41] extended this basic emotions list, by considering eight emotions of joy, trust, fear, surprise, sadness, disgust, anger and anticipation. Figure 2.1 presents the wheel model proposed by Plutchik, which describes emotions ranked by intensity, with the stronger ones in the centre and the weaker positioned at the flower blooms. The colour use represents that basic emotions can be combined to form complex emotions.

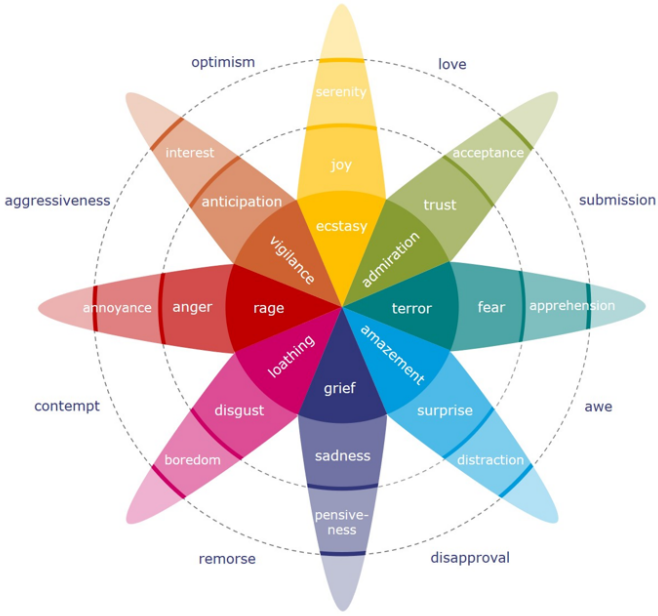


Figure 2.1: Plutchik's Wheel of Emotions, from [42].

Finally, Izard [43] defined ten basic emotions: joy, surprise, sadness, fear, shyness, guilt, anger, disgust and contempt. He hypothesized that these emotions were formed during human evolution, being a simple brain circuit.

Although these simplistic models would allow an easier view of emotions, they become unsatisfactory and limited when analysing complex emotions, which cannot be described with a discrete label and need a more quantitative evaluation [32].

2.1.1.2 Affective Dimensional Models

In 1989, the first dimensional model was proposed by Russel *et al.* [44], showing that there would have to exist some evaluation parameters that could describe emotions, like intensity or positiveness. These models suggest that emotions are not discrete but continuous, and, according to Lang *et al.* [45] , they can be projected into a two-dimensional space model, by *valence* and *arousal*. *Valence* is a measure of pleasure, ranging from negative/unpleasant to positive/pleasant. On the other hand, *arousal* describes the level of activation and intensity concerning emotional stimulation, varying from low/calmness to high/excitement, or even passive and active. In this way, as shown in Figure 2.2, different emotions can be plotted, having a correspondent value of *arousal* and *valence*. For example, sad has a negative valence, indicating that it is not a pleasant

emotion, and a passive arousal, proving a low activation level. On the other hand, angry is also a negative valence emotion but it presents a high level of activation and excitement, thus having an active arousal.

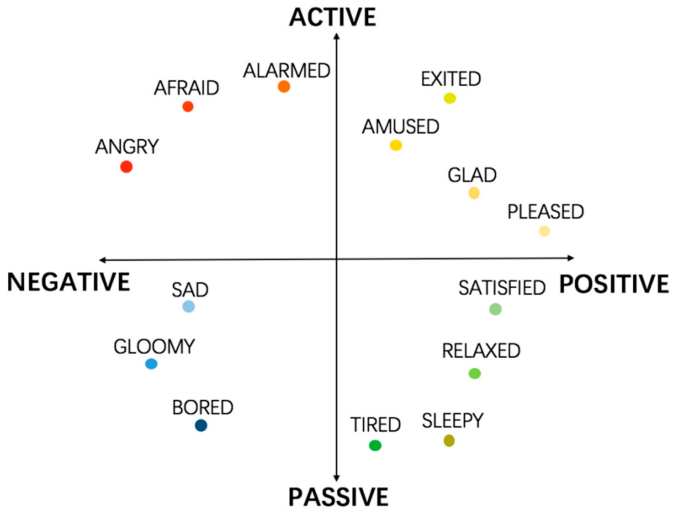


Figure 2.2: Two-Dimensional Space Model of Emotions, from [32].

This dimensional grid became highly attractive due to its higher flexibility in describing an emotion besides using just words, and the possibility of considering emotional variations over time. Other metrics emerged through time like *dominance*, as it can be seen in Figure 2.3 [46]. *Dominance* corresponds to someone's level of control, varying from submissive to dominant. Having into account other measurements to characterize emotions like this one, it becomes possible to obtain a more complete discrimination between emotions. For example, while fear and anger both have negative valence and active arousal, dominance allows to divide them into the submissive and dominant axis, respectively [32].

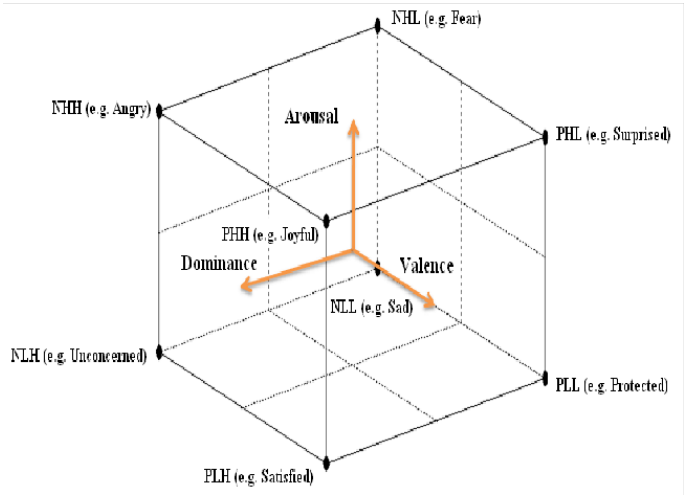


Figure 2.3: Three-Dimensional Space Model of Emotions, from [47].

This <arousal, valence, dominance> set is also known in the literature by different terminologies, like <evaluation, activation, power> or <pleasure, arousal, dominance> [48]. Furthermore, there are already research studies concerning a fourth dimension. Fontaine *et al.* reports consistent results using a set of four dimensions, <valence, potency, arousal, unpredictability> [49].

In addition, another affective categorization model appeared, inspired by Plutchik's studies on human emotions. The Hourglass Model [50] reinterprets Plutchick's model and organizes primary emotions around four independent dimensions, which present different levels of activation that allow to translate and have into account the entire scope of affective states. This model has an hourglass shape and is presented in Figure 2.4.

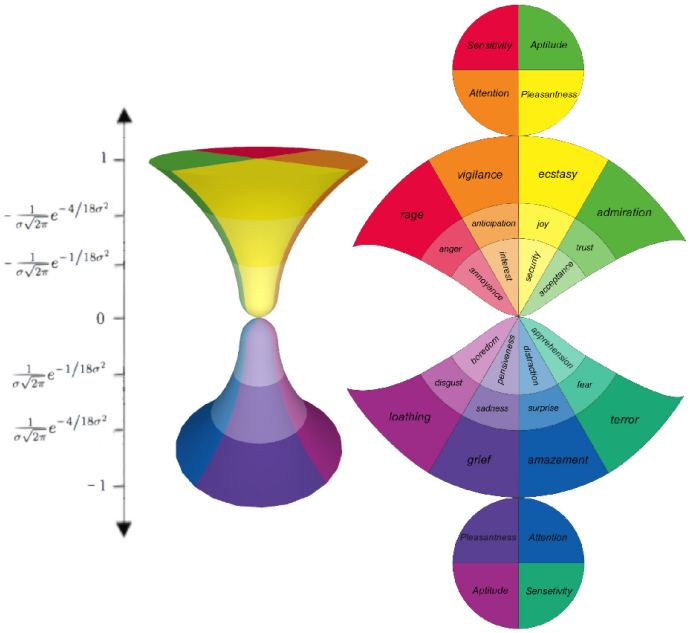


Figure 2.4: The 3D model and the net of the Hourglass of Emotions, from [50].

This 3D model represents emotions and affective states according to their strength, where the vertical dimension consists of emotional intensity and the radial dimension relates to the activation of different emotional configurations. In this way, this model attempts to reproduce the full range of emotions by considering Pleasantness, Attention, Sensitivity and Aptitude, being able to consider situations where up to four emotions are expressed and felt at the same time.

However, new and interesting models are being proposed and developed and we can only hope for more interesting and effective ways of differentiating and characterize these complex feelings to which we call emotions.

2.1.1.3 Discrete *vs* Dimensional Models

Concerning Discrete and Dimensional models, it is important to understand that both have advantages and disadvantages in their attempt to distinguish between emotions. Regarding the Discrete Emotional Models, as it can be easily understood, there are a variety of factors that can play an important role on someone's behavior or emotional response, which means that it is wrong to

consider that all people would act in the same way when exposed to the same environment or situation.

Furthermore, emotions are also complex processes, related to both physiological and neural activities, and, for this reason, it is limiting to consider a small list of basic emotions. On the other hand, even the Affective Dimensional Models are not enough to translate the emotional complexity already described. Besides assuming independence between axis, eliminating all possible relations between dimensions, which can also be a little uncertain and incorrect, these models do not consider the possibility of two or more emotions being felt at the same time. However, some new dimensional models are emerging, such as the Hourglass Model, trying to reproduce the entire range of emotions and considering situations where more than one emotion is being felt at the same time.

2.1.2 Emotion Recognition Modalities

Concerning Emotion Recognition and Affective states differentiation, it is essential to find a way of measuring the subject's emotional variations over time. As already mentioned, emotions are expressed by people in their daily life, through their bodily reactions, facial alterations or even the changes that can be detected in someone's voice or intonation. Yet, emotions are also closely related to neural activity and the body's physical and physiological states. In this way, different modalities can work as emotion indicators and be used to detect affective variations. Having this said, emotion recognition methods can be classified into two main categories. One is by using physical and more visible signals such as facial expressions, speech alterations, gestures or posture. The other has into account physiological signals, including EEG [28; 29; 51], ECG [24; 29; 30; 52], GSR [26; 53; 54], EMG [31; 55], HRV [37; 56], among others. The present section focuses on the current modalities used for emotion recognition, analyzing the advantages and disadvantages of each.

2.1.2.1 Facial Emotion Recognition

In everyday life, humans rely on their own perception of emotions through the analysis of people's facial expressions and body movement. Especially the latter plays an important role as the main non-verbal communication channel. The face conveys diverse information regarding someone's age, sex, identity, background or even what they are feeling. In this way, it is not surprising that various behavioral scientists showed interest in how facial expressions could indicate a large amount of information from someone [57].

At the beginning of 1970, computer scientists started to use the face as a biometric modality, since it is the main natural method to recognize a given person. Later, the idea was to analyze and synthesize facial expressions. Ekman and Friesen developed one of the leading methods in facial expression recognition, the Facial Action Coding System (FACS) [58]. This system describes facial expressions by considering individual muscle movements, denominated Action Units (AUs) - the smallest movements that can be distinguished - that allow detecting micro-expressions.

Ekman's work inspired many researchers to study facial expressions using both images and videos. With time, image processing techniques started to be used and applied in facial emotion detection. Nowadays, Facial Emotion Recognition (FER) techniques are normally composed of three major steps: preprocessing, feature extraction, and classification.

Preprocessing consists of a preliminary process that can be used to improve FER performance. It is normally done before the feature extraction process, facilitating both extraction and further classification. Regarding images, most common preprocessing methods include image clarity, contract adjustments, scaling and other possible enhancements [59]. After that, some features are extracted and fed into a classification system. Finally, the output consists of one of the pre-selected emotion categories or labels.

However, it is possible to divide FER techniques into two groups, considering if features are handmade or automatically produced through a deep neural network. As far as handcrafted features are concerned, they can be further discriminated in "feature-based" and "region-based". "Feature-based" consists of detecting specific facial structures like the eyes or the corners of the mouth. On the other hand, "region-based" focuses on specific regions of the face to detect expression variations. For example, this analysis can be done by taking into account the eye/eyebrow and mouth regions, discarding the rest of the face. Figure 2.5 presents the three main steps in conventional FER approaches: facial structures detection, feature extraction and expression classification [60].

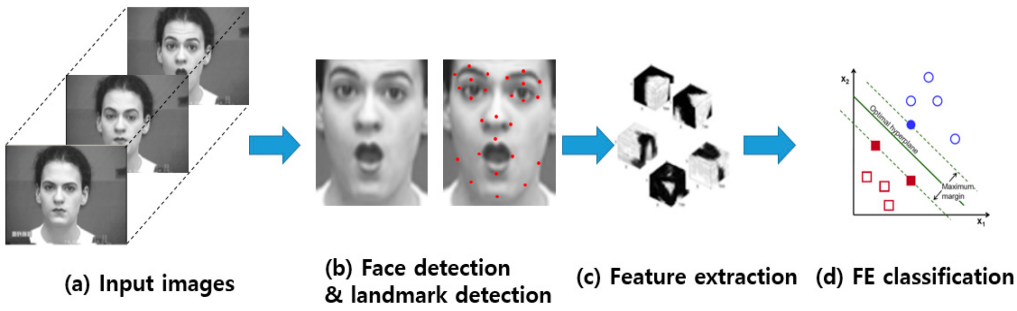


Figure 2.5: Procedure used in conventional FER approaches, from [60].

However, Deep Learning approaches present themselves as interesting techniques, in which features are automatically extracted, with no previous work or even knowledge regarding the extraction of specific and relevant features. In the literature, different techniques are presented using convolutional neural networks (CNN) [61], deep 3D convolution networks (3DCNN) [62] or even recurrent neural networks (RNNs) [63; 64].

Having this said, Deep Learning methods are able to reduce the dependence on correct preprocessing and feature extraction to obtain positive results. However, these methods ask for extensive datasets or methods like data augmentation, high computing power and are more time-consuming. To summarize, facial expressions are still one of the most chosen and used modalities

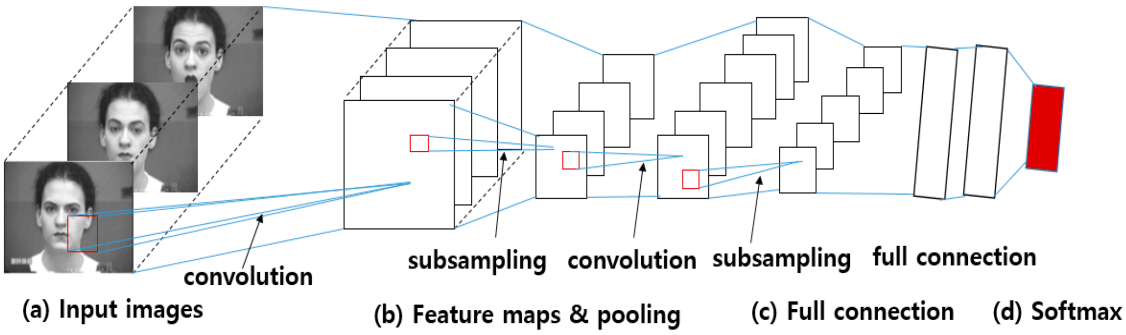


Figure 2.6: Example of a CNN-based FER approach, from [60]

for emotion recognition. It seems like a natural and practical way of analysing how someone is feeling, since humans commonly use it as such in their daily lives. However, as it is known, people are capable of faking reactions and emotions, by altering facial expressions [65]. This fact translates into a more untrustworthy modality as far as emotion recognition is concerned.

2.1.2.2 Speech Emotion Recognition

Speech is the most natural way of communication and interaction between humans. In addition, there is a lot of information regarding how people are feeling through their way of talking, not only by what someone says but also having into account the speech intonation, trembling or firmness. When people are nervous, their voice can become more shaking, however, when in an excited state of happiness, people tend to talk louder and wider.

These facts motivated different researchers to study speech alterations and relate them with possible emotion variations. Since the 1950s, different research has been developed on speech recognition, by converting human speech into words. However, when it comes to speech emotion recognition, research and achievements are still few [66].

Speech emotion recognition consists of extracting the emotional state of a given person by analyzing his or her speech. This goal can be achieved by using audio or transcriptions into text processing. However, it is not an easy task due to the variability introduced by different speakers, speaking styles and speaking rates [67].

Similarly to Facial Emotion Recognition and other classification problems, the initial step to be done, after possible preprocessing, is feature extraction, by identifying suitable features that may efficiently characterize emotions. These features can be *local*, by dividing the signal into small segments and analysing them in a stationary way, or *global*. Features like pitch and energy can then be extracted from each frame and further used to detect emotion variations. *Global* features are calculated as statistics of all speech features. Although there is some disagreement on which type of feature is better, most researchers consider that global features are more effective in terms of classification time and accuracy [68].

Speech features can also be grouped into four categories: continuous, qualitative, spectral and TEO-based features (Teager energy operator). In Figure 2.7, different examples of each category

are presented. However, most approaches combine features belonging to different categories.

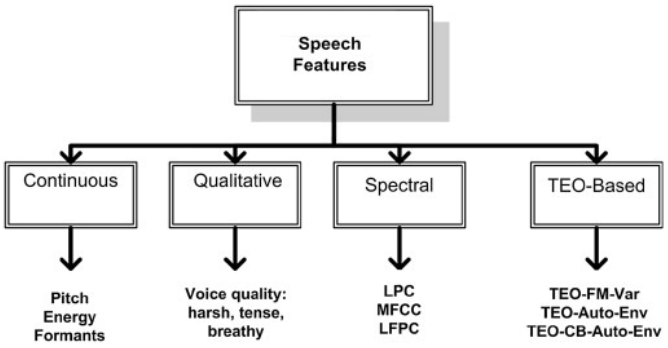


Figure 2.7: Five categories of speech features, from [68].

After having a front-end unit responsible for feature extraction, there's the need for a suitable classifier to perform the proper emotion recognition from each input considered. In this way, Figure 2.8 represents a possible method of speech recognition. Deep learning methods are also used for speech emotion recognition, avoid the need for manual feature extraction.

Some of the most commons classifiers used are Support Machine Vectors (SVM) and Artificial Neural Networks (ANN). However, HMM (Hidden Markov Model) is the most used classifier in speech emotion classification and recognition [69; 70].

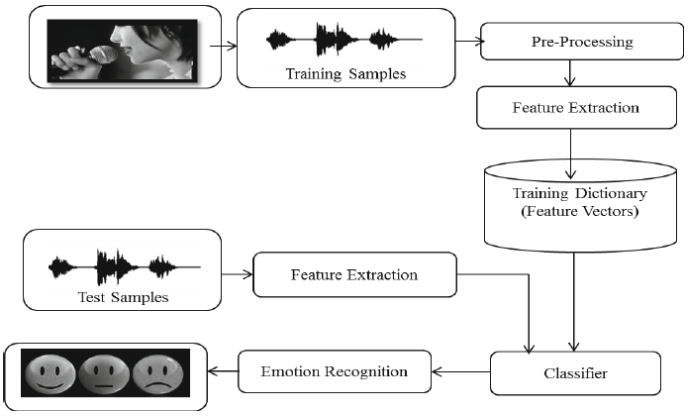


Figure 2.8: Conventional Architecture of Speech Emotion Recognition System, from [71]

Although some good results have already been achieved, this field is still attractive and full of opportunities to be improved, by tackling problems like speaker-dependent classification. However, speech emotion recognition is not always completely trustworthy. Although it cannot be so easily manipulated and altered as facial expressions, it is possible to control parameters such as rhythm or intensity, in order to mask some emotional states.

2.1.2.3 Body Gestures and Posture

Comparing with affect analysis from speech and facial expression variations, body gestures and posture are regularly overlooked and diminished as strong emotion indicators. However, the power of body language has been gaining more and more attention, and it is already common sense that body gestures can provide information regarding if someone is, for example, comfortable or nervous. According to Mehrabian's 7-38-55 principle, the percentage distribution of a message is 7% verbal signals and words, 38% strength, height and rhythm and 55% facial expressions and body posture [72].

According to [73], body language includes a variety of non-verbal indicators, from facial expressions to body posture, eye movement, or the use of personal space. The hands are also a great source of body language information [74], and nowadays, it is starting to receive some attention, being used by politicians during their speeches or debates. In the same way, details like the head positioning or chin lifting angle can also be a source of information, conveying emotions and intents [75].

More importantly, body gestures and posture are generally natural and unintentional, unlike facial expression or speech intonation, which can be purposely masked or faked. Thus, increased interest in using body language to emotion recognition started to grow and be further researched.

As far as body analysis is concerned, human modelling is an important preparation step, that may influence all the subsequent phases. The most common ways of modelling the human body are either as an ensemble of body parts as a kinematic model (see Figure 2.9). In the first way, different body parts are independently detected and some restrictions can be applied to refine the detection. On the other hand, the kinematic model consists of a collection of interconnected joints with predefined degrees of freedom identical to the human skeleton [76].

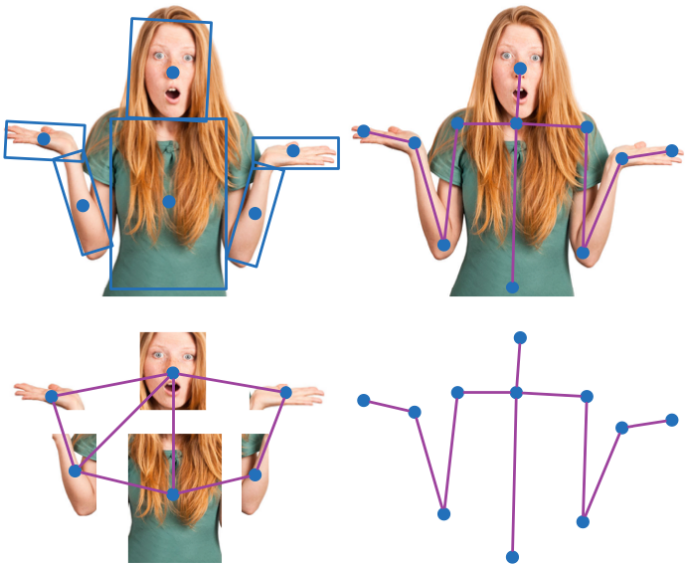


Figure 2.9: Two most common ways of modelling the human body: *model based on ensemble body parts* (left) and *kinematic model* (right), from [76]

Both human modelling and body posture detection are done by using a variety of possible methods, like the use of histogram of oriented gradient (HOG) features for object detection [77] or the use of support vector machines (SVM). One of the most known methods is the Deformable Part Models (DPM) [78], which consists of a set of parts and connections of a given geometry that are separately processed, instead of considering the totality of the body. More recently, deep neural networks started to be widely used, showing a rise in performance in pattern recognition.

Since the human body can be in movement and not static, there's also the need to model the temporal dimension and track the body. Probabilistic directed graphs like Hidden Markov Models (HMM) are highly used in body gesture recognition and activity analysis [79; 80], as well as dynamic Bayesian Networks [81].

The final step of Emotion Body Gesture Recognition is representation learning, by building a relevant representation that can be used to learn a mapping to the corresponding targets [76]. Finally, a classifier can use these body representations to extract meaningful information and perform emotion recognition (see Figure 2.10).

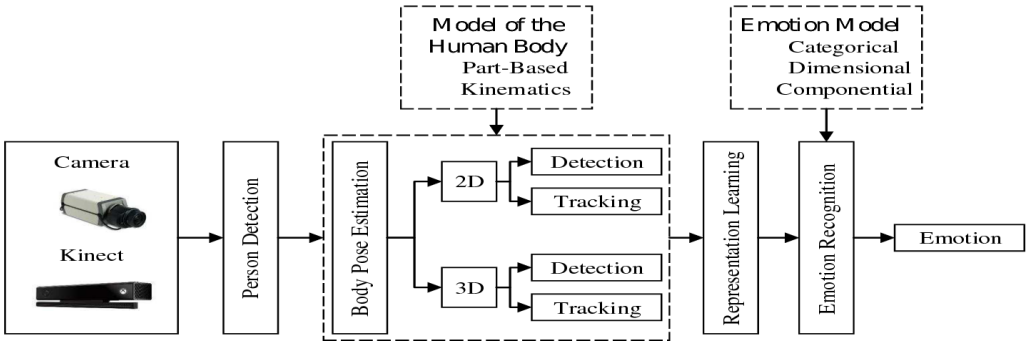


Figure 2.10: General overview of an Emotion Body Gesture Recognition system, from [76]

Gunes and Piccardi [82] decided to use Naïve representation from the upper-body to classify body gestures into 6 emotions. On the other hand, Saha *et al.*[83] used different methods, like compared binary decision tree (BDT), ensemble tree (ET), k-nearest neighbour (kNN) and SVM, in order to distinguish between 5 basic human emotional states. Kosti *et al.* [84] used a CNN for extracting features from the body and background and using them to recognize among 26 different emotions.

Besides the ones mentioned, other classifiers and methods were already used and further investigated. However, there's a high tendency of using body gesture and postures in combination with other emotion recognition modalities, like face or speech, in order to obtain more complete information and a better classification method [85; 86], since body gestures are not always complete and enough to interpret emotional states.

2.1.2.4 Physiological Signals

The nervous system is composed of two distinct parts: the central and the peripherical nervous systems (CNS and PNS). Considering the PNS, it is divided into the autonomic and somatic nervous

systems (ANS and SNS). The ANS is associated with the sensory and motor neurons, allowing the connection and communication between the CNS and all internal organs. In this way, physiological signals respond to both the CNS and the ANS, which implies their natural and unconscious nature. Nowadays, the autonomic nervous system (ANS) activity is viewed as a determinant component of the emotional response in a variety of recent theories concerning emotion [87]. Although there's still some disagreement regarding the degree of specificity of ANS activation in emotion, it can be inferred that emotions have an impact on ANS, which, in its turn, influences physiological signals.

Psychophysiology is the branch of psychology that correlates psychological states with physiological variations and measurements. Although someone can choose not to externally show his/her emotion, there's an inevitable change in physiological signals that cannot be hidden or avoided, since the ANS sympathetic nerves are activated whenever a person is positively or negatively stimulated [88]. This sympathetic activation leads to heart, respiratory and blood pressure rate variations, considered some of the most common reactions of the human body to a given emotion [89].

In 1983, Izard and Fridlund were responsible for one of the earliest works considering emotion recognition and physiological signals. They used Linear Discriminant Analysis on facial EMG activity, which became a landmark research since it was able to prove the existence of a correlation between physiological data and emotional states [90]. After that, a great number of studies regarding emotion recognition using physiological signals have been conducted, which can be found in the literature. However, it is still uncertain how emotion variations translate into actual pattern alterations in each physiological signal.

It is highly relevant to have an efficient and successful data collection when developing an emotion recognition system based on physiological data. Unlike facial or speech variations, physiological signals are harder to collect since they are highly prone to noise and ask for more complex setups. On the other hand, image and video can be easily acquired even by non-specialists [91]. Since the ANS controls physiological signals, there's the need to naturally induce a given emotion. For this, different emotion elicitation techniques can be used, like pictures [92], videos [12] or even music [27].

Regarding the specific methodology of emotion recognition using physiological data, it can be divided into two major categories, being the first the usage of more traditional machine learning methods, and the other using deep learning methods.

Like it was already mentioned in other modalities, automated classification tasks normally rely on these two options (see Figure 2.11). The first one implies handcrafted feature extraction and optimization, while deep learning methods learn from data that was practically unaltered. Especially when considering end-to-end approaches, preprocessing is almost non-existent since the network should be able to deal with the raw signals and extract all important patterns to recognize emotions. In this way, using raw signals makes the network more robust in its detection [19; 93].

In model-specific methods, in which feature extraction is done, there's the need to explore emotion-specific characteristics concerning each different physiological signal, in order to focus

on relevant features that may improve the model performance. On the other hand, deep learning approaches don't ask for such specific knowledge or challenging feature selection.

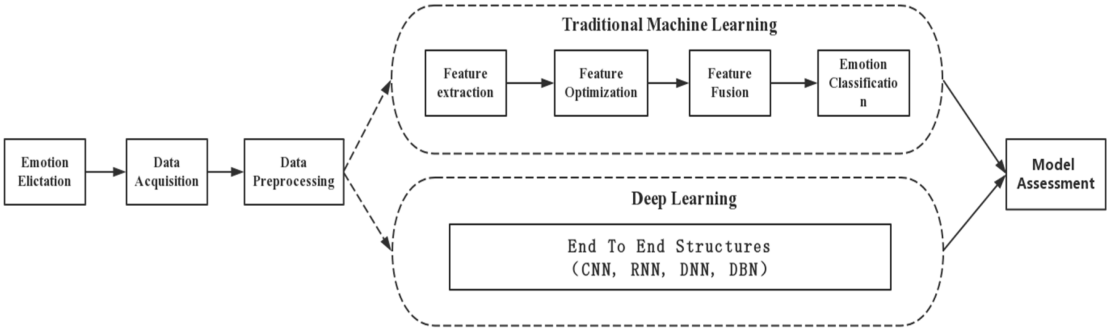


Figure 2.11: General overview of an Emotion Recognition process using physiological signals under target emotion stimulation, from [32]

Nevertheless, feature extraction still plays an important role in emotion recognition models found in the literature. Some of the most common feature extraction methods used are Wavelet Transforms (WT), Empirical Mode Decomposition (EMD) or even Fast Fourier Transform (FFT), among others. FFT can be applied to signals, for example, EEG, in order to calculate the spectrogram of each EGG channel [94]. On the other hand, WT is also highly used, since it provides a moving and non-fixed 'time-frequency' window, able to analyse the signal locally. EMD is another powerful tool, being able to decompose a signal according to its time characteristics.

After feature extraction, some features might be irrelevant or redundant and should be deleted to save some time and effort in training the model. Besides, these "repeated" features can also be one of the causes of overfitting. In this way, it becomes interesting to use some feature optimization methods available, guarantying a more effective training and performance of the model. [32]. Finally, regarding classification, there's a variety of classification models suitable for emotion recognition. In Figure 2.12 it is possible to find the main 7 classification models used.

Besides k-Nearest-Neighbour (kNN) and Support Vector Machines (SVM) [88], Linear Discriminant Analysis (LDA) [95] and Random Forest (RF) [96], other classifiers like Bayesian Networks [97] and Canonical Correlation Analysis (CCA) [98] are also used by researchers for classifying emotions.

In this way, although there's a variety of problems that need to be addressed, such as subject-dependency and the difficulties in experimental protocols and data collection procedures, physiological signals prove to be a strong modality as far as emotion recognition is concerned, since they are trustworthy and unconscious signals. Furthermore, there is scientific proof of their correlation with emotional states through ANS mediation. In particular, ECG is a promising signal, since it can be easily acquired in comparison with other physiological signals, such as EEG. In addition, it conveys a high amount of information that can be correlated with emotional states, and it is already a highly used and developed technique in the medical field.

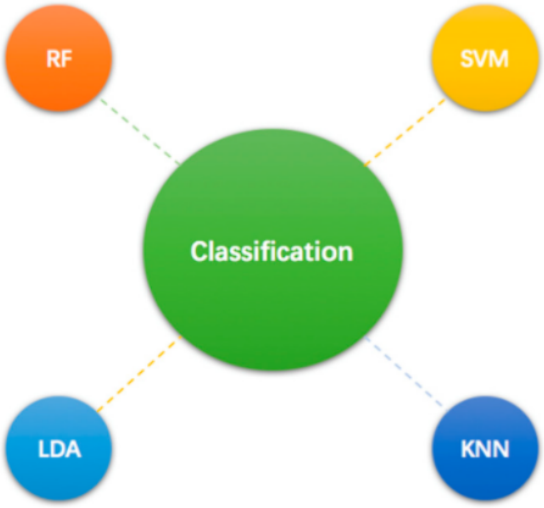


Figure 2.12: Most used classification models in Emotion Recognition methods using Physiological Data, from [32].

2.2 Electrocardiographic Signal

2.2.1 Anatomy and Physiology

The human heart is located in the mediastinum, a place within the thoracic cavity and between the lungs, slightly to the left of the sternum. This muscular organ is essential to human life since it works as a pump, by contracting and forcing the blood through the blood vessels of the body.

The heart has three major functions: (1) *generating blood pressure*, (2) *routing blood*, which allows for the separation between the pulmonary and the systemic circulation, and (3) *regulating blood supply*, in which the heart rate and its force contraction change in accordance to the human metabolic needs [99; 100].

Regarding its anatomy, the heart is enclosed in the pericardial membranes, composed of three layers. The more external is the *fibrous pericardium*, made of a dense fibrous connective tissue that protects the heart and anchors it to the surrounding structures. In the middle, the serous pericardium, a thin serous membrane, is folded into two layers: *parietal* and *visceral*. The latter is the one in contact with the surface of the heart, also known as *epicardium*, which means "upon the heart". The epicardium is considered to be the most inner layer of the heart wall, followed by the *myocardium* (middle layer), and the *endocardium* (outer layer). The myocardium is mainly composed of cardiac muscle, forming the bulk of the heart and the walls of the four chambers. The endocardium covers both the valves and chambers, preventing abnormal clotting [101].

The heart is formed by four chambers. The two upper ones are the right and left atria, which are separated by the interatrial septum. The lower chambers are the right and left ventricles, presenting thicker and stronger walls, and being separated by the interventricular septum. Caval veins carry blood from the body to the right atrium. From this atrium, the blood is then led to the right ventricle upon atrial contraction, passing through the right atrioventricular (AV) valve, also

called tricuspid. In the ventricle, the blood is pumped to the lungs, carried by the pulmonary artery. On the other hand, the left atrium receives blood from four pulmonary veins, that flows to the left ventricle. The valve separating these two chambers is the mitral or bicuspid, and after arriving at the ventricle, the blood is finally pumped to the body through the aorta, the largest human artery [102].

This sequence of events occurs in a cyclic way with each heartbeat, and it is dominated as the cardiac cycle. The electrical activity of the myocardium regulates this cycle, which is stimulated to contract without any external stimulation [103], and the electrical impulse follows a specific order and route, throughout the myocardium [99; 104] (see Figure 2.13):

1. **Atrial depolarisation** - The sinoatrial node (SA), considered the natural pacemaker of the heart, generates the impulse and is the fastest structure in the myocardium to depolarize. The electrical impulse follows its path, reaching both atria and the AV node;

2. **Atrial depolarisation complete** - After its depolarization, the atria contraction begins. Conduction through the AV node is ten times slower than through the surrounding heart tissue, delaying the impulse for 0.1s. This allows the atria to finish its contraction before the impulse gets into the ventricles and it also protects the ventricles from high atrial rates during possible atrial arrhythmias. The electrical conduction occurs through the AV node to the bundle of His, that divides into left and right bundle branches, carrying the electrical impulse into the left and right ventricles, respectively;

3. **Ventricular depolarisation** – Upon reaching the apex, the impulse spreads onto the myocardium cells through the Purkinje fibers, beginning the ventricular depolarization. Simultaneously, the atria begin to repolarise;

4. **Ventricular depolarisation complete** – the ventricles depolarise completely, and their contraction start immediately after the depolarisation;

5. **Ventricular repolarisation** – The ventricles start to repolarise;

6. **Ventricular repolarisation complete** – the ventricles repolarise completely and the cardiac cycle starts again.

The cardiac cycle is a quick process, taking a total time of approximately 0.22 s (220 ms) and, as it can be perceived, it consists of an electrical current that is generated and conducted through the heart. This electrical activity may be detected and recorded by an electrocardiograph, obtaining a graphic record of the heart activity: the electrocardiogram (ECG) [101].

A typical ECG is composed of three distinct waves: the P wave, the QRS complex and the T wave. The P wave results from the depolarisation of the atria, whereas the QRS is associated with the depolarisation of the ventricles and the simultaneous repolarisation of the atria. Finally, T-wave translates the ventricular repolarisation. The PQ or PR interval is the time between the beginning of the P wave and the QRS complex, during which the atria contracts and begins to relax. After the PQ interval, ventricles begin to depolarize.

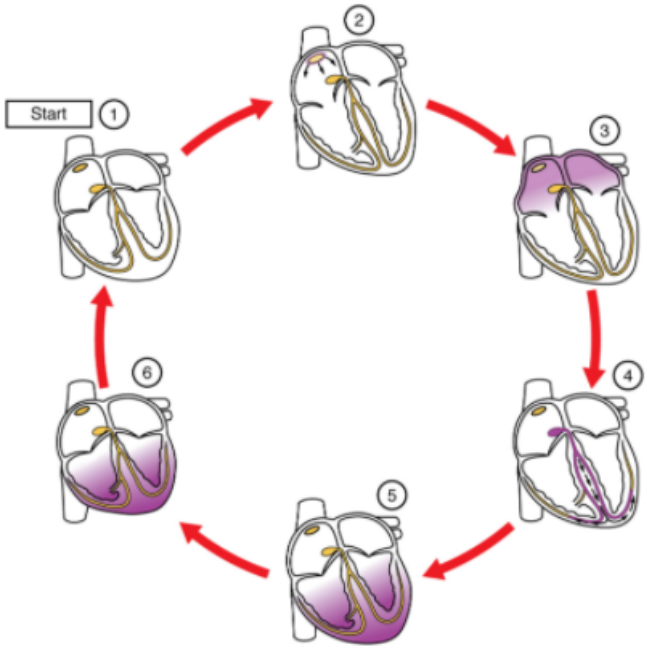


Figure 2.13: **The cardiac conduction system**. (1) The sinoatrial (SA) node and the conduction system are at rest. (2) The SA node generates the impulse, which spreads across the atria. (3) After reaching the atrioventricular node (AV), there is a delay of approximately 0.1s, allowing the atria to complete pumping blood before the impulse is transmitted to the atrioventricular bundle. (4) The impulse then travels through the atrioventricular bundle and bundle branches to the Purkinje fibers. (5) The impulse spreads to the contractile fibers of the ventricle. (6) Ventricular contraction occurs and the cycle is ready to start again, from [105]

The QT interval, which goes from the beginning of the QRS complex until the end of the T wave, represents the time needed for ventricular depolarization and repolarization. Within this interval, the S-T segment represents the time between the moment in which ventricles are completely depolarised until their repolarisation [100; 101].

All these waveforms are registered in a repeating rhythm, the sinus rhythm, originated by the SA node. Furthermore, in some people, generally with slower heart rates, a fourth waveform known as U-wave may appear. Some researchers believe that it represents late stages of ventricular repolarisation or even a post-repolarisation phenomenon [104].

2.2.2 Variability

Although in normal conditions the ECG presents the same major deflections and waves through each heartbeat for all subjects, there's a high degree of variability considering this physiological signal. There are two different kinds of variability in the ECG: *intrasubject*, which occurs when variations are detected between heartbeats of the same person, and *intersubject*, consisting of variations between heartbeats of different people [99]. According to J. R. Pinto *et al.* [33], these types of variability can be associated with different aspects being the most relevant:

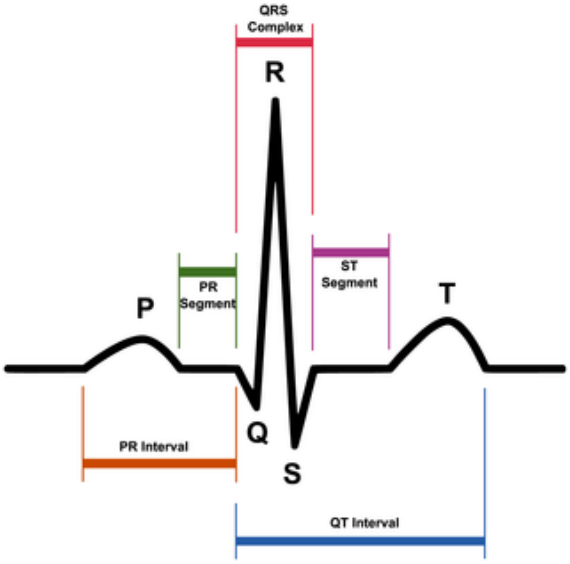


Figure 2.14: **Electrocardiogram** The main waves and intervals in a heartbeat of an electrocardiogram, from [106]

1. **Heart Geometry** - The geometry of the heart depends on its size, shape and positioning. All these features highly influence the electrical current path during the heartbeat, the heart and muscle cells to be recruited and depolarised, and the average amount of time of a single heartbeat [107; 108];

2. **Physical Exercise and Meditation** - Athletes with high levels of physical training generally have larger hearts and thicker myocardium, indicating how physical exercise can influence the general heart and ECG characteristics. In addition, both exercise and meditation affect the heart rate, which translates into variations especially visible on the interval between the QRS complex and the T wave [109];

3. **Individual Characteristics** - Overall features like age, weight, height and pregnancy can affect the heart orientation and position. These shifts change the orientation of the electrical current conduction vectors, which alters the ECG waveforms [110];

4. **Cardiac Conditions** - Heart diseases or other medical conditions are also visible in the ECG waveform, since they may interfere in the general electrical conduction dynamics [111];

5. **Position and Shape of the Organ** - The body positioning like standing or laying down lead to significant variations in the position and shape pf the organ. Altering the position of the heart in the thorax will also change its reference position with the electrode placement. In this way, different body positions result on different ECG signals collected [110];

6. **Emotions and Fatigue** - Physiological and psychological states highly influence the autonomous nervous system, which translates into variations in heart rate considering different states of fatigue or emotion, from excitement to calmness [23];

7. **Electrode characteristics and placement** - The ECG acquisition methods and general specifics like the type, size and number of electrodes, as well as their positions on the chest and limbs, have an influence on the quality of the signal collected. The mispositioning of electrodes is also a source of variability, since it can change the general overview and perspective of the electrocardiographic signal [108; 110].

All the factors presented result in visible differences in the ECG data collected. Regarding the first three mentioned (Heart Geometry, Physical Exercise and Meditation and Individual Characteristics), they highly contribute to inter-subject variability, which is potentially positive to the development of biometric systems based on the ECG. However, concerning emotion recognition, subject-dependent settings are one of the most difficult problems that affect the already developed systems, since this can result in different ECG signals and patterns corresponding to the same emotion felt by different subjects.

On the other hand, as mentioned, emotions are one of the factors that can also produce variability. However, this variability is mainly intra-subject and depends on the emotional state, which allows for the detection of different emotions through time. The nerve-endings of the Autonomic Nervous System (ANS) within the cardiac muscle have a major effect on the cardiac response, namely on the heart rate. The sympathetic system has fibers that run along the atria and the ventricles. When activated, they stimulate the myocardium to increase the heart rate. On the other hand, this rate and general cardiac workload are reduced by the parasympathetic system [23].

Concerning a stressful state, the sympathetic system overtakes the parasympathetic and different effects occur like the increase of the conduction rate, the dilation of coronary blood vessels, and the increase in perceptiveness to internal and external stimuli [112]. All these effects result in specific patterns and variations that can be detected in the ECG, thus, identifying emotional changes.

Nevertheless, the physiology and anatomy behind these biological processes also differ considering the subject, translating into distinct ECG patterns for the same emotion depending on the subject, which can be one reason for the subject-dependency problems found in a large scale of emotion recognition algorithms [23].

Furthermore, other factors may negatively influence emotional detection. For example, body posture during data acquisition may influence HR measurements, consisting of one of the most common features to detect arousal. Thus, all these variability aspects have to be considered since they can either ease or hamper the task at hands.

2.2.3 Acquisition and Experimental Setups

The ECG signal can be used for a variety of different applications, like medical diagnostic, biometric recognition or, in this case, emotion recognition. These purposes are highly distinct and imply completely different experiments and requirements. In this way, they differ in the most generally used techniques for ECG measurement and collection. In this section, the most common ECG

acquisition methods are presented, first for medical diagnostic purposes and finally concerning emotion recognition tasks.

2.2.3.1 Medical Diagnostic Acquisition Settings

For medical diagnostic purposes, there are some previously defined electrode configurations, that allow for standardized techniques to be applied, ensuring effectiveness and collection quality as well as comparable and reproducible results. Having this said, the two most common methods are the standard 12-lead configuration and the corrected orthogonal ECG configuration.

Standard 12-lead configuration

A 12-lead-electrocardiogram is the most standard lead system used in clinical practice, nowadays. A lead is an indication of the electrical activity of the heart from a given angle. This system uses a total of 12 leads to acquire the ECG signal, showing 12 perspectives of the heart's activity through two electrical planes – vertical and horizontal - by using a total of 10 electrodes [113; 114].

These 12 leads are generally divided into three categories: three bipolar limb leads, three monopolar limb leads and six monopolar precordial leads. The limb leads give information regarding the vertical plane whereas the precordial provide information in the horizontal plane.

The bipolar limb leads are part of Willem Einthoven's legacy, winner of the 1926 Nobel Prize for his advances in electrocardiography. Using an equilateral triangle (Einthoven's triangle) coordinate system (see Figure 2.15), the idea is to capture the projection of the cardiac dipole. The vertices are the left arm (LA) and the right arm (RA) writs and the left leg (LL), where each electrode is positioned. The right leg (RL) works as a point zero, where the electrical current is measured. Because of this, RL is considered as a grounding electrode that helps minimize ECG artifacts and it is not considered as a lead in the 12-lead system, since it doesn't come up in the ECG readings [114]. In this configuration, the signals are obtained by measuring the electric potentials between RA and LA (Lead I), RA and LL (Lead II) and between LA and LL (Lead III).

These same three electrodes are also used to capture the monopolar limb leads, also known as augmented limb leads, that correspond to the measurement of the cardiac dipole in the direction of each limb electrode. In this way, in the direction of LA, it is obtained the aVL lead, for RA it's obtained the aVR and finally the for LL we have the aVF lead [99].

The six monopolar precordial leads, or transverse leads (V1 to V6), are positioned of the chest and provide information about the heart's horizontal plane. Like the unipolar limb leads, the precordial are also unipolar and require only a positive electrode. The negative pole of these 6 leads is found at the center of the heart. Figure 2.16 presents the complete standard 12-lead configuration.

Corrected Orthogonal Configuration (Frank Leads)

The Frank electrode system captures the three-dimensional extent of the heart dipole, using three different channels and each one corresponding to an orthogonal direction: x, y, z. In this way,

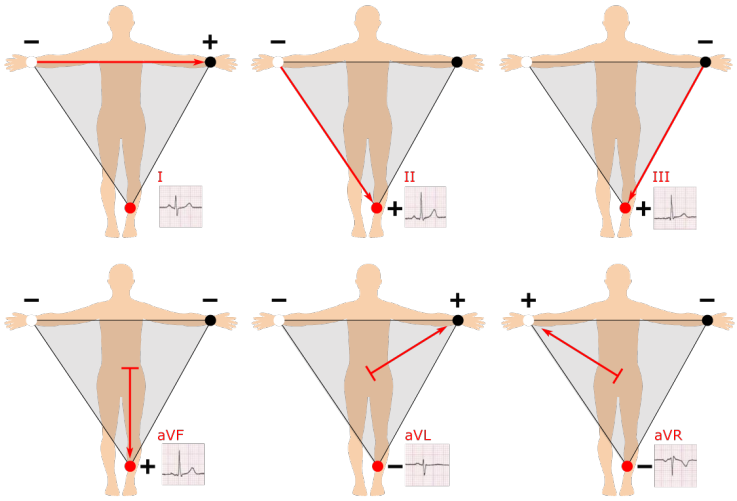


Figure 2.15: **Einthoven's triangle**: bipolar (I, II, III) and monopolar (aVR, aVL, aVF) limb leads, from [114].

potential differences are measured not only in the frontal plane, through the limb leads, but also along the antero-posterior axis of the body [115][116].

The Frank Leads system uses seven different electrodes: I, E, C, A, M, F, H. Five of them are located on the same axial plane (A, E, M, C). I and A are positioned below the right and left axillae, E and M are on the intersection of the sagittal and axial plane, being E positioned anterior and M posterior to the subject, and C is located halfway between A and E. Finally, H is on the back of the neck and F is on the left ankle [99]. Through the electrical circuit presented on Figure 2.17, three outputs are obtained: V_x, V_y and V_z, being proportional to the dipole components p_x, p_y and p_z, respectively [116].

2.2.3.2 Emotion Recognition Acquisition Settings

Regarding the ECG acquisition for emotion recognition experiments, only signals analogous to the bipolar limb leads are normally used, which indicates a lesser complex collection than the medical standard acquisitions. This lead choice can be because limb leads are generally more comfortable for the user, especially Lead I. In this way, they are the more relevant and generally applied in real emotion recognition applications. In [16], three sensors were attached to the participant, two electrodes were placed on the wrist and a reference in the ulna bone (in the arm), working as a ground electrode. This type of setup allows an accurate measurement of the heart rate (HR) and consequently of the heart rate variability (HRV), highly used and correlated with emotional states.

In [9], the ECG was recorded using a SHIMMER^{TM} [117], a wireless sensor able to obtain Lead II (RA to LL) and Lead III (LA to LL) vectors, although they only used Lead II for further development. Regarding [14], the ECG was recorded using a SHIMMER 2R, and three electrodes. Two of them were placed at the right and left arm and the third at the ankle, working as a reference. This set up provided a precise identification of HR and the full QRS complex.

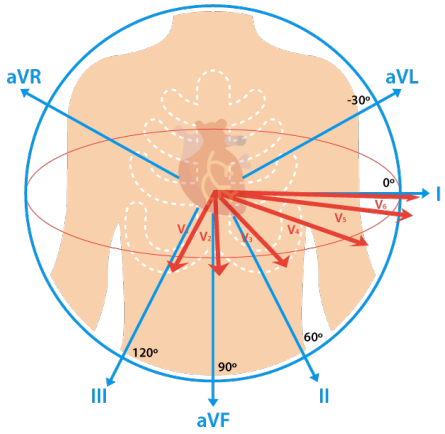


Figure 2.16: Standard 12-lead configuration, from [114]

In other cases, instead of placing the electrodes on the wrists or arms, some experiments also place them below the collar-bone [15; 18].

However, when in emotion recognition research, experimental setups also need to consider the emotion elicitation, since the ECG signal should reflect some emotion variation.

In emotion recognition, two types of induction experiments can be considered: **passive** or **active**. In a passive experiment, subjects are passively exposed to some emotion stimuli, while in active experiments, subjects participate actively during the experiment [11]. Concerning the emotion annotation, it can be a **self/explicit assessment** (internal annotation), in which the subject describes the emotion he/she is feeling, or an **implicit assessment** (external annotation), where the affective state is externally evaluated by close observation of the subject's behaviour or physiological signals. However, these two methods both show some uncertainty since it depends on the person observing or evaluating. In some research works, there's the combination of both techniques to ensure correct labelling of the data [14].

For emotion elicitation, passive experiments can use videos, pictures or even music records as external stimulus for emotion elicitation. One example widely used is The International Affective Picture System (IAPS), a picture set that provides rates of affect for a large set of emotional states (see Figure 2.18) [11; 92]. The volunteers are generally asked to sit conveniently while the stimuli are presented and the emotion assessment is performed during the stimuli or after.

Concerning the active experiments, the stimulus dynamically interacts with the volunteers. In [11], a commercial video game was used, increasing the difficulty and gradually absorbing the player in the game. After getting acquainted with it, the ECG starts to be measured as well as a facial expression recording. In the end, the volunteers should watch the video with their facial expressions while playing the game, and continuously report the emotions they felt at each time. Concerning this specific experiment, the subjects reported the level of arousal, as can be seen in Figure 2.19. However, these active experiments are rather new in the Emotion Recognition field, and most researchers opt to use passive experiments with pictures, videos or music as elicitation techniques.

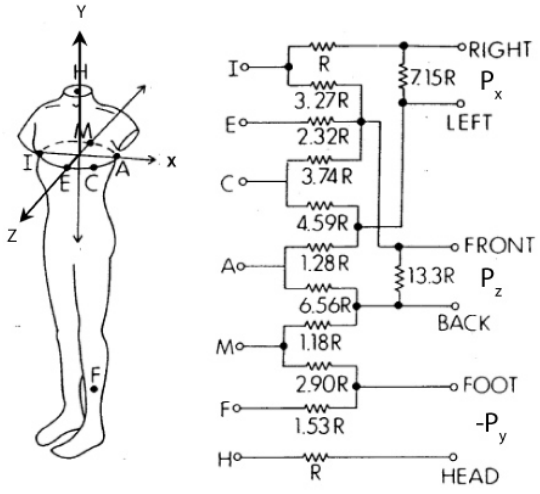


Figure 2.17: **Corrected Orthogonal Configuration: Leads and electrode placement**. Left: electrode placement; right: lead measurement computation), from [115].

2.3 Conclusion

The electrocardiogram is a cyclic signal that represents the electrical activity of the heart. As mentioned earlier, because the heart is directly connected to the ANS (parasympathetic and sympathetic nervous systems), it is possible to correlate emotional states with changes in the ECG signal.

Although the electrocardiogram is prone to high variability, which may have its origin in various factors and lead to false emotional interpretations, techniques should be applied to reduce these problems, especially regarding subject dependence.

In terms of acquisition methods, traditional emotion recognition experiments use simpler and less complex ECG measurement techniques than those used in the field of medical diagnostics. Wireless sensors are widely used, as well as a reduced number of electrodes, allowing a more comfortable and relaxed setup for the subjects, who can move more freely.

After presenting and analysing some available ECG databases in chapter 3, some of the most commonly used approaches and techniques in the literature for Emotion Recognition using ECG are presented in chapter 4.

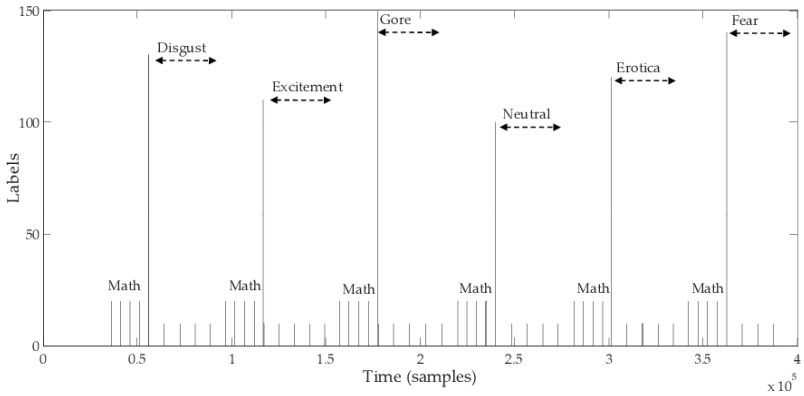


Figure 2.18: **Passive Experiment**. Every picture of the IAPS photo-set labelled, showing emotion variations through time, from [11].

Figure 2.19: **Active Experiment** Game and face expression recording, used for self-assessment of arousal, from [11].

Chapter 3

ECG Signal Databases

3.1 Introduction

A database is an organized collection of significant data that can be used for a given purpose. In all machine or deep learning problems, databases are an essential part of the system's development process, giving significant examples to train a given model correctly.

Having a well-structured signal collection is fundamental for the development of an accurate system, that may work in general real-life applications. In order to achieve a good database, Pinto *et al.* [99] identified a list of factors to take into account, like the *electrode number* and the *electrode positioning*, the *sampling frequency* and the *subjects health*.

The *electrode number* can have a direct impact on the system's performance as well as the *electrode positioning*. Zhang and Wei [118] concluded that the further away the electrodes were placed from the heart, the worse the results would be. *Sampling Frequency* proves to be important, as well, since lower sampling frequencies translate into a lower loss of information in the conversion of the analogical heart signal to digital.

Regarding the volunteers, a lot of factors have to be taken into consideration, like the *subjects health*, since heart conditions are one of the causes of ECG variability, and the *number of subjects*. A signal collection with the presence of a high number of subjects guarantees subject diversity as well as a high amount of data, both important factors concerning a machine or a deep learning problem. Regarding emotion recognition, a high number of subjects will allow to divide different subjects into training and testing sets tackling the subject-dependency issue, by having a test set of subjects "unseen" by the algorithm.

In emotion recognition tasks, the *label reliability* is also an important matter, since the algorithm's performance depends on how well it is trained. However, emotional states are sometimes difficult to evaluate in a quantitative way. Thus, as already explored in Chapter 2, Section 2.2.3.2, there are two main methods of labelling emotions and their combination is highly used to ensure a higher level of trust in the labelling process.

The *number of records* is also a parameter to consider in a well-structured database. Machine learning and especially deep learning algorithms need to have a high amount of data available.

With a low number of records or subjects, overfitting is prone to happen, since there is not a representative amount of data of each class to allow quality learning. Furthermore, small databases, with reduced data amount and signals' variety, can also lead to other problems such as signal and subject dependency. Finally, aspects like the *acquisition settings* are also relevant and should be in accordance with the purpose of the research, as already seen in Chapter 2, Subsection 2.2.3. In emotion recognition, the type of sensor and the electrodes used are important aspects. Wearable sensors and a low number of electrodes that guarantees a quality signal collection allow an easier and more comfortable data acquisition for the subjects involved.

3.2 Available databases

Concerning emotion recognition using physiological signals, there are some available databases (see Table 3.1) that are further presented and analysed. However, due to the difficulty of experimental setups and other limitations such as setups costs, the databases presented are not considerably large. However, emotion recognition is still a recent field, which indicates that, with time, more databases will become available for different investigations to be conducted.

AMIGOS

AMIGOS - A dataset for Multimodal research of affect, personality traits and mood on Individuals and GrOupS, was collected by Miranda *et al.* [14], with the purpose of creating a database for personality research based on neurological and physiological signals that studies participants in individual and group settings, exposing the differences between emotional reactions when alone or part of an audience.

The experimental scenarios involved two distinct settings. In the first one, 40 healthy volunteers watched 16 short emotional videos and, in the second, 37 participants watched 4 long videos, 17 alone and the other 20 in a group (5 groups of 4 people). This experiment allowed a significant number of data and the participation of a variety of volunteers. Simultaneously, five different modalities were recorded: audio, visual (body and face), EEG, ECG and GSR.

Regarding the emotion assessment, an internal annotation was performed, in which the volunteers self-assessed their own affective states in terms of arousal, valence, control, familiarity, liking and basic emotions, and an external annotation, by evaluating the emotion each subject was experiencing. The combination of labelling methods also ensures proper labelling for the AMIGOS database.

The ECG signal was recorded with the wearable sensor Shimmer $2R^5$ extended with an ECG module board, using a sampling frequency of 256Hz and a 12-bit resolution. Three electrodes were used, two of them on the right and left arm and the third in the internal part of the left ankle.

ASCERTAIN

ASCERTAIN - A multimodal databASe for impliCit pERsonaliTy and Affect recognition [14], is a database that intends to connect personality traits and emotional states through physiological responses. In this way, ASCERTAIN contains five personality scales (Extraversion, Neuroticism, Agreeableness, Conscientiousness and Openness) and emotional self-ratings (Arousal, Valence, Engagement, Liking and Familiarity).

58 healthy volunteers watched 18 affective videos while their EEG, ECG, GSR and facial activity were recorded using off-the-shelf sensors, during a 90-minute session. Thus, this database contains a high number of healthy subjects and long record sessions, making ASCERTAIN useful for machine and deep learning applications. After viewing each clip, the subjects made a self-assessment of their emotional state using the affective ratings, evaluating, using only a type of data labelling.

For the ECG measurement, two electrodes were placed at each arm crook, and the reference and third electrode was positioned at the left foot.

DECAF

DECAF - A multimodal database for DECoding user physiological responses to AFFective multimedia content was recorded by [17].

Thirty volunteers listen to 40 one-minute music records and watched 36 movie clips and after each session, the participants did their self-assessment ratings, for arousal, valence and dominance, from 1 (low) to 9 (high). In addition, an implicit assessment was also developed, allowing a thorough data labelling. Although it doesn't present a number of subjects as high as other databases, it contains two distinct experimental sessions (music and movie clips), which increases the number of data available.

The modalities considered were the near-infra-red (NIR) facial videos, horizontal Electrooculogram (hEOG), ECG, trapezius-Electromyogram (tEMG), EEG and magnetoencephalogram (MEG), which also acquires brain signals. In this way, DECAF enables comparisons between the EEG and MEG modalities, as well as movie and music stimuli for emotion recognition.

The ECG signal was recorded at a sampling rate of 1KHz, subsequently downsampled to 256Hz. The set up was simple by placing three sensors in the subject's body, two electrodes on the wrist and the reference on a bone part of the arm (ulna bone).

DREAMER

DREAMER is a multimodal database that consists of EEG and ECG signals recorded during emotion elicitation experiments [9].

This database has records from 23 participants while they were presented with audio-visual stimuli, consisting of 18 videos. In this way, in terms of subject and records number, it is a small database that can show some limitations.

After each stimulus, subjects self-assessed their affective state in terms of arousal, valence and dominance, by evaluating them from 1 (low) to 5 (high), only considering an explicit assessment.

Both EEG and ECG were recorded with portable wearable and wireless equipment, indicating the possibility of applying affective computing methods in everyday technology. ECG was recorded at a 256 Hz sampling rate, using a SHIMMER^{TM} sensor, that is able to produce Lead II and Lead III vectors.

WESAD

WESAD - a database for WEarable Stress and Affect Detection is a multimodal database, using both physiological signals and motion data, recorded from a device in the wrist and chest, which allowed performing measurements on the subjects' daily basis. The goal of the database was to elicit three affective states - neutral, stress and amusement.

A small set of subjects, 15 in total, experienced three different scenarios: sitting while reading a magazine, watching 11 funny videos and going through a Trier Social Stress Test (TSST), corresponding to the neutral, amusement and emotional stress elicitation, respectively. Simultaneously, ECG, face, body, heart rate variability and skin conductance were measured.

Considering the small number of subjects and the not so common emotional states tackled, this database use is a little bit more limited, since most research groups use the traditional affective dimensional parameters, like arousal, valence, dominance, among other.

For emotion assessment, questionnaires were developed and presented to each subject, in order to obtain a detailed description of the emotion felt. In this way, although an implicit assessment wasn't also conducted, the self-assessment developed was more detailed and thorough.

For data collection, as already mentioned, chest and wrist-worn devices were used: *RespiBAN Professional* and *Empatica*, respectively. The ECG signal was recorded at a sampling rate of 700 Hz, using three standard points.

SWELL-KW

The SWELL Knowledge Work (SWELL-KW) is a dataset developed for stress and user modelling research [18].

Twenty-five volunteers performed typical knowledge work, like writing reports, reading e-mails, searching for information about a given topic or making an oral presentation. Meanwhile, the working environment was altered with different stressors such as email interruptions, generally increasing the work pressure.

In this way, three states were considered: neutral (working with no external alterations and no limitations in terms of time), the stressor time pressure (reducing the time deadline to 30 minutes) and the stressor interruptions (sending a variety of eight emails, some important and others irrelevant).

Concerning the subject ratings, different questionnaires were applied to evaluate mental effort and demand, physical and temporal demand and the level of stress. Considering emotion

evaluation, the subjects reported their affective states in terms of valence, arousal and dominance, evaluating them from 1 (low) to 9 (high). This kind of labelling is complete, offering evaluation regarding different parameters like stress and demand. However, concerning emotions, there's some uncertainty since only the subjects' own evaluations were considered.

During the experiment, facial expressions, body posture, ECG and skin conductance were measure. The ECG measurement was done at a sampling frequency of 2048HZ and using a Mobi device (TSMI) with self-adhesive electrodes, positioned across the heart. One of them was placed bellow the right collar-bone and the other bellow the chest, with the grounding electrode below the left collar-bone. SWELL provides both raw and preprocessed ECG data, calculating the heart rate and heart rate variability, which can be useful in some researches where this type of processed data is needed.

MAHNOB-HCI

MAHNOB-HCI is a multimodal database for affect recognition and implicit tagging [15]. 30 healthy volunteers participated in the development of this database, however 3 of them didn't conduct all the experiments. In this way, 27 subjects from both genders and different cultural origins, participated in two distinct experiments.

The first one consisted of watching 20 emotional videos while self-reporting their own emotions, using arousal, valence, dominance and predictability as emotion parameters, and emotional keywords (discrete emotions: disgust, amusement, joy, fear, sadness and neutral). On the other hand, the second experiment used short videos and images without any tag and, after, with correct or incorrect tags. The participants were instructed to analyse each tagging and agree or disagree with it.

In this way, self-assessment is used regarding the first experiment. In the second, the subjects also perform an evaluation on a given image or video tagging. Although the first videos are evaluated through an internal assessment, the second experiment works as an implicit/external annotation. The subjects evaluate the label given to each image and understand possible variations from their opinion to the real image/video annotation. Thus, this database can also be useful to understand how subjects opinion and emotional states may differ and how reliable their assessment is.

Regarding the number of recordings, MAHNOB presents a wide range of data available, even with only 27 volunteers, since each one is presented with 20 emotional videos, 28 images and 14 short-videos.

During each experiment, EEG, eye gaze, audio, visual (face and body), and different physiological signals (ECG, GSR, skin temperature, RA) were collected. The ECG was acquired using three sensors attached to the subjects' body. Two of the electrodes were placed on the right and left corners of the chest, below the clavicle bone and the third on the abdomen, below the last rib. This type of setup provides a precise identification of the heartbeats and it is also simple and easy for the subjects to wear.

3.3 Conclusion

In this section, various and important factors for selecting a good and well-structured database have been analysed. Aspects such as the number and positioning of electrodes, the reliability of labels and the number of data and subjects are important as they determine how well a particular database is suited for a specific research purpose.

During this dissertation, different databases are used, namely DREAMER, AMIGOS and MAHNOB-HCI, since all of them respect the basic factors that were analysed, such as a wide range of available data and correct acquisition settings. They also provide different experiments with relevant information from which different hypotheses can be tested and conclusions can be drawn. For example, concerning AMIGOS, as mentioned, there is access to both external and internal annotations, which can be interesting for further experiments. Moreover, both AMIGOS and MAHNOB-HCI use two types of emotion assessment, discrete (disgust, amusement, joy, fear, sadness and neutral) and dimensional models (arousal, valence, dominance and predictability). All of these aforementioned features make them complete databases suitable for a variety of research purposes. However, as already referred, none of these databases is extremely large, specially DREAMER, since it presents a reduced number of subjects and experiments per subjects. This fact can turn out to be problematic in some situations and lead to the limitations above mentioned. Nonetheless, these databases are used in most deep learning approaches found in the literature, and using the same databases is an important factor when trying to reproduce a given methodology.

Table 3.1: Summary of the most commonly used ECG Signal databases for Emotion Recognition (N.S. - number of subjects; N.E. - number of electrodes).

Database	Author/year	N.S.	Sampling Rate (Hz)	N.E.	Electrode Placement	Stimuli	Emotion Assessment
AMIGOS	J. A. Miranda-Correa *et al.* [14] Queen Mary University of London, University of Trento, Italy 2017	40	256	3	Arms Left Ankle	16 short emotional videos 4 long videos	**Self/Explicit Assessment** (arousal, valence, control, familiarity, liking, emotional keywords) **Implicit Assessment** (arousal and valence)
ASCERTAIN	Ramanathan Subramanian *et al.* [119] 2016	58	–	3	Arms Left Foot	18 affective videos 90-minute sessions	**Self/Explicit Assessment** (arousal, valence, engagement, liking, familiarity)
DECAF	Abadi *et al.* [16] University of Trento, FBK, ADSC, Semantics Inovation Lab, Telecom Italia 2015	30	1000 downsampled: 256Hz	3	Wrists Arm (boney part)	40 1-minute music records 36 movie clips	**Self/Explicit Assessment** (arousal, valence) **Implicit Assessment** (arousal, valence)
DREAMER	Stamos Katsigiannis and Naeem Ramzan [9] University of the West of Scotland 2015	23	256	3	Lead I and Lead II vectors	18 affective videos	**Self/Explicit Assessment** (arousal, valence,dominance)
WESAD	Schmidt *et al.* [120] Corporate Research, University of Siegen, Germany 2018	15	700	3	Chest	neutral amusement stress: TSST	**Self/Explicit Assessment** (questionnaries)
SWELL-KW	S. Koldijk *et al.* [18] Radbound University, Delft University of Technology 2014	25	2048	3	Chest	neutral stressor time pressure stressor interruptions	**Self/Explicit Assessment** (arousal, valence,dominance)
MAHNOB-HCI	Soleymani *et al.* [15] Imperial Colllege London, Intelligent Behavior Understanding Group 2012	27	1024 downsampled: 256	3	Chest	20 emotional videos Image Tagging	**Self/Explicit Assessment** (arousal, valence,dominance, predictability, emotional keywords) **Implicit Assessment**

Chapter 4

Prior Art

4.1 Introduction

For more than a century, since the appearance of the first electrocardiogram, this electrical signal has been developed and constantly improved to obtain more reliable results and measurements. Nowadays, the main application of ECG remains in the field of medical diagnosis.

On the other hand, Emotion Recognition to human-computer interaction first gained attention in the 1980s. Although most research in this area uses facial expressions [121], emotions are known to strongly influence the Autonomous Nervous System (ANS) activity, which is responsible for regulating a variety of body parameters. In this way, physiological signals began to be considered as possible indicators of emotional fluctuations.

Thus, the use of ECG for emotion detection is still quite recent compared to other modalities such as face or voice. However, as a physiological signal, ECG patterns can translate changes in emotion, mainly through heart rate and heart rate variability [32].

The most common methods using ECG for emotion recognition have well-established steps: (1) ECG data collection, (2) Preprocessing, (3) Feature Extraction and (4) Classification. However, as already approached in Chapter 2, Section 2.1.2.4, in the literature, the methods can be divided into two main groups, considering feature extraction. On the one hand, the researcher manually engineers *handcrafted features*, computing and extracting them, before feeding them to the classifier. On the other hand, *deep learned features* are automatically obtained from a given algorithm.

In this Chapter, the most common handcrafted and non-handcrafted methods are described and analysed, considering their advantages and limitations to Emotion Recognition tasks. Before this, a brief overview of preprocessing methods is offered in Section 4.2.

4.2 Preprocessing

The ECG is commonly contaminated with artifacts that may difficult its use and application to different tasks. The powerline interference (50-60Hz), the electrode contact noise, surrounding mus-

cles contraction, external electronic devices interference and motion artifacts are some examples of unwanted noise. In this way, in both *handcrafted* and *non-handcrafted* features, pre-processing can be applied to clear the ECG signal.

According to [122], the bandpass able to maximise the QRS energy is between 5-15 Hz. In this way, the signal is generally filtered to remove the effects of noise and baseline wander. [123] used a 3^{rd} order Butterworth filter with 0.002 Hz and 100Hz cut-off frequencies, whereas in [124] the signal crosstalk removal was performed by applying a notch filter. Considering [125], a high pass filter was adopted to eliminate baseline wander, with a 4Hz cut-off frequency. Other possible alternatives to filtering are transforms that may also cancel the ECG noise. The Discrete Wavelet Transform (DWT) decomposes the signal into different sets, capturing both frequency and location information was used in [126].

In [127], three different denoising techniques were tried and compared: Independent component analysis (ICA), Principal Component Analysis (PCA) and Multiscale Principal Component Analysis (MSPCA). PCA is a statistical linear technique that uses an orthogonal transformation to obtain a dataset with no correlated variables, only keeping the principal components (PCs). On the other hand, ICA is a non-linear technique that can be considered an extension of PCA, and finds linear transformations of multivariate random signals into additive and independent subcomponents. Regarding MSPCA, it extracts relations between different variables and decorrelates them in case they are related.

According to the type of signal, the acquisition settings and the characteristics of both time and frequency domain, different preprocessing methods can be applied (filtering, DWT, ICA, among others). Especifically for filters, different low pass filters like elliptic, adaptive and Butterworth are also highly applied to ECG signals. In addition, some artifact components can be removed through visual detection, requiring more expertise and knowledge [32]. However, it is important to be careful when preprocessing ECG signals, in order to retain the useful information for the task at hand.

4.3 Handcrafted Methods

Feature extraction is able to transform the ECG signal into a list of important and relevant features for the following classification. These features can be further categorized according to the domain they are extracted as frequency or time [128]. Furthermore, features can also be distinguished as linear or non-linear, fiducial or non-fiducial. Statistical features have into account measurements like the mean, variance and other computations that can consider the whole signal or be only computed between fiducial points. Fiducial-based approaches use specific points such as P Q R S or T, measuring their interval, amplitude, time, among others. On the other hand, non-fiducial approaches consider the signal as a whole, or segments of it, in order to extract features related to the waveform morphology [99]. Fiducial features are generally computed in the time domain; however, there are some examples of frequency-fiducial approaches. On the other hand, non-fiducial are widely computed in both the time and frequency domain.

In this section, approaches are divided concerning the extracted features domain (time or frequency) and further analysed the fiducial and non-fiducial approaches in each. In Table 4.3.3, a literature review is presented and the following subsections analyse and describe some of the researches taken into account.

After evaluating the different features that can be considered, a description and analysis of the most common machine learning classifiers are also provided.

4.3.1 Time-domain Analysis

As already mentioned, heart rate and heart rate variability are the main changes visible on the ECG signal associated with changes in emotion. Therefore, time-domain features are usually associated with heart rate variability (HRV), using different methods to detect and extract information from it. Some examples of this kind of features are the heart rate [123], R-R intervals and R peak value [129], the standard deviation of NN intervals (SDNN) [95], root mean square for standard deviation (RMSSD) [12], among others.

4.3.1.1 Fiducial Features

Fiducial Features are associated with specific markers and points of a given signal. Regarding the ECG, there are different fiducial-based approaches in order to extract emotion information, by using the features mentioned above, besides considering other relevant points such as the P and T waves.

When the cardiac rhythm is constant, the heart rate can be determined by the interval between two successive QRS complexes. Thus, different approaches choose to detect the QRS to acquire time domain HRV information, like the R peaks and R-R intervals [129–133] . The R wave can be easily detected, since it presents the higher amplitude in the signal [134]. To detect the R peaks, different algorithms can be applied, like the Pan and Tompkin's algorithm [122]. Hamiton's [135], Slope Sum Function [136] and Christov [137]. However, although the QRS complex defines the activation of the heart to human emotional states, there are some difficulties in emotion recognition since this indicator is common to a wide range of emotions. Nevertheless, it still represents the most used time-domain, fiducial features for emotion recognition purposes.

In [138], three methods were applied to analyse the HRV, including time and frequency domain analysis, as well as statistical methods. Concerning the time domain, features like the mean, STD and the coefficient of variation of the RR intervals were considered. In the same way, Sriramprakash *et al.* [139] used a total of 13 features related to HR and HRV, being 10 of them in the time-domain. Regarding HR, the mean, median, mean absolute deviation (MAD HR) and the standard deviation of Heart Rate (STD HR) were computed, whereas the HRV statistical features considered were the RMSSD, the average of NN intervals (AVNN), the NN50, among others, presented in Table 4.3.3.

In [140], the R-R intervals are calculated and further filtered with a median filter, that removes the R-R intervals out of a given range. The intervals are further used to calculate time-domain

HRV features, such as meanHR, meanRR, SDNN, SDSD, RMSSD, pNN20, pNN50, etc. In addition, [141] decomposed the ECG signal and detected the P-QRS-T wave. After identifying R peaks, Q and S are the left and right minimum of R, respectively. The P and T waves were located by looking for peaks between consecutive QRS complexes. With the accurate detection of the P-QRS-T wave detected, different fiducial features were extracted, such as the R wave mean, median, maximum, range and amplitude, and analogous characteristics concerning the QS interval, as well as the median amplitude value of P wave.

Finally, in [142], different fiducial features were considered and extracted, like the PQ, QS and ST intervals, the R-mean, median, standard deviation (STD), minimum and maximum value and the same features concerning the PQ interval.

Although the majority of studies focus on the analysis of the QRS amplitudes, R-R intervals and the duration between consecutive QRS complexes, some research also pays attention to QT/QTc dispersion, and [143] provides information that allows correlating this interval with levels of anxiety and anger.

More methods using time-domain and fiducial features are presented in Table 4.3.3.

4.3.1.2 Non-Fiducial Features

Non-fiducial features consider the totality of a signal or segments of it. By not being limited to specific fiducial points, this kind of approach offers various methods and can be applied in both time and frequency domain. In the time domain, some examples are the computation of statistical features from ECG segments or the totality of the signal such as the mean, median or variance. Furthermore, specific methods like EMD and BEMD also allow obtaining time-domain features.

Considering the method applied in [24], both local binary pattern (LBP) and local ternary pattern (LTP) in the time-domain were used to extract directions and patterns from the ECG signal. Although these are techniques most commonly applied to images and emotion recognition from facial expressions recognition, it was adapted to ECG, by dividing the signal into different frames and detecting possible patterns and information conveyed in each segment, with high accuracy.

Agrafioti *et al.* [23] approach consisted of extracting information about oscillation activity of a signal. For this, it was used a feature extraction based on Empirical Mode Decomposition (EMD). EMD performs a signal decomposition, obtaining Intrinsic Mode Functions (IMFs) and the oscillatory activity conveyed in these time-domain signals is extracted, by applying a local oscillation computation. Furthermore, in this research the oscillatory activity was also extracted in the frequency domain, by applying a different technique, further described in Subsection 4.3.2.

Considering [144], both time and frequency domain features were extracted. Taking into account the non-fiducial time-domain features, statistical measurements were computed, such as the mean, standard deviation, first and second difference and the maximum and minimum of the ECG signal. In addition, [145] also extracted time-domain features, such as amplitude, frequency, linearity and variability.

4.3.2 Frequency-Domain Analysis

In order to extract features from the frequency domain, signals need to be transformed from their natural time-domain to this one. Although some features in the frequency domain can be extracted by taking into account some specific points in the ECG signal, the most common features in this domain are non-fiducial. Some of the most used techniques are the Fast Fourier Transforms (FFT), Continuous Wavelet Transform (CWT), Discrete Wavelet Transform (DWT), the Hilbert transform and the Power Spectral Density (PSD) [132; 146].

In [147], a total of 13 frequency-domain HRV related features were calculated. The power spectral density (PSD) was computed by applying a fast Fourier transform (FFT) to the R-R intervals. The PSD analysis allowed to compute the power of specific frequency ranges (very-low-frequency range (VLF), low-frequency range (LF) and high-frequency range (HF)) and peak frequencies. These features are highly used, as well as the LF/HF ratio, being present in a variety or researches [12; 138; 142; 148–150]. In [151], the features extracted are Fourier coefficients, obtained by applying a FFT to the ECG signal.

On the other hand, in [123], Discrete wavelet transform (DWT) was applied to extract features by using four wavelet functions: Daubechies6, Daubechies7, Symmlet8 and Coiflet5. [26] also applies DWT, namely db4 and coiflet5, and Discrete Cosine Transform for feature extraction. Continuous wavelet transforms may also be used [152].

Concerning the already discussed method of [23], oscillation activity was also calculated by applying a Hilbert Transform after the EMD. Regarding [153], EMD in combination with Hilbert transform was also used, as well as EMD combined with Discrete Fourier Transform (DFT), comparing the results when using these different feature extraction methods.

4.3.3 Non-Linear Analysis

Besides the feature extraction techniques already mentioned and analysed, concerning both frequency and time domain, there are other possible methods for extracting features.

The Poincaré Plot of RR intervals measures the quantitative beat-to-beat correlation between adjacent RR intervals and it is a metric used to evaluate HRV. The point cloud obtained in this plot is generally characterized by its length through the line of identity (SD_2), which represents the standard deviation of the continuous long-term beat-to-beat RR interval variability, and its breadth along this line (SD_1), indicating the standard deviation of the instantaneous beat-to-beat RR internal variability. The ratio between SD_1 and SD_2 is also a valuable measure and these parameters can be computed by the following formulas [138]:

$$SD_1^2 = \frac{1}{2} SDSD^2 \tag{4.1}$$

$$SD_2^2 = 2SRR^2 - \frac{1}{2} SDSD^2 \tag{4.2}$$

$$SD_{12} = \frac{SD_1}{SD_2} \tag{4.3}$$

These features are extracted in different approaches [132; 138; 140; 147] In the latter, [147], there are also autocorrelation-related features, namely the maximum autocorrelation coefficient (*ACFcoef*) and the reciprocal of the lag time (*ACFfreq*) autocorrelation. Furthermore, two ECG-derived respiration (EDR) parameters are also computed: respiratory rate (*RSPrate*) and Coherence between final EDR and RR intervals (*Coherence*).

Concerning [150], SD_1 and SD_2 were also nonlinear features extracted. However, other methods were applied in order to obtain other features such as the Sample Entropy (*SampEn*) and the Approximate Entropy (*ApEn*), which consist of two entropy measures of the HRV. *ApEn* detects the changes in underlying behaviour that was not translated and reflected in the peak appearance and amplitude. On the other hand, the *SampEn* provides a good evaluation of the signal regularity and dynamics. DFA correlations, another nonlinear method, considers two, short and long term, fluctuations in $\alpha 1$ and $\alpha 2$ features, respectively. Finally, the Correlation Dimension measures the complexity of the time series, as well as its strangeness, by the *D2* feature.

Concerning statistical methods, the Kurtosis coefficient and the Skewness value are also used [129; 138; 154] and enable the evaluation of distribution probabilities of HRV. The Kurtosis allows evaluating the shapes of the probability, whereas the Skewness provides information regarding the probability distribution. Entropy is also a statistical parameter that can characterize the randomness of the dataset.

$$Kurto(X) = \frac{\sum(X-\mu)^4}{\sigma^4} \tag{4.4}$$

$$Skew(X) = \frac{\sum(X-\mu)^3}{\sigma^3} \tag{4.5}$$

$$Entropy(X) = -sum(px\log_2(p)) \tag{4.6}$$

Selvaraj *et al.* [155] considered a skewness-based Hurst and kurtosis-based Hurst features. The Hurst parameter analyses the smoothness of a time series, by having into account self-similarity and correlation properties, as well as its long-range dependence [155]. This parameter can be obtained by the already mentioned methods, such as the EMD and the WT; however, this research uses the already mentioned High Order Statistics (HOS) features, extracted from the QRS complex using two methods: Rescaled Range Statistics (RRS) and Finite Variance Scaling (FVS).

Table 4.1: Summary of the feature extraction and features selection methods used on the surveyed approaches

			Feature Extraction			
Paper	**Year**	**Database**	**Time-domain**	**Frequency-domain**	**Non-Linear**	**Feature Selection**
Hjortskov *et al.* [148]	2004	Private	–	LF, HF, LF/HF	–	–
Kim *et al.* [91]	2004	Private	menaHRV, SD HRV	LF, HF	–	–
Kappeler-Setz [133]	2007	Private	RR-peak, RR-interval, HR, SDNN, RMSSD, pNN50, triangular index	LF, HF, LF/HF,	Wilcoxon signed rank test Krushkal-Walls test / –	
Kim and Andre [95]	2008	Private	HRV, SDNN, NN50, PNN50	PSD, LF, HL,LF/HF	SD_1 , SD_2, SD_{12}	pLDA
Schubert *et al.* [156]	2009	Private	HR, HRV, RSA	HF, LF, LF/HF	D2	–
Xu and Liu [141]	2009	Private	P-QRS-T features (meanR, max R, STD QS...)	–	–	Hybrid Particle Swarm Optimization (HPSO)
Boonnithi and Phongsuphap [149]	2011	Private	mRR, mHR, SDRR, CVRR, RMSSD, pRR20, pRR50	VLF, LF, HF, nVLF, nLF, nHF, dLFHF, SMI, VMI, SVI	–	–
Guo [52]	2011	Private (Augsburg)	-	WT, eigenvectors (max, STD wavelet coefficients)	–	–
Taelman *et al.* [152]	2011	Private	–	LF, HF	–	–
Bong *et al.* [131]	2012	Private	HR, MRAmp, mRRI	–	–	–
Agrafioti *et al.* [23]	2012	Private	Local Oscillation (EMD, BEMD)	HHT	–	SFFS, FP
Murugappan *et al.* [123]	2013	Private	–	LH (STD, power), HF (STD, power), HF/LF, LF+HF	–	–
Selvaraj *et al.* [155]	2013	Private	–	–	Hurst, $\text{Hurst}_{skewness}$, $\text{Hurst}_{kurtosis}$	–
Xun and Zheng [142]	2013	Private (Augsburg)	PQRST waves, PQ, QS, ST (max, min, mean, median, STD, range) SDNN, pNN50	PSD, HF	–	ANOVA, SFS, SBS

Table 4.1: Summary of the feature extraction and features selection methods used on the surveyed approaches

			Feature Extraction			
Paper	**Year**	**Database**	**Time-domain**	**Frequency-domain**	**Non-Linear**	**Feature Selection**
Mikuckas *et al*. [132]	2014	Private	HR, SDNN, SDANN, RMSSD, $SDNN_{index}$, SDSD, NN50, pNN50%	Total power, VLF, LF, LF n.u, HF, HF n.u, LF/HF	SD_1, SD_2	Manual
Valenza *et al*. [92]	2014	Private	RRmean, RRSTD	instantaneous spectrum and bispectrum, HF, LF, LF/HF	dominant Lyapunov exponent	–
Walter *et al*. [145]	2014	Private	meanRR, RMSSD, slopeRR	–	–	Forward selection Backward selection
M *et al*. [153]	2014	Private	EMD	DFT, HHT	-	-
Godin *et al*. [12]	2015	DEAP MAHNHOB-HCI	avHR, avRR, SDT HR, RMSSD, SDSD	VLF, LF, HF	–	Correlation, Fisher score, Bayes classification
Ménard *et al*. [151]	2015	Private	–	fourier coefficients	–	–
Ferdinando *et al*. [30]	2016	MAHNOB-HCI	BEMD	dominant features	–	forward-floating search
Ping Gong [144]	2016	Private (Augsburg)	mean, median, STD, 1^{st} dif, 2^{nd} dif, max, min, max ratio, min ratio	HF, LF, LF/HF	Lyapunov exponent, ApEn, SampEn, Correlation Dimension, Complexity	C4.5 decision tree
Guo *et al*. [138]	2016	Private	meanRR, CVRR, SDRR, SDSD	LF, HF, LF/HF	kurtosis, skewness, SD_1, SD_2, SD_{12}	PCA
Tivatansakul and Ohkura [24]	2016	Private (Augsburg)	LBP, LTP	-	-	—
Ferdinando *et al*. [154]	2017	MAHNOB-HCI	BEMD	Spectrogram analysis, dominant features	kurtosis, skweness	LDA, NCA MCLM
Gjoreski *et al*. [140]	2017	MAHNOB-HCI , DEAP, DECAF, ASCERTAIN Cognitive Load Driving Worload	meanHR, meanRR, SDNN, SDSD, RMSSD, pNN20, pNN50	LF, HF, LF/HF	SD_1, SD_2, SD_{12}	—

Table 4.1: Summary of the feature extraction and features selection methods used on the surveyed approaches

			Feature Extraction			
Paper	**Year**	**Database**	**Time-domain**	**Frequency-domain**	**Non-Linear**	**Feature Selection**
Goshvarpour *et al.* [26]	2017	Private	–	DCT, db4, coiflet5, DWT, MP coefficients	–	PCA, LDA, Kernel PCA
Hsu *et al.* [147]	Private	SDNN, RMSSD, NN50, pNN50, SDSD, HR	VLF, LF, HL, LF/HL (and others based)	EDR, SD_1, SD_2, SD_{12}, ACF_{coef}, ACF_{freq}	SFFS-KBCS-based, GDA	
Minhad *et al.* [129]	2017	Private	R-peak and RR interval measurements, HR	–	kurtosis, skewness	Manual
Sriramprakash *et al.* [139]	2017	SWELL-KW	HR (mean, median, MAD, STD), HRV (RMSSD, AVNN, SDANN, SDNN, NN50, pNN50	LF, HF, LF/HF	-	Manual
Marín-Morales *et al.* [150]	2017	Private	meanRR, STQ RR, RMSSD, pNN50, RR triangular index, TINN	VLF, LF, HF (peak, power), Total Power	SD_1, SD_2, ApEn, SampEn, DFA $\alpha 1$, $\alpha 2$, D2	PCA
Wei *et al.* [146]	2018	MAHNOB-HCI	mean, standard of amplitude of P,R,T, HRV, RA	PSD (max, mean,STD) of HRV, RA	–	–
Xiefeng *et al.* [157]	2019	Private	HRV	–	SD_{12}	GA

4.3.4 Dimensionality Reduction and Feature Selection

Dimensionality Reduction and Feature Selection are two methods used for reducing the number of features extracted, since too many features result in high computational costs, the risk of curse of dimensionality and low classification accuracies [26]. However, these two methods are slightly different, since *feature selection* consists of simply selecting and excluding some features, without changing them. On the other hand, *dimensionality reduction* transforms features into a lower dimension.

Some of the most common and traditional techniques are PCA and LDA. PCA constructs a linear transformation by considering the principal components (principal eigenvectors) of the data, without any loss of information [158]. In this way, only some Eigenvalues are significantly high and the others are considered very small, not contributing to the data variations and, thus, the latter are the features to be removed. This technique is applied in a variety of researches involving emotion recognition [26; 52; 150].

On the other hand, LDA is a supervised technique that looks to find a linear mapping M of the data that can maximize the linear class differentiation in a low-dimensional space [159]. This technique is also highly present in the literature [26; 154]. In addition to these two methods, others can be applied, such as the Kernel PCA (K-PCA), a reformulation of the traditional PCA, in which a kernel function is used, resulting in a nonlinear mapping [160]. [26] used all three techniques, PCA, LDA and K-PCA, comparing their effectiveness and obtained the best recognition rate when using PCA. In addition, PCA also showed to be computationally less expensive when compared to other methods, such as K-PCA.

In [138], PCA also proved to be a good technique for feature selection. When using 13 HRV features to classify two and five emotion states, an accuracy of 70.4% and 52% was achieved, respectively. However, with five features selected, these accuracy levels increased to 71.4% and 56.9%.

Ferdinando *et al.* [154] developed research to explore different supervised dimensionality reduction. For this, three different techniques were considered, namely LDA, NCA (Neighbourhood Component Analysis) and MCML (Maximally Collapsing Metric Learning). NCA is a non-parametric method with the final goal of finding a transformation matrix that will determine the dimension of the transformed features. Considering MCML, a simple geometric hypothesis is used, by considering that all points belonging to the same class are mapped to a single location in the feature space, whereas all the other points are mapped to other locations [154; 161]. Considering this research, the results showed that the supervised Dimensionality Reduction based on NCA increased the accuracy from 55.8% to 64.1% and 59.7% to 66.1 % for a 3-class problem in valence and arousal, respectively. In this way, NCA showed to be superior when compared with the other two methods.

Concerning [23], two techniques were combined, the SFFS (Sequential Floating Forward Search) algorithm and the FP (Fisher Projection), while in [144] a C.45 decision tree was used for feature selection and the results were compared with not using a feature selection method. As

a result, the small set of selected features lead to more precise and accurate detection than a larger feature set.

Furthermore, [12] made use of 3 feature selection methods: correlation, fisher score and Bayes classification. The first evaluates the correlation between features, which indicates if they are linear dependent, and provides information regarding the features to remove. Fisher Score evaluates the ratio between inter-class and intra-class variability. The features to be chosen are the ones with higher inter-class and lower intra-class variability, i.e, the ones with a higher score. Finally, in Bayes Classification the features are selected using a selective forward search with Bayes classifier, by combining features and evaluating which set results in the highest classification accuracy.

Finally, in [162], different sets of features and classifiers are used for emotion recognition, presenting, for each case, the results obtained with and without feature selection. For all of the cases taken into account, feature selection improved the results obtained, showing to be a useful technique that may ensure more accurate emotion recognition systems.

4.3.5 Decision Methods

After the features are extracted and further selected, a decision method or classifier is needed, in order to do the proper recognition. As already presented in Section 2.1.2.4, Figure 2.12, some of the most common classifiers for emotion recognition based on physiological signals are SVM, KNN, LDA and RF. Considering specifically the ECG, SVM and kNN seem to be the two most present methods of decision [123; 129; 146; 154].

k-Nearest Neighbours (k-NN) is a non-parametric algorithm, which means that it avoids assumptions about the shape of the class boundary and thus it easily adapts to nonlinear boundaries [163]. This algorithm estimates how likely a feature vector is to be a member of a class or other, depending on its distance. Thus, this method attributes a feature vector to the most verified class among the k closest stored templates.

On the other hand, Support Vector Machines (SVM) are non-probabilistic binary linear classifiers. Given a labelled training set, SVM computes an optimal classification boundary that maximises the width of the gap between the two categories [163; 164]. However, other decision methods are used in the literature and will be further presented.

Furthermore, in a classification task the number and type of classes are also important aspects to take into account. Regarding emotion recognition, as explained in Section 2.1.1, there are discrete and dimensional models, resulting in discrete emotion classes (such as joy, anger, sadness) or parametric classes (such as arousal, valence, dimension). In this way, this section will evaluate different methods concerning both emotion models.

4.3.5.1 Discrete Emotion Classes

Although it is considered a rather limited mode of evaluating a person's emotional state, there is a lot of literature available using discrete emotion classes, since these simple models allow an easier and more direct way of assessing specific and concrete emotions. Furthermore, in parallel with

Emotion Recognition, there is also Stress Detection research using ECG, since stress is deeply connected with HRV, which leads to the use of identical features and methods. In Table 4.4.1, different studies are presented in this field [132; 139; 148]. Concerning [131], it tried to distinguish between 3 levels of stress (Negative, Neutral and Positive) by using time-domain fiducial features and both SVM and kNN classifiers. By investigating three different training and testing feature vectors, the best results were obtained for HR, MRAmp and mRRI, with a maximum classification accuracy of 61.7% for three classes and 69.6% for two. Another example is a pain intensity recognition algorithm, developed in [145]. In this way, it can be understood that ECG can have different applications concerning human physiological and psychological states.

However, concerning specifically the use of ECG to Emotion Recognition, in 2011, Guo *et al.* [52] used ECG recordings conveying four types of emotions: joy, anger, sadness and pleasure, collected while a subject was listening to music by the Augsburg University in Germany. After extracting the maximum value and standard deviation of wavelet coefficients, both RBF and BP neural network classifiers were used and compared. Although those are deep learning methods, there was a previous feature extraction in order to enhance the network performance. In this way, Guo was able to obtain 87.5% and 91.67% classification rates for BP and RBF neural networks, respectively, concluding the superiority of the RBF method. Nonetheless, it is important to mention that all the data used during this experiment belonged to a single subject, which could lead to a high subject-dependency methodology, not performing well when other people may be concerned.

Later, in 2016, Guo *et al.* [138] decided to develop a machine learning methodology based on 2 and 5 discrete emotion states. By using PCA as the feature selection method and SVM for the classification task, it was obtained a 71.4% accuracy for two classes and 56.9% for five classes. This result not only indicates a higher precision by deep learning methods, further developed and analysed in Section 4.4.2, but also the discrepancy between the accuracy obtained when the number of classes was reduced. This result is expected, having into account that more classes require a more powerful method, able of finer distinction between emotional states.

In [147], three different classification tasks were considered: valence (negative/positive), arousal (high/low) and discrete emotion tagging. By considering a total of four emotions in the last set mentioned, the classification rate was 61.52%, significantly low when compared with the 82.78% for valence and the 72.91% for arousal. Concerning the difference that exists between the arousal and valence level detection, it can be related to the fact that some emotions present similar levels of arousal while being completely different emotional states, which may be responsible for poor learning and predicting levels. Finally, Kim *et al.* [91] also observed a similar result, by using 3 and 4 emotional states and obtaining 78% accuracy for the first and only 62% for the latter. However, it is important to mention that, unlike the previous research [138], this system shows a recognition ratio considerably higher than the chance probability (33.3% and 25% for 3 and 4 emotion categories) for both classes. One of the reasons for this is the fact that Kim *et al.* uses a multimodal approach, which translates into a more complete data and the possibility of extracting more efficient patterns between different modalities. Nonetheless, Kim *et al.* [91] considers a signal division into segments, and further randomly splits them into train and test data, without

considering more challenging data divisons.

As already mentioned, different factors play an important role in algorithm performance, from the features extracted and the method of extraction to the classifier or the training and testing settings. In [155], a total of 4 classifiers were applied, namely Regression Tree, Bayesian Classifier, kNN and FkNN (Fuzzy k-Nearest Neighbours). Besides, this study also used two different methods of computing the Hurst parameter feature (HOS and traditional Hurst methods), as seen in Section 4.3.3. As a result, Selvaraj *et al.* noted that the best results were obtained by using a combination of FVS and HOS methods in a random setting with FkNN classifier (92.87%), followed by kNN (87.72%). Random Tree obtained results similar to kNN, however Bayesian Classifier performed poorly in almost all the classes and experiments (with varying features and settings), indicating that the independent probabilistic assumptions done by the classifier were not a good fit and do not suit the classification of emotional states.

Furthermore, the feature extraction methods used were also an important factor analysed in this research. When compared with all the results obtained by isolating different methods, it can be concluded that combining both non-linear analysis and HOS leads to finer emotional recognition. Finally, two different settings were considered: random and subject independent. As was expected, the best result for the random setting (92.87%) was significantly higher than for the subject independent, by achieving a maximum accuracy of 76.45%. This problem is present in all emotion recognition research, and other classification problems, due to subject-dependent factors, already mentioned in Section 2.2.2 and continues to be one of the major challenges to be mitigated in the AI field. In [153], a subject-independent emotion recognition system was developed, obtaining an accuracy of 52% for the recognition of 6 emotional states. Although better classification rates could be achieved using subject-dependent settings, these systems are unable to perform well with unseen data, which makes subject-independent approaches more reliable. In addition, signal-independent settings should also be considered when using segments as inputs, so as to understand if there is also signal dependency.

Another important aspect that can have an influence on the global performance achieved is the emotion elicitation technique. In [129], different emotion elicitation methods were used and compared, such as digital images, audio visual and audio stimuli. In ECG signals, it was understood that the use of digital images produced the lowest emotion classification rate (60.34%), followed by audio elicitation. On the other hand, the selected clips for the video and audio-video stimuli were the most effective on evoking emotion, with an accuracy rate higher than 68%, using SVM as the classifier. This research used a limited set of data, which can even lead to believe that, by augmenting the ECG data, the accuracy rate can even increase.

Finally, although kNN and SVM were the most used classifiers, other research showed good results by applying different decision methods. By defining only two discrete emotion states (joy and sadness), [141] used a Fisher Classifier and obtained 88.43 % accuracy. The success found in this research can be somehow associated with the high volume of data available and its variability, since 391 subjects were part of the experiment. However, in [123], both kNN and LDA were used, obtaining an average classification of 69.75% and 67.81%, respectively. Besides the difference

concerning the features extracted, time-domain fiducial for the first, and frequency-domain non-fiducial for the other, [123] used five different emotional states, which increases the recognition difficulty, and only a 20 subject experiment was carried.

In [142], the comparison made was between three different feature selection methods (ANOVA, SBS, SFS) and three classifiers (SVM, LDA, Fisher), reinforcing the knowledge that these methods can influence the overall algorithm capacity to detect emotions. The best results were obtained by combining SBS and SVM, for a classification rate of 88%. However, when combining all feature selection methods with SVM classifier, an accuracy of 92% was achieved. On the other hand, it was perceived that joy was always more easily (or equally) detected than pleasure, which can have a variety of reasons, namely the type of features obtained.

Finally, comparing the type of features mentioned in Section 4.3, one cannot specify that frequency or time-domain features are more accurate or prone to obtain better results, since the performance is also dependent on all the other factors already mentioned.

4.3.5.2 Dimensional Emotion Classes

Although dimensional emotion models are still more recent than discrete emotion characterization, there are already scientific studies that choose to describe an emotion as a continuous parameter, by defining it through different emotion-related parameters such as *valence* and *arousal*. In [147], as already mentioned in the previous section, both dimensional and discrete classes were considered, using a LS-SVM classifier and obtaining considerably high accuracies for a 2-class valence (positive/negative) and arousal (low/high) (82.78% and 72.91%, respectively), when compared with other methodologies such as [9], where 3 different binary classification schemes were defined for arousal (low/high), valence (unpleasant/pleasant) and dominance (control/empowered) and obtaining a mean average classification of 62% for the three parameters. However, these worse results can be due to the small database used, with only 23 subjects and low levels of variability. Furthermore, a 10-fold cross-validation was used in order to validate the user-independent classification performance, meaning that, the eighteen samples for each participant were randomly divided into ten groups, using 1 group for testing and the others for training. This implies that train and test sets consider signals from the same person, not having into account subject-independent settings. Although the same signal is not in both train and test sets, there is no subject independence in these data division scenarios. Nevertheless, DREAMER can be a good database to be used, since it also considers the dominance metric, which can provide better emotion recognition performance, and the small database problem can be reduced with some possible data augmentation techniques.

Soleymani *et al.* [15] considered three states of arousal (calm/medium/excited) and valence (negative/neutral/positive), obtaining a classification rate of 46.2% and 45.50% by using ANOVA as a feature selection method and SVM for the classification task. Nonetheless, when a feature fusion was considered (EEG and Gaze), results improved more than 30%, indicating the potential of a multimodal emotion recognition system.

However, features play an important role since they convey the information that may enable to distinguish between different classes. These three pieces of research had into account typical

and traditional features, concerning time-domain fiducial features based on specific ECG points and frequency-domain such as spectral power in VLF, LF, HF and LF/HF, as well as total power. However, other features less direct may also offer good results. In 2016, Ferdinando *et al.* [30] proposed a new set of features not yet tried, based on the statistical distribution of dominant frequencies, and using kNN to classify a 3-class problem in valence and arousal (low-medium-high). In the research, Ferdinando compared the results with similar algorithms using features from standard Heart Rate Variability analysis. As a result, the accuracies for valence and arousal increased 13% and 12%, respectively by using the statistical distribution of dominant frequencies. Furthermore, the proposed features also showed better performance when compared with features based on the statistical distribution of instantaneous frequency, calculated using Hilbert transform of IMFs, after applying EMD and BEMD.

In 2017, Ferdinando *et al.* [154] focused on proving the importance of the feature selection method, by using LDA, NCA and MCML in a set of features based on the statistical distribution of dominant frequencies after applying BEMD. The combination of a feature selection method (NCA presenting the best results) and a different feature extraction resulted in an improvement in both arousal and valence detection. However, in both approaches, Ferdinando *et al.* considered 5-second ECG segments as inputs, randomly divided into train/test sets, which can lead to higher results than when considering more realistic data splits.

Nonetheless, these approaches indicate that features should be suited to the purpose of the research. Kim and André [95] tried to identify the significant features for each classification problem, investigating the class-relevant feature domain for a given emotion. For arousal, the feature obtained from the time/frequency analysis of the HRV time series seems to be decisive, while features from the MSE domain of ECG signals are fundamental for a correct valence detection. This fact can put in perspective the results found in different researches that maybe didn't adjust the features to each metric. By taking into account the relevant features for each task, it was possible to achieve an average recognition accuracy of 98% for arousal, 91% for valence and 87% for 4 emotion classes, using pLDA as the classifier.

In addition, Kim and Andre [95] also developed a EMDC scheme, testing it in subject-dependent and subject-independent settings. For the first, a classification rate of 95% was achieved, whereas for the subject-independent setting the accuracy was 70%. The authors believe that the main reason for this can be associated with the non-emotional individual contexts of each subject, rather than possible ANS differences among emotions. By identifying the user prior to starting the emotion recognition process, an accuracy of 99% was obtained, proving how emotions are intrinsically associated with their subject and reinforcing the subject-dependency problem for the development of robust and subject-independent recognition systems. On the other hand, Goshvarpour *et al*. [26] were able to achieve satisfying results for both subject-independent and subject-dependent settings. By using PCA and PNN, with a sigma of 0.01, a recognition rate of 100% was achieved in all classification schemes. Although the overall results were still better in subject-independent settings considering other schemes, [26] illustrated that it is possible to use feature extraction and selection methods that enable high accuracy rates for unseen data.

Furthermore, emotion elicitation is also relevant. For the first time in 2012, by using dimensional models, Agrafioti *et al.* [23] differentiated between active and passive arousal experiments, concluding that there are higher chances of ECG reactivity to emotion when the induction method is active. This type of discoveries shows that it is possible to fit the method used to the characteristics of the classes to be distinguished, which can potentially allow for more accurate and precise hand-crafted methods.

Concerning the dimensional and discrete emotion classes, both allow obtaining good results, as it was already reviewed. However, dimensional models will be able to evolve as the recognition methods also evolve, allowing for the detection of more complex and similar emotions, and even the overlapping of emotion in the same moment.

4.4 Deep Learning Methods

4.4.1 Introduction

Deep Learning is a subset of machine learning and consists of neural networks with the ability to learn the input data and its increasingly abstract representations. In this way, there isn't the need for performing feature extraction since the network is able to detect and recognize the patterns conveyed in the data as it goes deeper into the network. Deep learning normally outperforms traditional machine learning, when it comes to large datasets, which adds a requirement to the deep learning method. However, easy steps such as windowing and the extraction of overlapping segments of the data, or even data augmentation can enlarge the dataset and ensure good performance. However, on the other hand, deep learning is more computationally challenging, since a large number of parameters is being trained. Nonetheless, it still presents itself as a groundbreaking field, with highly promising results and possible applications to Emotion Recognition systems.

There are different types of neural networks, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Generative Adversarial Networks (GANs), among others. Convolutional Neural Networks are one of the most used and widespread deep neural network models, consisting of a special class of feedforward neural networks, that eliminates manual feature extraction. A given input image or signal goes through a set of filters in the convolutional layer, followed by a pooling layer. At each convolutional layer, feature maps are obtained through successive convolutions between the input and the weight vectors (filters). Normally, in the end, fully connected layers work as final learning phases, where the features are mapped into the predicted outputs. On the other hand, Recurrent Neural Networks are a type of architecture mainly used to deal with sequential data, forming a unidirectional cycle between units. RNNs are able to accumulate past information, since it includes "memory" [165]. With the development of RNN models and the increase of complexity, some authors considered that RNN architectures were unable to identify and learn long-term dependencies in the sequence data. To overcome this problem, a widely used type of RNN appeared, called Long Short-Term Memory Networks (LSTMs) [166].

Table 4.2: Summary of the handcrafted methods used on the surveyed approaches (ordered by year and alphabetically by the surname of the first author). Legend: V - Valence; A - Arousal; D - Dominance; L - Liking; Em - emotions.

Paper	**Year**	**Database**	**Subjects**	**Modalities**	**Emotion Elicitation**	**Emotion Classes (Dimensional/ Discrete)**	**Classes**	**Classifier**	**Results**
Kim *et al*. [91]	2004	Private	50	ECG, SKT, EDA	Audio, Visual, Cognitive	Discrete	3 4	SVM	3 Em: 78.4% 4 Em: 61.8%
Kim and Andre [95]	2008	Private	3	ECG, EMG, RSP, SC	Music	Dimensional	4	EMDC	95% 70% (subj. independent)
Xu and Liu [141]	2009	Private	391	ECG	Movie clips	Discrete	2	Fisher	88.43%
Guo [52]	2011	Private (Augsburg)	1	ECG	Music	Discrete	4	BP NN RBF NN	87.5% 91.67%
Agrafioti *et al*. [23]	2012	Private	44	ECG	IAPS (passive) Video game (active)	Dimensional	2	LDA	89%**
Bong *et al*. [131]	2012	Private	5	ECG	Movie clips	Discrete	3	kNN SVM	61.65% 61%
Soleymani *et al*. [15]	2012	MAHNOB-HCI	27	ECG, GSR, RPS, SC	videos, images	Dimensional	3	SVM	A: 46.20% V: 45.50%
Murugappan *et al*. [123]	2013	Private	20	ECG	Movie clips	Discrete	5	LDA kNN	67.81% 69.75%
Selvaraj *et al*. [155]	2013	Private	60	ECG	Movie clips	Discrete	6	Regression tree Bayesian kNN FkNN	85.71% 65.82% 87.72% 92.87%
Xun and Zheng [142]	2013	Private (Augsburg)	80	ECG	Movie clips	Discrete	2	SVM	92%
Valenza *et al*. [92]	2014	Private	60	EEG, ECG	IAPS	Dimensional	2	SVM	A: 71.21% V: 75%
M *et al*. [153]	2014	Private	30	ECG	Movie clips	Discrete	6	kNN LDA	52% (kNN + DFT)
Godin *et al*. [12]	2015	DEAP MAHNHOB-HCI	32 27	GSR, ECG PPG, RA, SKT, EMG, EOG	short videos	Dimensional	2	Naive Bayes	varies with no. features

Paper	Year	Database	Subjects	Modalities	Emotion Elicitation	Emotion Classes (Dimensional/ Discrete)	Classes	Classifier	Results
Ménard *et al*. [151]	2015	Private	35	EEG, EMG, ECG	visual, dynamic	Dimensional	6	SVM	89.58%
Katsigiannis and Ramzan [9]	2015	DREAMER	23	ECG	video clips	Dimensional	2	SVM (+ RBF kernel)	A: 62.33% V: 61.84% D: 61.84%
Ferdinando *et al*. [30]	2016	MAHNOB-HCI	27	ECG	video clips, images	Dimensional	3	kNN	A: 59.7% V: 55.8%
Ping Gong [144]	2016	Private (Augsburg)	1	ECG, EMG, RSP, SC	Music	Discrete	4	C4.5 Decision Tree	95%
Guo *et al*. [138]	2016	Private	25	ECG	video clips	Discrete	2 5	SVM	2 Em: 71.4% 5 Em: 56.9%
Tivatansakul and Ohkura [24]	2016	Private (Augsburg)	1	ECG	Music	Discrete	3	LBP LTP	84.17% 87.92%
Ferdinando *et al*. [154]	2017	MAHNOB-HCI	27	ECG	video clips images	Dimensional	3	kNN	A: 66.1% V: 64.1%
Gjoreski *et al*. [140]	2017	MAHNOB-HCI , DEAP, DECAF, ASCERTAIN Cognitive Load Driving Worload	181 (total)	ECG, BVP, PPG (R-R)	video clips, images	Dimensional	2	AdaBoosting GB RF SVM	57% 58% 58.5% 65% 67%
Goshvarpour *et al*. [26]	2017	Private	11	ECG	Music	Discrete Dimensional	5 3	PNN	100%
Hsu *et al*. [147]	2017	Private	61	ECG	Music	Discrete Dimensional	2 4	LS-SVM	A: 72.91% V: 82.78% 4 Em: 61.52%
Miranda *et al*. [14]	2017	AMIGOS	40	ECG	video-clips	Dimensional	2	NB	A: 55.1%* V: 54.5%*
Minhad *et al*. [129]	2017	Private	19	ECG	digital images audio visual audio	Discrete	5	SVM	60.34% 71.92% 68.36%
Sriramprakash *et al*. [139]	2017	SWELL-KW	25	ECG, GSR	Different tasks (for stress)	Dimensional	2	SVM (+ RBF kernel)	92.75%

Paper	Year	Database	Subjects	Modalities	Emotion Elicitation	Emotion Classes (Dimensional/ Discrete)	Classes	Classifier	Results
Marín-Morales *et al*. [150]	2017	Private	60	ECG	virtual environments	Dimensional	2	SVM	A: 75% V: 71.21%
Wei *et al*. [146]	2018	MAHNOB-HCI	27	ECG	videos, images	Discrete	5	SVM	68.75%
Xiefeng *et al*. [157]	2019	Private	16	ECG	IAPS, IADS	Dimensional	4	SVM	V dimension: 96.86% A dimension: 88.54% A-V synthesis: 81.25%
Hsu *et al*. [27]	2020	Private	61	ECG	Music	Dimensional	2 4	LS-SVM	A: 72.91% V: 82.78% 4 Em: 61.57%

*F1 Score **Best in V axis

Based on the literature reviewed, it can be understood that ECG is less explored when compared to EEG. However, some researches using ECG alone or in a multi-modal approach were analysed. For these, different methodologies were applied mainly based on the architectures previously mentioned (RNNs, CNNs), although other approaches were also found, such as self-supervised methods.

4.4.2 Deep Learning Approaches

As previously mentioned, the absence of very large-scale datasets with physiological signals is a limitation as far as deep learning applications are concerned. Such datasets are available considering other modalities such as face and audio, since they are more easily acquired. In this way, different works used various databases publicly available to increase the amount of data. Concerning [140], 7 different databases were used: AMIGOS, DEAP, DECAF, MAHNOB-HCI, Driving Workload dataset, Cognitive Load Dataset and ASCERTAIN, both isolated and combined, in order to obtain emotion recognition from R-R intervals. Besides, as seen in Section 4.3.5.2, classical machine learning methods were also used and compared. For each database alone, the DNN proposed in this research was able to outperform different ML classifiers, namely AdaBoost, GB, RF and SVM. Furthermore, by a pre-processing where the 7 databases were merged into common spectro-temporal space, an overall accuracy of 75% was also obtained, which indicates the superiority of deep learning methods in all experiments.

Furthermore, the same superiority was found in [167]. Santamaria-Granados *et al.*, developed an interesting method using the AMIGOS database by considering a convolutional neural network for automatic feature extraction and fully connected network layers for emotion prediction. In addition, other techniques were also applied, such as dropout, in order to reduce the probability of overfitting. Finally, this study also compared results obtained with other classical ML algorithms and methods, such as the one used by the authors of the database [14]. The DCNN applied was able to achieve an accuracy and F1 score of 81% and 76%, respectively, for arousal and 71% and 68% for valence, overperforming all other methods concerned. In addition, this research used the preprocessing of the peaks as the input vector for CNN, making possible the identification of morphological features that showed to be suitable for emotion recognition. Nonetheless, the data split chosen by Santamaria-Granados *et al.* was also not the most realistic and the reported results correspond to random data splits, which may indicate that for more challenging settings, the model performance would decrease.

Furthermore, still concerning the lack of sufficient data for well-performed deep learning methods, [29] also used 4 databases (AMIGOS, DEAP, MAHNOB-HCI and DREAMER), in a multimodal approach, considering ECG, EEG, GSR and face expressions. Initially, the signals went through a simple pre-processing involving filtering and artifact removal, and after that, ELM was used as the classifier. For facial and EEG modalities, LSTM was also applied in order to learn temporal dynamics of the features extracted. Concerning the evaluating, each signal was considered separately to classify emotion in 4 different metrics: valence, arousal, liking and emotion categories. After that, the same metrics were evaluated in multimodal approaches, obtaining a

mean accuracy of 79.09%, 79.03%, 79.70% and 55.43% for arousal, valence, liking and emotional states, respectively, using combining EEG, ECG/PPG and GSR signals. Considering the improvement in accuracy performance compared with single modalities, these results also translate the power of multimodal approaches for emotion recognition.

On the other hand, another limitation considering the currently available databases has to do with the labelling process, since sometimes there is a reduced amount of data annotated or the process of labelling is not very trustworthy, as seen in Chapter 3. Self-supervised Learning is a kind of unsupervised learning where the data provides its own supervision, which means that the network uses automatically generated labels instead of human-annotated ones. By pre-training a model on unlabelled data and then fine-tune it on a smaller labelled set, interesting results can come to light. In [168] and [169], Sarkar *et al.* used self-supervised methods, obtaining positive results. Concerning [168], the proposed architecture consists of a multi-task CNN trained with automatically generated annotations. Then, the weights go to the emotion recognition network, where fully connected layers do the classification task. This proposed method obtained an accuracy of 84% and 85.8% for valence and arousal classification, respectively, in the AMIGOS database. Concerning SWELL, even higher classification rates were obtained for stress, arousal and valence classes. These results were compared with other ML and DL methods, such as the ones already mentioned [167]. Furthermore, the method was also applied with full supervised learning, obtaining lower results and proving the self-supervised potential.

Similarly, [169] also performed well in all of the four different databases used, obtaining accuracies higher than 85% for all, when considering two or three arousal/valence classes. Furthermore, the method distinguished between 9 levels of arousal and valence for AMIGOS and SWELL, and 5 levels concerning DREAMER, obtaining average accuracies of 79%, 93% and 76%, respectively. However, as stated by the author, there are also some limitations concerning this method, such as its poor performance for subject-independent settings, which [169] considers it could be related to some need of more specific and hand-crafted features that can be subject-invariant. In addition, 10-second segments are considered during the entire methodology, with random 90%-10% train/test split, which indicates that the sets used are not signal-independent.

However, in order to combat the need for more specific features, there are some other approaches available. End-to-end learning is a deep learning process in which all of the parameters are obtained and learned inside of the deep learning network. Common methods, either from ML or even pre-processed data fed into a deep learning architecture, rely on hand-crafted features that may not provide the performance necessary for real-life applications. With End-to-end approaches, raw signals are directly fed into the network, that is able to learn an intermediate representation of the raw input in a way that better suits the task [25]. This fact mitigates the need for a high knowledge concerning the most specific and relevant features for each classification problem.

Keren *et al.* developed the first study on End-to-End learning of emotion from physiological signals, which proved to yield large improvements compared with hand-crafted methods on the RECOLA database. In [93], both DREAMER and AMIGOS were used, achieving a peak accuracy

of 90%. In this approach, the input IBI time series data segments are fed to 1D convolutional layers and a bi-directional LSTM. Then, the output of both these streams is concatenated before entering a final dense layer, where the prediction is made. Furthermore, a Bayesian framework is used for modelling the uncertainty in valence predictions. These types of probabilistic neural networks may have a high applicability in environments where confidence plays an important role in decision-making, enhancing the value of such approaches.

Dar *et al*. [170] proposed a different approach, using a combination of LSTM and CNN architectures to improve the recognition of four emotional classes (HVHA, HVLA, LVHA and LVHA). This approach presents an accuracy of 99.0% for AMIGOS database and 90.8% for DREAMER. Furthermore, an average accuracy of 98.73% is achieved with ECG right-channel modality. On the other hand, 1-second segments are considered in this approach, for both AMIGOS and DREAMER databases, which are then randomly divided into 70% train and 30% test sets. Having this said, Dar *et al*. [170] presents an interesting approach through the combination of CNN and LSTMs networks. However, this model should also be tested under more realistic data splits so as to fully analyse this methodology's robustness.

In this way, having into account the state-of-the-art presented, deep learning proves to be an innovative field with promising results for more accurate and automatic emotion recognition methods, that can someday be used in real-life contexts. However, for this to happen, some methodologies should be first analysed under more strict data divisions, since performance usually drops in these cases. Furthermore, specific techniques can be then applied to improve the overall performance for these realistic settings.

4.5 Summary and Conclusions

In this chapter, various studies on the state of the art have been analysed, allowing some conclusions to be drawn. Regarding the handcrafted methods, several approaches have already been developed, showing that good results can be obtained by considering different features. However, nonlinear features can be explored more as some studies have shown their promising application. Similarly, it is shown that feature selection is a useful method by removing unnecessary or irrelevant features and enabling high performance of the developed algorithms.

Hand-crafted methods are much more explored and present in the literature, and most researchers have chosen to use discrete emotion classes. On the other hand, if we consider deep learning approaches, dimensional metrics have always been the first option to describe and recognize different emotional states. Although deep-learning methods for emotion recognition using physiological signals have not yet been widely explored, existing studies have yielded promising results that illustrate the potential of these powerful methods to lead to increasingly automatic ways of emotion recognition. Nonetheless, most of these approaches consider more flexible and random data division scenarios, which can emphasize some performances, obtaining higher results than it would be obtained if signal and subject independence were ensured. However, as

Table 4.3: Summary of the deep learning methods used on the surveyed approaches (ordered by year and alphabetically by the surname of the first author). Legend: V - Valence; A - Arousal; D - Dominance; L - Liking; Em - emotions.

Paper	Year	Database	Subjects	Modalities	Emotion Elicitation	Emotion Classes (Dimensional/Discrete)	Classes	Classifier	Results
Chao *et al*. [171]	2015	RECOLA	46	ECG, speech, face, EDA	video clips	Dimensional	2	LSTMs-RNNs	A:71.6 %* V:61.8&*
Kächele *et al*. [172]	2015	RECOLA	46	ECG	video clips	Dimensional	2	CCC-NN + RF	A: 0.546** V: 0.479**
Chen and Jin [173]	2015	RECOLA	46	ECG	video clips	Dimensional	2	RNNs-BLSTMs	A:73.9%* V:56.7%*
Gjoreski *et al*. [140]	2017	MAHNOB-HCI , DEAP, DECAF, ASCERTAIN Cognitive Load Driving Workload	181 (total)	ECG, BVP, PPG (R-R)	video clips, images	Dimensional	2	fully connected DNN	75%
Keren *et al*. [25]	2017	RECOLA	46	ECG, HR, EDA, SCL, SCR	video clips	Dimensional	2	CNNs + RNNs	A: 0.430** V:0.407**
Lin *et al*. [174]	2019	WESAD	15	ECG, EDA, EMG, RESP, SKT chest, ACC, BVP	different tasks	Discrete	3	MMSF (RF)	85%
Yin *et al*. [54]	2017	DEAP	32	EEG, physiological signals, Face	music video	Dimensional	2	MESAE	A: 77.19 V:76.17
Gjoreski [175]	2018	ASCERTAIN DEAP Driving Workload Cognitive Load MAHNOB-HCI AMIGOS	191	GSR, ECG, BVP	video clips, images	Dimensional	2	DNN	70.3%
Kawde and Verma [176]	2018	DEAP	32	EEG, physiological signals, Face	music video	Dimensional	2	SAE + DNN	A: 73.08% V:78.84% D: 65.38%

Paper	Year	Database	Subjects	Modalities	Emotion Elicitation	Emotion Classes (Dimensional/Discrete)	Classes	Classifier	Results
Siddharth *et al*. [177]	2018	AMIGOS	40	ECG	video clips	Dimensional	2	ELM	A: 57.94
							4		V: 58.73%
							8		D: 55.56%
									L: 69.05%
									4 Em: 28.57%
									8 Em: 28.57%
Santamaria-Granados *et al*. [167]	2019	AMIGOS	40	ECG, GSR	video clips	Dimensional	2	DNN-FCN	A: 81%
									V: 71%
Siddharth *et al*. [29]	2019	AMIGOS	40	EEG, ECG, GSR	video clips	Dimensional	undefined	LSTMs	A: 79.09
		DEAP	32						V: 79.3%
		MAHNOB-HCI	27						L: 79.70
		DREAMER	23						Em: 55.43%
Dar *et al*. [170]	2020	AMIGOS	40	ECG, EEG, GSR	video clips	Dimensional	4	CNN-LSTM	99.0%
		DREAMER	23						90.8%
Harper and Southern [93]	2020	AMIGOS	40	IBI	video clips	Dimensional	2	CNN + LSTM +	90%
		DREAMER	23					Bayes	86%
Chao *et al*. [178]	2020	AMIGOS	40	EEG, ECG, GSR	video clips	Dimensional	2	BLSTM + Attention + DNN	67.8%
Sarkar and Etemad [168]	2020	SWELL	25	ECG	video clips	Dimensional	2	Self-Supervised	96.15%***
		AMIGOS	40					CNN	84.9%***
Sarkar and Etemad [169]	2020	AMIGOS	40	ECG	video clips	Dimensional	2	Self-Supervised	88.2%***
		DREAMER	23				2	CNN	85.45%***
		WESAD	15				3		96.6%***
		SWELL	25				2		95.77%***
Oh *et al*. [179]	2020	Private	53	HRV, RSP	video clips	Discrete	6	CNN	94.02%

*F1 Score **CCC
***A, V mean

mentioned, some approaches also evaluated their methodologies in these settings, showing considerable performance decrease, as seen in [155].

Having this said, this dissertation pretends to focus on the development of deep learning architectures, building robust networks that can handle noisy signals and still recognize all the important patterns for the task at hand. For this, it is possible to use different deep learning architectures, such as RNNs so as to highlight the temporal information conveyed in the ECG signal, which can be useful for emotion recognition. Furthermore, although deep learning has shown promising results that outperform hand-crafted methods, it becomes important to understand if the literature reported results are as promising when realistic data divisions are concerned and what is the true impact of applying these settings to a model that can achieve high performances for random data splits. In conclusion, this dissertation will focus on the development of a robust and accurate deep learning architecture, analysing its performance for increasingly realistic data splits.

Chapter 5

Comprehensive Comparison of Prior Art Approaches

5.1 Introduction

After analysing the most common methods applied for emotion recognition using ECG signals in Chapter 4, the initial work developed focused on replicating some of those methods. As mentioned, this dissertation main goal is to develop a deep learning architecture able to obtain high emotion recognition performances by considering dimensional classes, namely valence and arousal. In this way, having the already available techniques presented in the literature in consideration, it is possible to use them as a basis to build the following work.

In addition, motivated by the limitations already mentioned throughout this dissertation, especially in Chapter 4, which consist of the recurrent use of signal segments and random splits of train and test sets, using exclusively these setups for model evaluation, a more thorough analysis should be done, considering other settings to assess model performance concerning more strict and realistic scenarios.

Furthermore, besides replicating some approaches, some possible alterations were taken into account, which will also be described in this chapter. Taking into account the promising results obtained when considering temporal information or other forms of aggregation, techniques such as LSTMs and MLPs were also explored, based on some literature methodologies.

5.2 Methodologies

5.2.1 Santamaria-Granados *et al.*'s Deep Convolutional Neural Network

In addition to the overview of the fundamentals and prior art methods on emotion recognition, several preliminary tasks were also conducted, allowing us to get acquainted with Pytorch working mode, the available databases, and some possible techniques and approaches for emotion recognition.

In this way, the first replicated approach was developed by Santamaria-Granados *et al.* [167]. It consisted of a simple neural network with a convolutional and fully-connected layers so as to recognize the levels of arousal and valence.

5.2.1.1 Methodology

Although this first approach had into account an already developed technique, there were some differences regarding the methods used. First, Santamaria-Granados *et al.* [167] considered both the ECG and the GSR signals, while in this work only the ECG signal was used. Furthermore, some initial preprocessing was also applied in the published work, whereas in this methodology the signal was only normalized before being fed to the deep learning network.

Furthermore, while Santamaria-Granados *et al.* [167] used 2-minute ECG segments as input, in this initial replication smaller segments were considered, since 2 minutes is an extended period of time in which more than one emotion could be felt, which would hinder the emotion recognition task. In addition, all the recordings in MAHNOB-HCI contain 30 seconds before and after the emotion elicitation. Thus, the first 30 seconds of each ECG signal were discarded. On the other hand, concerning the final 30 seconds, with the end of emotion elicitation, the emotional state can still be felt and prolonged, so this segment was considered. In Santamaria-Granados *et al.*, the database used was AMIGOS, which explains why this procedure was not considered. It is important to mention that the database used was not the same since it was not already available when this methodology was replicated. Furthermore, as the first network developed, some variations were tried, in order to test and to understand what could work better in emotion recognition considering further and future implementations. Since these were the first experiments tested, different experiments were conducted and analysed, to find the most appropriate kind of setups for emotion recognition tasks.

After the normalization, the DCNN proposed by Santamaria-Granados *et al.* [167] involved a sequence of Convolutional Neural Network (CNN) layers and pooling layers, used to extract the relevant features of the ECG signal automatically. The output of this sequence is then fed to a total of 3 fully connected layers (FCN), to predict the affective state, as presented in Figure 5.1.

As already mentioned, different experiments were conducted, considering variations in the number of classes for both valence and arousal, as well as the train and test settings. In the following sections, each experiment is described, and the results are further presented and discussed.

5.2.1.2 Experimental Settings

A variety of experiments was conducted, and they can be distinguished by the ECG segment size considered, the presence or absence of overlap, the number of classes for arousal and valence recognition, and the training and testing sets considered. Furthermore, other alterations were also applied, to determine how they may influence the final classification, such as the optimizer and loss function used.

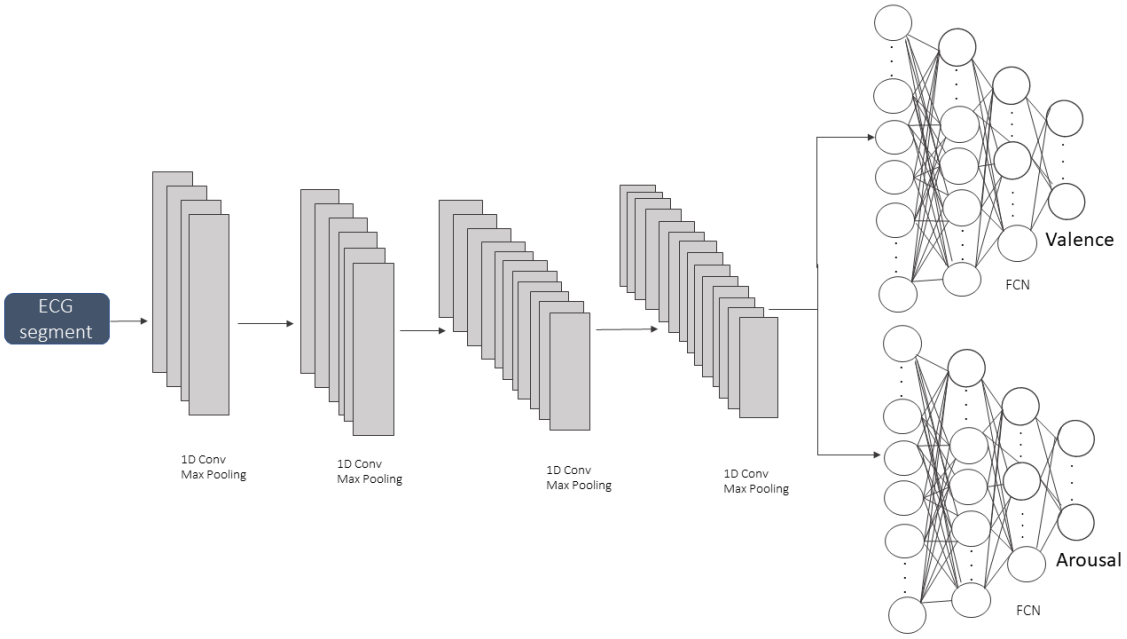


Figure 5.1: Scheme of the deep Learning architecture implemented for emotion recognition.

Experiment 1 - Impact of Model's Complexity and 30-second Segment Overlapping

In the first experiment, the size of the filters used in the CNN layers was 5, while the number of filters differed in each layer. The first experiment used 16, 16, 32 and 32 filters in each one of the 4 CNN layers. Considering the FCN, 3 layers were firstly applied, considering 150, 100 and 50 neurons, respectively. As a regularization technique to decrease the overfitting, a dropout of 0.5 was added in the fully connected layers. Furthermore, during the supervised training, a Cross-Entropy loss was minimized with the Root Mean Square Propagation (RMSProp) optimizer and the initial learning rate was set to 0.001. The cross-entropy loss function was obtained by considering an individual loss for each task (arousal and valence classification) and then computing the sum between them.

Finally, an 80%-20% division was done for the training and testing sets, and a 90%-10% for the training and validation sets. It was done a division by videos, i. e., this first setting guaranteed that the videos present in the training set were not in the testing set.

Using 30-second segments with no overlap and 9 classes for both arousal and valence, it was obtained a high level of overfitting, presented in Section 5.3.1.1. Due to these unsatisfactory results, we tried to decrease the model's complexity by reducing the number of filters in the CNN layers and the number of FCNs layers. Furthermore, dropout was also increased to 0.8. However, all these changes didn't lead to the results expected.

In this way, in order to increase the amount of data, 30-second segments with 15 seconds of overlap were used, which doubled the number of segments available. However, the results remained unsatisfactory and 5-second overlaps were then considered. Furthermore, other changes were also applied to this last approach, such as batch normalization, Adam optimizer instead

of RMSProp, and two separate loss functions for arousal and valence, allowing to define distinguished class weights, to avoid imbalanced classes. In addition, the ReduceLROnPlateau() function from PyTorch was also applied, to reduce the learning rate if the model didn't show any improvement after a given number of epochs.

Experiment 2 - Impact of the decrease in the number of classes and segments sizes

For the 2^{nd} experiment, the classes considered for each task were reduced from 9 to 3 and 2. Since the ECG signals in MAHNOB-HCI were labelled from 1 (low) to 9 (high), both valence and arousal were evaluated as negative (1,2,3), neutral (4,5,6) or positive (7,8,9) for the 3-class problem, and as negative (1,2,3,4) or positive (5,6,7,8,9), for the binary classification. This approach was conducted using 3 different segment sizes: 30, 15, and 10 seconds, in order to study the effect of the input segment size.

Furthermore, the changes regarding the loss function and optimizer were also applied as well as a 5-second overlap for each one of the 3-sized segments. Each CNN layer was composed of 16 filters of size 5, and only 1 FCN layer was applied, receiving the CNN output and predicting the affective states.

In this experiment, 2 different settings were considered. The first was equal to the one used in Experiment 1, dividing the sets in videos, while the second considered a subject-independent approach, by forming a test set with "unseen" subjects. This division was applied to identify a possible subject-dependency problem and evaluate the differences in performance between the two settings.

Experiment 3 - Impact of random data division and no overlapping

Finally, the last experiment was closer to the method developed in [167], by considering a random 90%-10% train-test and train-validation split. However, besides this random setting, a 2^{nd} setting was also used, considering a subject-independent division. In this way, 10-second ECG segments were extracted from each recording, without any overlap between segments, and the model's specifications concerning the CNN and FCN layers were kept the same. Finally, also two different tasks were considered for both valence and arousal, consisting of three-class and two-class problems, and the results are further analysed in 5.3.1.3.

5.2.2 Dar *et al.*'s CNN and LSTM-Based Emotion Charting

Considering the unsatisfactory results obtained with the replication of the work of Santamaria-Granados *et al.*, it became logical to consider different architectures that could be able to perform better in recognizing emotions.

Having this said, Dar *et al.* [170] developed an interesting technique by combining CNNs and LSTMs in order to recognize emotions. Different physiological signals were used, namely ECG, EEG and GSR. Since all of them are continuous-time signals with high memory content, LSTMs present themselves as promising neural networks, able to exploit relevant features. LSTMs

have the power of selectively remembering patterns for a long time, which becomes interesting to extract important features from physiological signals.

Dar *et al*. [170] exploited both multi and unimodal approaches, by considering the signals individually and combined. During this replication, the only signal used and compared was the electrocardiogram.

The proposed algorithm consists of three main steps: pre-processing, classification and multimodal fusion, which was not tested, as already mentioned. Before entering the neural network, all physiological signals needed to be pre-processed. Since EEG is significantly different from the other peripherical signals and was not used, only the processes applied to the ECG signal are further described and analysed.

The pre-process applied to the data is divided into basic and specific pre-process techniques. First, all signals are downsampled, loss-pass filtered and segmented into 1-second segments. After that, considering the specific pre-processing, the baseline is removed from all of these segments, and, finally, a z-score normalisation is applied.

5.2.2.1 Datasets Preparation and Pre-processing

The method presented in [170] uses two publicly available datasets, already described in Chapter 3: DREAMER and AMIGOS. The AMIGOS dataset presents both raw and pre-processed data. The pre-processed one is firstly downsampled from 256Hz to 128Hz and subsequently filtered with a loss-pass filter of 60Hz, in order to remove the high-frequency noise components. In this way, the approach uses the already pre-processed data of AMIGOS, and applies the same pre-processing steps to the DREAMER dataset, which only offers raw ECG data.

According to Dar *et al*. [170], DREAMER presents signals with high length variation, since it uses recordings with significant duration differences. Thus, the authors consider that there are some trials much larger in length to be significant for specific emotions. For that reason, considering this dataset, only the last 60 seconds of each trial are used, unlike AMIGOS where the total number of instances is exploited. Having this said, after filtering the ECG signal and considering only 64 seconds of each trial, 2 seconds from the start and the end of the signal are removed to counter the effect of the filter at the edges, and thus obtaining 60 seconds in total (7680 samples). Since DREAMER has data from 23 different subjects while watching 18 different videos and collecting 2 ECG channels, the total size of ECG data after this basic pre-processing is given by 7680 x 23 x 18 x 2.

After the basic pre-processing, there is the need to remove the baseline. Both DREAMER and AMIGOS datasets have baseline signals, where there is no emotional stimulus provided. In this way, it works as a subject-specific emotional ground-truth, since each person presents a different resting emotional state. Thus, removing this neutral baseline activity from all ECG signals is a specific pre-processing step that can have benefits for emotion recognition.

AMIGOS has a 5-second baseline segment, recorded at the beginning of each trial, while DREAMER has an independent baseline recording of 57 seconds before each experiment. For

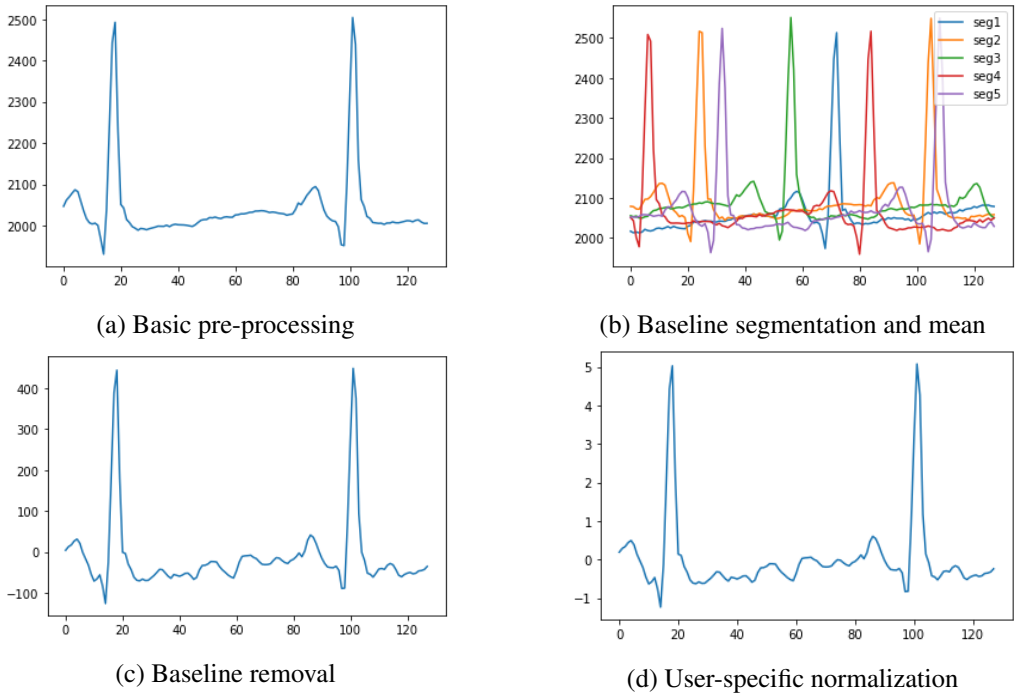
this purpose, each baseline signal was divided into 1-second segments, followed by the mean computation of all the segments in order to obtain the mean baseline activity of each signal.

$$meanBL = \frac{1}{S}(\sum_{s=1}^{S} BL_s) \tag{5.1}$$

The 60-second signals obtained by the basic pre-processing are then divided into 60 segments of 1 second each. After that, each segment of emotional activity is subtracted from their corresponding mean segment of baseline activity, so as to remove the neutral emotional effect.

$$blrSig_s = Sig_s - meanBL \tag{5.2}$$

For the DREAMER dataset, the baseline was computed from the mean of the 57 segments of baseline activity. Finally, the last pre-processing step was to normalize the ECG signal using Z-score normalization. Figure 5.2 present a given ECG segment after its basic pre-processing steps, as well as 5 different baseline segments, its removal and, finally, the Z-score normalization of the baseline removed signal. Normalization allows signals do be converted to a common scale, by having a unity standard deviation and zero mean.



(a) Basic pre-processing

(b) Baseline segmentation and mean

(c) Baseline removal

(d) User-specific normalization

Figure 5.2: Pre-processing ECG steps for data preparation, starting with basic pre-processing 5.2a, baseline mean computation 5.2b, its subtraction to the ECG signal 5.2c and user-specific normalization 5.2d.

Furthermore, contrary to most of the affective computing literature based on physiological signals, which classify valence (high/low) and arousal (high/low) as separate binary classification tasks, the approach defined four classes of emotions, having into account the levels of arousal and

valence combined: High Valence High Arousal (HVHA), High Valence Low Arousal (HVLA), Low Valence High Arousal (LVHA) and Low Valence Low Arousal (LVLA). Since AMIGOS and DREAMER present a different number of classes, they used different thresholds to distinguish from high and low levels. For AMIGOS, which has nine classes, low valence and low arousal consisted of labels equal to 5 or lower, while above 5 it is considered to be a high level of valence or arousal. Considering DREAMER and its five classes, low levels of arousal or valence consisted of labels equal or inferior to 3, while high levels corresponded to 4 or 5 annotations.

5.2.2.2 DNN Architecture

The architecture developed by Dar *et al*. [170] focuses upon the combination of LSTM and 1D convolutional layers, since LSTM is expected to fully exploit all the potential of time-series data. Figure 5.3 presents the architecture considered, where one-second signals are captured as sequences, and go through two 1D-convolutional layers, followed by ReLU activation and max-pooling layers, in order to extract temporal features from the time-series data. The first convolutional layer contains 16 filters, with a kernel size of 3 and the second has 32 filters of the same size. Concerning the max-pooling layers, it presents a size of 2 with a stride of 1. Then, these extracted features are flattened for the LSTM layer, with 128 hidden nodes. This layer learns the dependence between extracted temporal features that, afterwards, go through three dense layers, which provide the learning of prediction probabilities from extracted features. The first dense layer contains 256 hidden nodes, followed by 128 and 4, in the last one. Between each fully-connected layer, a dropout of 50% is applied, discarding 50% of random features and avoiding overfitting. Finally, a softmax layer is applied for the classification of HVHA, HVLA, LVHA and LVLA classes of emotion.

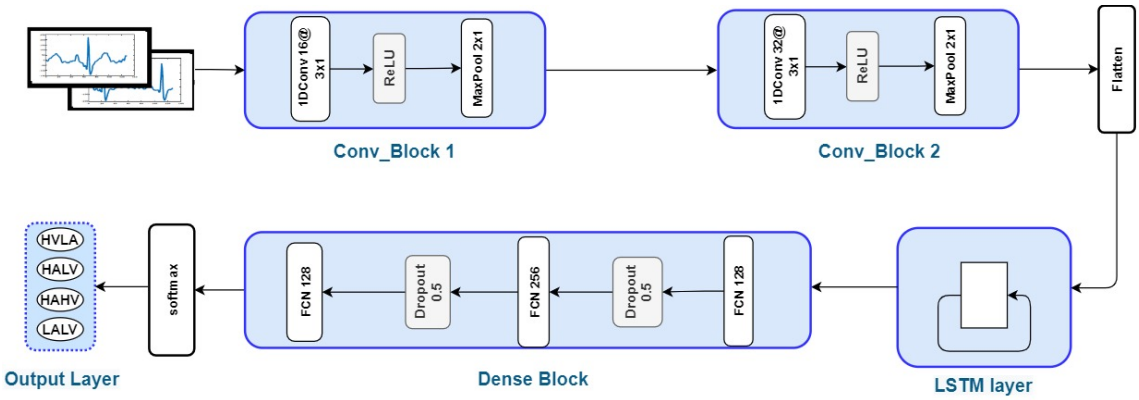


Figure 5.3: Deep neural network (1D-CNN + LSTM) design

Experimental Setup

Furthermore, in order to obtain the results, pre-processed data is then randomly split into 70% training data and 30% test data. The training parameters used by Dar *et al*. [170] were also

considered, such as a minimum batch size of 240, an initial learning rate of 0.001 with ADAM optimizer and a gradient squared decay factor of 0.99. However, different parameters of this approach were not available, such as the number of segments to consider for the LSTM layer. Since Dar *et al*. [170] refers that only the last 60 seconds of each signal from DREAMER database is used, in the replication it was considered 60 segments of 1-second at each time in the network. Thus, after pre-processed as 1-second signals, they were reset as 60-second signals again, received by the network and finally divided into 60 1-second segments before entering the convolutional layers. In the LSTM layer, all these segments belonging to the same signal were then aggregated. For this reason, the batch size used was 4, since each signal corresponded to 60 segments (60x4 = 240). Furthermore, other parameters remained not defined by Dar *et al*. [170], such as the padding used in each layer. All these factors make it more difficult to replicate a given approach, which can result in a discrepancy of results.

5.2.3 Sarkar and Etemad's Self-supervised ECG Representation Learning for Emotion Recognition

After the previously described methods were replicated, the idea was to focus on a different paper, with promising and reliable results, a distinctive methodology and a clear presentation of the data used, its processing, and all the relevant training parameters, such as batch size and learning rate, but also the number, size and stride of each layer. Thus, the choice was to replicate the approach used by Sarkar and Etemad [169].

As already analysed in Chapter 4, the majority of machine and deep learning techniques for ECG-based emotion recognition use fully-supervised learning methods, which means that the model is trained from scratch for every classification task. Furthermore, these models are often extremely specific for the task at hands, which leads them not to generalize well for other possible tasks. In addition, fully-supervised learning requires large datasets, with human-annotated labels, since small datasets often results in poor model performances with deep learning networks. [169].

In order to tackle all these problems, Sarkar and Etemad [169] present a self-supervised learning approach, inspired by the success obtained using multi-task self-supervised learning in other fields [180].

In self-supervised learning, the network is trained using automatically generated labels instead of human-annotated ones, resulting in more generalized features than task-specific ones. Because of this, this kind of models can then be reused for other tasks, improving the performance and decreasing the computation time, compared to fully supervised learning. Finnally, since labels are automatically generated, it is possible to use smaller datasets and also obtaining good results, since this approach will work as a type of data-augmentation. Different techniques can be used to obtain these automatic labels. For example, in [181], high-level spatiotemporal features were learned from unlabelled videos and the network was trained using rotated video clips to predict rotation. After that, the trained model was fine-tuned on the action recognition datasets [169]. Their research improved more than 15% compared to the fully-supervised approach. Furthermore, both [180; 182] also present self-supervised learning methodologies, obtaining promising results,

which points out the advantages of using self-supervised learning in other domains. Sarkar and Etemad [169] were the first to apply and propose its use for ECG representation learning, first in [168] and then further improved with some adjustments and alterations in [169].

Sarkar and Etemad [169] implemented the proposed architecture using Tensorflow, and shared the implementation in Gitlab [183], which provides a high level of confidence and helps in the replication to Pytorch, in this thesis. In the next subsections, the methodology will be described, presenting the main differences between the replication and the actual approach.

The proposed methodology can be divided into two stages of learning: first, learning ECG representations and, finally, using the previously trained network, capable of extracting ECG representations, to learn another network to recognize emotions. Thus, two sets of tasks were defined, referred to as T_p and T_d, where T_p represents the set of signal transformation recognition tasks and T_d learns to recognize emotions. T_p learns robust and generalized features from unlabelled ECG through recognition of signal transformations. After that, in the second phase, the original ECG signal and the human-annotated emotion labels are used, and T_d is learned. In this final stage, the network uses the frozen convolution layers of the self-supervised network in order to learn emotional classes upon supervised learning of the fully-connected layers at the end of the network [168].

Figure 5.4 presents an overview of the proposed and replicated self-supervised solution. This approach uses four publicly available datasets for the self-supervised learning: AMIGOS [14], DREAMER [9], WESAD [120] and SWELL [18]. By combining 4 different datasets, Sarkar and Etemad [169] use a high amount of data, which results in good performances of both the self-supervised learning and the emotion recognition tasks. However, during this replication, there were only used DREAMER and AMIGOS, since the quality of the self-supervised learning does not sharply decrease by using fewer datasets since the amount of data is already considerably high.

In the following subsections, the method's architecture is further described, presenting the details of each learning stage and the hyperparameters considered in the network.

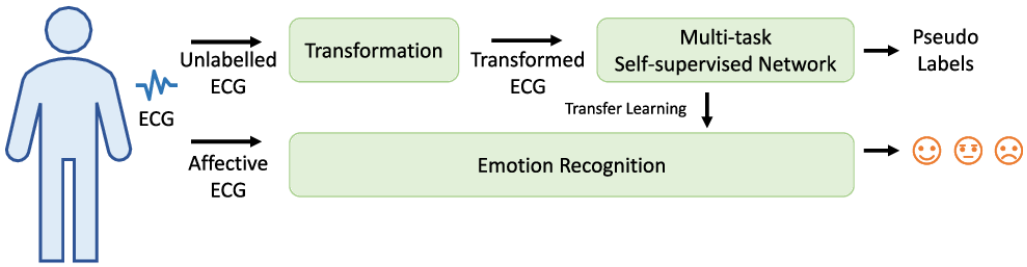


Figure 5.4: Overview of of the proposed framework for self-supervised emotion recognition, from [169]

5.2.3.1 Self-Supervised Learning

Sarkar and Etemad [169] propose a self-supervised deep multi-task CNN, that should be able to recognize different signal transformations applied to the original ECG. Six transformations are

applied to the input signal, resulting in seven different recognition tasks, including the original signal recognition.

Transformation Dataset

Sarkar and Etemad [169] used each of the four already mentioned datasets individually to perform the self-supervised task, and then combined them all. However, during the replication, only DREAMER and AMIGOS were used for the signal transformation recognition, which consisted of the optimization of convolutional weights for the emotion recognition network.

Nonetheless, it is important to mention that, since the code is available in [183], it was possible to extract features from the frozen convolutional layers directly. Thus, during this replication, both methodologies (using the extracted features directly into the emotion recognition network and replicate the entire process, including the self-learning tasks) were applied and their results shall be presented and discussed, afterwards.

In this way, there is the need to create a dataset composed by original signals and their transformations for the self-supervised task. These six transformations are noise addition, scaling, negation, temporal inversion, permutation and time-warping:

- *Noise addition:* Random noise from a Gaussian distribution, $N(t)$, is added to the original signal. $N(t)$ has mean 0 and the standard deviation of $\sqrt{E_{N_{avg}}}$, where $E_{N_{avg}}$ is the average power of $N(t)$. The noise-added signal is generated as $S(t) + N(t)$ and the noise amount used was fixed at 15;

- *Scaling:* The original magnitude of the signal is altered through the multiplication of an integer, β. This scaling factor is manually assigned and value 1.1 was used in the self-supervised learning tasks;

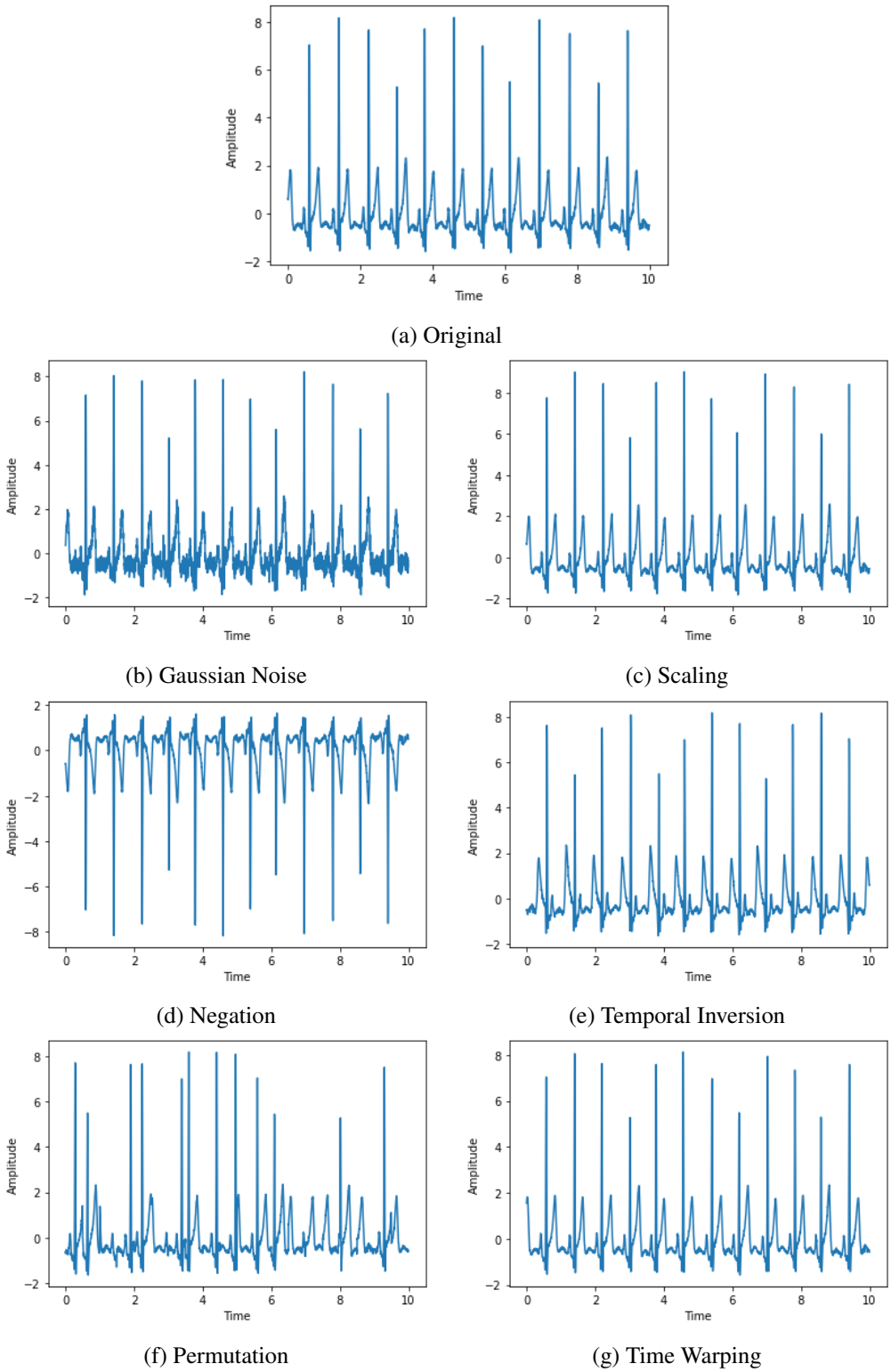
- *Negation:* The amplitude of the ECG signal is multiplied by -1, which causes a spatial inversion of the time-series data amplitude;

- *Temporal Inversion:* The original ECG is a temporal signal, where $t = 1, 2, ..., N$, and N is the length of the signal. Temporal Inversion consists of inverting the time-series data, obtaining a signal where $t = N, N-1, ..., 1$;

- *Permutation:* The ECG signal is divided into m segments and shuffled, which alters the temporal location of each segment. The value of m was set to 20;

- *Time-warping:* Randomly selected segments of the ECG signal are stretched or squeezed, along the time axis. For the self-learning task, it was used 9 time-warping segments, a stretching factor of 1.05 and a squeeze factor of $1/1.05$.

Each of these transformations is applied to the original signals, as seen in Figure 5.5, and each one has a specific purpose. For example, permutation creates signal discontinuities by altering the natural order of the segments. The model becomes capable of ignoring this kind of discontinuities,



(a) Original

(b) Gaussian Noise

(c) Scaling

(d) Negation

(e) Temporal Inversion

(f) Permutation

(g) Time Warping

Figure 5.5: **Signal Transformation for Self-Supervised Learning** with six different signal transformations.

becoming more robust. On the other hand, negation, for example, inverts the signal waves, while gaussian noise uses the addition of high-frequency noise. All of these transformations that could difficult the signal analysis are learnt by the model, which is then able to ignore these types of noise and alterations, focusing on the important patterns of the ECG signal. The outcome is then stacked, resulting in the input matrix X, while the labels are 0,1,....6, according to the transformation applied, and being 0 the original and unchanged ECG data. Data and labels are then shuffled and re-ordered in a random way.

Self-Supervised Learning Architecture

The proposed network for the ECG representations learning, using the created dataset described in 5.2.3.1, consists of 3 convolutional blocks (conv-blocks) and 7 task-specific layers (one for each transformation, encompassing the original signal), composed of 2 dense layers. The conv-blocks are formed by 2 1D convolution layers, followed by a ReLU activation function and a max-pooling layer of size 8. The number of filters used in these 3 conv-blocks increase from 32 to 64 and, finally, 128. In turn, the kernel size decreases from 32 to 16 and 8, respectively. After the 3 convolutional blocks, a global max-pooling is applied.

Before the fully connected layers, the extracted features are flattened and then fed to 2 fully connected layers with 128 hidden nodes and ReLU activation functions, each. Furthermore, a dropout of 60% is also used, in order to prevent overfitting.

Finally, the output of each one of the 7 task-specific layers goes through a sigmoid layer, since it is a binary classification (is or not transformed with transformation x). Table 5.1 presents the general architecture structure and the hyperparameters already mentioned.

During the model training, it was used 10-fold cross-validation and a random division of 90%-10% of the data into training and testing sets, respectively. However, in replication, cross-validation was not applied. Furthermore, ADAM optimizer was used with a learning rate of 0.001 and a batch size of 128. This network was trained for 100 epochs.

Table 5.1: Signal transformation recognition network architecture and hyperparameters, from Sarkar and Etemad [169].

Module	**Layer Details**	**Feature Shape**
Input	–	2560×1
Shared Layers	$[conv, 1 \times 32, 32] \times 2$ $[maxpool, 1 \times 8, stride = 2]$ $[conv, 1 \times 16, 64] \times 2$ $[maxpool, 1 \times 8, stride = 2]$ $[conv, 1 \times 8, 128] \times 2$	2560×32 1277×32 1277×64 635×64 635×128
	global max pooling	1×128
Task-Specific Layers	$[dense] \times 2$ $\times$ 7 *parallel tasks*	128
Output	–	2

5.2.3.2 Emotion Recognition

Considering the second stage of learning (T_d), the model uses the original ECG signals and the emotion labels y_i for the classification. The weights of the convolutional blocks obtained in the self-supervised learning task optimization are frozen and used in this emotion recognition network, since the latter contains convolutional layers similar to those used in the network for learning ECG representation.

Data Pre-processing

Since Sarkar and Etemad [169] make use of four different datasets, which were collected in different environments and using different hardware, there was the need to minimize the effects of these variations, which was done through the use of three pre-processing steps. Firstly, SWELL and WESAD were downsampled to 256Hz, which, in this replication was not done. After that, the baseline wander was removed from all the datasets, by applying a high-pass IIR filter, with a pass-band frequency of 0.8Hz. Finally, it was performed a user-specific z-score normalization.

After the preprocessing is done, every signal is segmented into 10 seconds and stacked into an array. This segmentation step has no overlap, so as to avoid potential leakage of equal small segments of data between the training and testing sets.

Emotion Recognition Architecture

The proposed network for emotion classification is also made of 3 convolutional blocks, like the self-supervised learning network. Following the conv-blocks, there are fully connected layers with 512 hidden nodes. The weights of the shared convolutional layers of the signal transformation recognition network are frozen and then transferred to the emotion recognition network. On the other hand, the fully connected layers are not received from the transformation recognition network and, thus, are trained in a fully supervised manner, using the labelled dataset, shown in Figure 5.6.

In this way, these fully-connected layers are fine-tuned according to the dataset used. Sarkar and Etemad [169], used three dense layers for SWELL and WESAD, while 4 dense layers with L2 regularization (0.001) were applied for AMIGOS and DREAMER. Furthermore, concerning DREAMER, a dropout of 20% was used while in AMIGOS the dropout rate was 40%. For the emotion recognition task, only DREAMER dataset was considered during the replication. The final output was then taken from a sigmoid (binary classification) or softmax (multi-class classification) layer, having into account the number of classes considered.

Finally, similar to the signal transformation recognition network, it was also used a 10-fold cross-validation and a 90%-10% division of data in training and testing sets. Furthermore, Adam optimizer was used, with a learning rate of 0.001 and the same batch size of 128. The emotion recognition network was trained for 200 epochs, when using the frozen convolutional layers. In addition, it was also tested fully-supervised learning, without transferring the weights from the

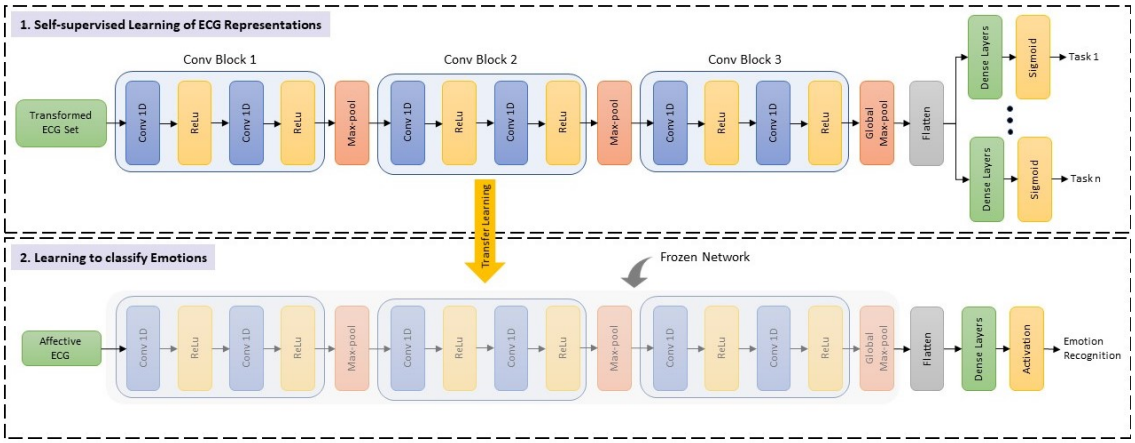


Figure 5.6: Self-supervised Learning Architecture, from [169]. First, a multi-task CNN is trained using automatically generated labels to learn ECG representations. The optimized weights are then transferred to the emotion recognition network, where the fully connected layers are trained.

self-supervised learning task, in which the network was trained for 250 epochs. However, in this replication, fully supervised was only tested to confirm if the methodology was being correctly implemented and its results will not be further analysed.

5.3 Results

5.3.1 Santamaria-Granados *et al.*'s Deep Convolutional Neural Network

5.3.1.1 Experiment 1 - Impact of Model's Complexity and 30-second Segment Overlapping

As already mentioned in the experimental settings, the first experiment conducted did not have satisfactory results, obtaining a training accuracy of 98.58% and 98.21% for arousal and valence, respectively, and only a testing accuracy of 8.05% and 17.98%, as presented in Table 5.2. This considerable overfit could be due to the model complexity, since four different convolution layers with a high number of filters were used, in addition to the 3 FCN layers, also with a great number of neurons. For this reason, the alterations described in Section 5.2.1.2 were performed, in order to decrease this complexity. However, occasional changes in the dropout value or the number of layers did not improve the results, still obtaining similar accuracies.

On the other hand, other changes were done, applying batch normalization, using Adam Optimizer and considering two separate Loss Functions. With these alterations and 30-second segments with 5-second overlaps, an accuracy of 41.59% and 36.82% for arousal and valence, respectively, was obtained, indicating that the model may need more data to adequately learn the task and avoid overfitting.

Table 5.2: Classification rates (%) for the first approach of Experiment 1

		Valence	Arousal
No Overlap	train	98.21	98.58
	valid	12.42	11.81
	test	17.98	8.05
Overlap 5sec	train	71.99	70.09
	valid	54.42	53.79
	test	36.82	41.59

5.3.1.2 Experiment 2 - Impact of the decrease in number of classes and segments sizes

By reducing the number of classes from nine to three or two, it was expected that results could be improved. However, that improvement was not verified for any of the settings considered, as presented in Table 5.3 and Table 5.4.

Using 30-second segments, for the first setting, a classification rate of 41.59% and 36.82% for arousal and valence was obtained for the 3-class problem. Considering 2 classes, an accuracy of 55.43% for arousal and 46.58% for valence was achieved. These results are only slightly above random guess results (33.3% and 50% for 3 and 2 emotion categories, respectively), which proves that the model is not recognizing the relevant ECG patterns and learning to distinguish between emotions.

It was expected that the results would be worse for the second setting due to the subject-independent division. For three classes, it was obtained a classification rate of 37.89% for arousal and 32.88% for valence, whereas in the two-class task, accuracies of 62.22% and 52.70% were achieved. Furthermore, there was noticeable overfitting in both settings, since the training accuracy was around 70% for the 3-class problem and 80% when two classes were considered. In this way, it can be concluded that the model was unable to learn in both settings.

Although the model complexity was reduced, other reasons can lead to the overfit present in both settings, like noisy signals with a lot of irrelevant information, and the small amount of data available. Furthermore, excessively long segments can lead to overfitting, since larger amount of input information can make it more difficult for the model to identify small patterns needed for the task at hand . In this way, it was considered and tested small segments of 15 and 10 seconds, as already mentioned.

For 15-second segments, the results were similar, as it can be seen in Table 5.3. In the 3-class problem, classification rates were slightly above 33% , whereas for two classes, accuracies were around 50%. The same happened for the 10-second segments, obtaining an accuracy of 37.35% and 38.53% for arousal and valence, in the 3-class problem and 47.92% and 53.63% for two classes. Concerning the second setting, results were also identical for both segments considered (see Table 5.4).

Table 5.3: Classification rates (%) for the 1st Setting of Experiment 2.

		Classes			
		2		3	
Segment size		**Valence**	**Arousal**	**Valence**	**Arousal**
10 sec	train	66.72	74.04	58.90	62.12
	valid	58.39	62.64	45.69	47.76
	test	49.14	57.03	38.53	37.35
15 sec	train	72.34	78.46	62.67	64.20
	valid	55.43	61.22	50.31	48.03
	test:	55.31	51.67	37.19	36.44
30 sec	train	80.87	84.75	71.99	70.09
	valid	63.15	69.76	54.42	53.79
	test:	46.58	55.43	36.82	41.59

Table 5.4: Classification rates (%) for the 2nd Setting of Experiment 2.

		Classes			
		2		3	
Segment size		**Valence**	**Arousal**	**Valence**	**Arousal**
10 sec	train	69.65	77.28	58.27	61.87
	valid	59.07	67.69	45.18	48.81
	test	53.63	47.92	37.90	35.54
15 sec	train	71.36	77.41	61.00	62.95
	valid	54.59	64.87	34.78	36.68
	test:	54.21	52.67	34.78	36.68
30 sec	train	79.15	83.63	66.82	68.95
	valid	61.49	66.24	53.72	50.87
	test:	52.70	62.22	32.88	37.89

5.3.1.3 Experiment 3 - Impact of random data division and no overlapping

By developing a less thorough division of the data, the idea was to achieve a better recognition performance, similar to the results obtained on the published work [167]; however, the results were identical to the previous ones. The difference in the results obtained can be associated with the databases used, since MAHNOB-HCI was used in this preliminary work, while Santamaria-Granados *et al*. [167] chose to use AMIGOS. The latter presents a total of 40 subjects, whereas MAHNOB-HCI only contains recordings from 27 subjects. Thus, AMIGOS presents a high amount of data, which is an important factor concerning deep learning tasks, as discussed in Section 5.2.1.1. Furthermore, this database also includes an experiment with long videos, which means that higher segments can be considered to train the network accurately.

Nevertheless, some results were slightly better. Considering the 3-class task for the first setting, a 41.95% accuracy was obtained for the valence recognition. On the other hand, for the 2-class problem, an accuracy of 61.47% was obtained considering the first setting, as it can be seen in Table 5.5. For the second setting, similar results were also achieved (see Table5.6).

However, as it can be understood, the same problems encountered in the previous experiments were also present in this one. Although the overfit is somewhat more reduced, it is still present.

Table 5.5: Classification rates (%) for the 1^{st} Setting of Experiment 3.

		Classes			
		2		**3**	
Segment size		**Valence**	**Arousal**	**Valence**	**Arousal**
10 sec	train	70.97	75.69	60.86	63.39
	valid	54.92	59.28	40.45	43.10
	test	52.40	61.47	41.95	37.84

Table 5.6: Classification rates (%) for the 2^{nd} Setting of Experiment 3.

		Classes			
		2		**3**	
Segment size		**Valence**	**Arousal**	**Valence**	**Arousal**
10 sec	train	72.84	76.44	61.17	68.91
	valid	51.43	60.99	43.87	42.53
	test	56.43	60.62	40.31	31.01

Concerning the general results obtained in this first replication, it can be understood that the implemented methods did not reach the desired levels of performance, which can have different causes, such as the overfit. Overfit is present in every experiment, especially in the first one. By reducing the model complexity, this overfit also decreased, proving that high levels of model complexity are sometimes the cause of worse results. However, overfit was still present in the rest of the experiments, which can be due to the use of noisy ECG signals since no preprocessing was done before feeding them to the network. These irrelevant artifacts can impair the detection of essential patterns for emotion recognition. The small amount of data can also be an important factor, taking into account that deep networks require a high amount of data available to achieve good performances. In this way, larger databases are often in need for the network to be well trained and perform accordingly.

Furthermore, regularization techniques were also applied, such as L2 regularization, in order to automatically regulate the learning rate throughout the model training. However, neither this technique nor dropout was able to eliminate overfit.

Having this said, the difference in the results obtained, compared with the ones reported by Santamaria-Granados *et al.* [167], can be due to the differences in the replication, since the methodology applied was only based and not completely similar, having differences in the length of the segments considered and even the database used. However, it is also important to underline the lack of clarity in the approach presented since many relevant details were not well specified and had to be externally decided or empirically tested.

5.3.2 Dar *et al.*'s CNN and LSTM-Based Emotion Charting Using Physiological Signals

As already mentioned, the work was replicated using only the ECG signal. For this, two random splits of each lead were considered for both the replication and Dar *et al.* [170] approach. Table 5.7 presents the results obtained in both cases. Although both DREAMER and AMIGOS were equally prepared and tried out, the results for this second database are not presented since Dar *et al.* [170] mentions that the totality of each signal is used, which difficults the choice of the number of segments to consider in the LSTM layer.

Table 5.7: Summary results for DREAMER database, compared with the original results reported by Dar *et al.* [170].

		Accuracy (%)		
Approach	**Lead**	**1st Random Split**	**2nd Random Split**	**Average**
Dar *et al.* [170]	ECGl	90.46	88.31	89.39
	ECGr	90.03	90.96	90.50
Our Results	ECGl	32.80	38.92	35.86
	ECGr	43.72	41.60	42.66

As one can understand by the results presented, the work's results were not successfully replicated, obtaining accuracies considerably lower. However, every detail explained in the methodology was carefully followed and verified. For this reason, the poor results can be due to the already mentioned lack of information regarding a variety of parameters such as padding values and, most importantly, the number of segments aggregated at each epoch by the LSTM layer. This information becomes crucial and can have a high impact on the results obtained, since the model receives and processes data at a different pace. Furthermore, the temporal relations and patterns extracted by this Long Short-Term Memory layer depend on the number of segments aggregated at each moment.

It is also relevant to mention the poor data division, with completely random 70%-30% splits of segments that may belong to the same signal. This kind of division does not ensure that there is no leakage of data between training and testing sets, which means that the model is probably trained with segments belonging to the same signal as other segments present in the test set. Thus, although the results are considerably high in Dar *et al.* [170] approach, it may not hold up when it comes to a real-life application, where new and unseen data would be tried out.

In summary, the lack of clarity in all of the steps performed during this approach, as well as the unreal data division with no concerns with real conditions, led to the conclusion that this paper, while achieving high accuracy levels for the four emotion classes considered, is not replicable nor fit to be implemented and used in everyday life.

5.3.3 Sarkar and Etemad's Self-supervised ECG Representation Learning for Emotion Recognition

As mentioned, the replication of this paper has some changes when compared with the original, mainly in the datasets considered for both tasks and the fact that cross-validation was not used. Furthermore, besides fully replicating this TensorFlow approach into PyTorch, the code available in Gitlab [183] made it possible to first try this approach by directly obtaining the extracted features from a given dataset when it went through the frozen convolutional network of the self-supervised approach. However, as it can be understood, fully replicating both the pre-training and the emotion recognition network allows a higher level of control over the methodology and opens opportunities for further additions or alterations to it.

Having this said, the following Section will present the results firstly obtained by using the features directly and secondly by replicating both the self-supervised learning and the emotion recognition networks in PyTorch. For each experiment, three random splits were considered, concerning both ECG leads, and their average and standard deviation results are further presented.

5.3.3.1 Feature-based results

By using the code available in Gitlab [183], it is possible to extract the features from the frozen convolutional layers, using the optimal weights from Sarkar and Etemad [169]. From this point on, the features are already prepared to enter the fully connected layers. Concerning this methodology, DREAMER was prepared and tested using two classes for high/low classification of both arousal and valence. Furthermore, as mentioned, both leads were tested with three random splits, computing the mean afterwards. Considering the approach replicated, the results are presented without defining the lead used and using only one random split of data.

As it can be seen in Table 5.8, the results were fairly similar to the ones presented in the literature, showing a small decrease of 6% when it comes to arousal recognition. Concerning valence, the mean accuracies obtained for both leads were around 8% lower than the one reported by Sarkar and Etemad [169]. However, it is not clear the reasons why there is this difference, since every step of data pre-processing was carefully followed. Nonetheless, the paper only describes the filter used, not mentioning, for example, its order. Other aspects concerning the data preparation could be considered since some details were unspecified, such as if the complete signals are used, or only small and sectioned fractions of them. Furthermore, since the fully-connected layers had to be trained from scratch, their initialization was probably different from the one of Sarkar and Etemad [169], which could also justify the difference in the results obtained. However, it can be considered that the results are close enough to conclude that the replication was successful when extracting the features and then use them to train the fully connected layers to recognize emotion.

5.3.3.2 Signal Transformation Recognition

In order to have a quality pre-training for the emotion recognition task, the signal transformation recognition was performed, following the methodology described in Section 5.2.3.1. However,

Table 5.8: Summary results for DREAMER database considering high/low arousal and high/low valence, using the extracted features directly, compared with the original results reported by Sarkar and Etemad [169].

		Accuracy (%)			
		Arousal		**Valence**	
Approach	**Lead**	**Average**	**Standard Deviation**	**Average**	**Standard Deviation**
cite	n.d	85.9	-	85.0	-
Our results	ECGr	80.55	0.99	77.49	1.23
	ECGl	79.63	0.84	77.05	1.07

as mentioned, the datasets used for this pretraining were DREAMER and AMIGOS, excluding WESAD and SWELL. This choice is based on the fact that the chosen datasets were already available and ready to be used due to their applications in other approaches, such as [167] and [170]. Furthermore, according to Sarkar and Etemad [169] the performance is not strongly affected by the reduction of datasets used, considering that each available signal suffers 6 different transformations, which in the end leads to a six-fold dataset increase. In addition, Sarkar and Etemad [169] performed this transformation recognition task using each dataset individually and all of them combined, which proved not to exist relevant accuracy differences.

In Table 5.9 the accuracy results for the recognition of each transformation are presented and compared with the paper replicated, demonstrating that the pretraining was successful.

Furthermore, in Figure 5.7 it is possible to analyse the train and validation loss curves through epochs, indicating a good training that resulted in high accuracies analysed in Table 5.9.

Table 5.9: **Summary of the results for the Signal Transformation Recognition task**, using DREAMER and AMIGOS datasets, compared with the results obtained by Sarkar and Etemad [169] when combining all four datasets.

Transformation	**Accuracy (%)**	
	All datasets combined in Sarkar and Etemad [169]	**DREAMER + AMIGOS (our results)**
Original	98.00	97.73
Noise Addition	99.50	99.53
Scaling	98.20	98.03
Temporal Inversion	99.80	99.87
Negation	99.80	99.73
Permutation	99.80	99.86
Time-warping	99.70	99.93
Average	**99.20**	**99.24**

5.3.3.3 Emotion Recognition

Using the weights of the convolutional layers obtained in the signal transformation recognition task, only the fully connected layers of the emotion recognition network had to be trained. In Figure 5.8, the results obtained are compared to the ones by Sarkar and Etemad [169], presenting a decrease in performance of nearly 14% concerning arousal and 17% for valence classification.

Furthermore, when comparing with the results obtained when using the extracted features directly, the results were also inferior, presenting a decrease close to 5% for arousal and 8% for valence.

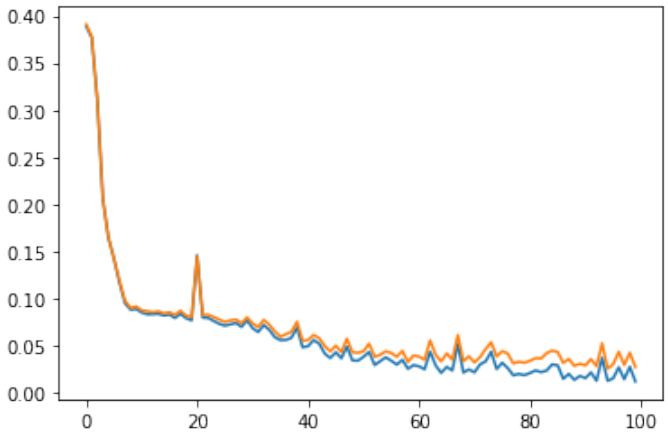


Figure 5.7: Train and validation loss curve during pretraining.

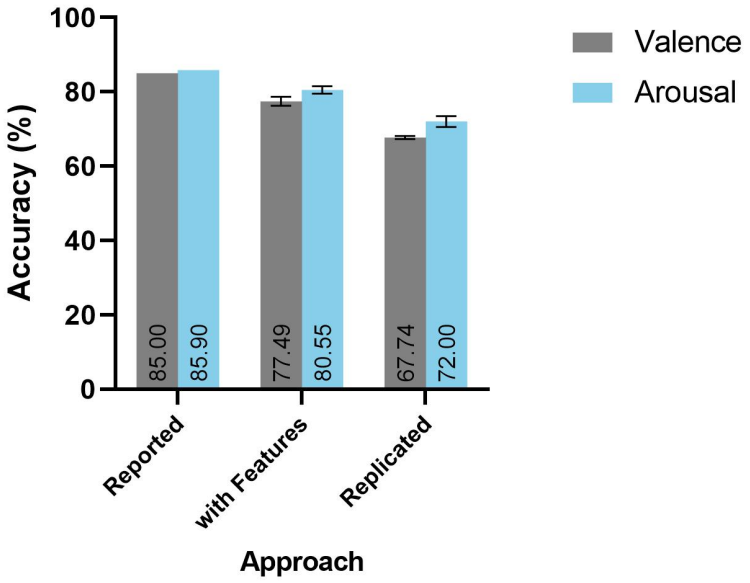


Figure 5.8: Emotion Recognition results considering the ones presented by Sarkar and Etemad [169] and the two methodologies used. For those methodologies, the accuracy considered is from lead ECGr.

The possible reasons for the accuracy decrease verified between the use of extracted features and the ones reported by Sarkar and Etemad [169] were already discussed in Section 5.3.3.1 and are also valid to justify some of the variance between the results from Sarkar and Etemad [169] and the ones obtained with the replication. Although the pre-training results were highly satisfactory, as analyzed in Section 5.3.3.2, the convolutional weights obtained might not be exactly equal to the ones from Sarkar and Etemad [169], since this training is randomly initialized and may result

in different optimal weights. This idea is supported considering that the results obtained with the directly extracted features, using Sarkar and Etemad [169] convolutional weights, were higher than the ones obtained with the full pre-training replication in PyTorch.

Furthermore, taking into account that there were used two different deep learning libraries, TensorFlow and PyTorch, respectively, there are some intern and specific differences of processing and work mode that could also play an important role in the distinct results obtained.

In addition, as mentioned, the data pre-processing developed by Sarkar and Etemad [169] was not completely clear, namely in aspects such as filtering and signals used to train the model. Sarkar and Etemad [169] state that every DREAMER ECG signal has 60 seconds which does not correspond to the dataset reality, since each signal has a different length considering the emotion elicitation video used. Thus, there is some doubt if Sarkar and Etemad [169] pre-processed or previously selected the signals to use. Furthermore, taking into account the incomplete code presented in Gitlab [183], the code to access signals from the different datasets does not seem to agree with the way datasets are saved, which also indicates some previous organization and possible selection of the data to use. However, during this replication, each signal was used in its totality and there were no alterations applied to the datasets in question.

5.4 Summary and Conclusions

This chapter focused on presenting and analysing the state-of-the-art approaches that were replicated, serving as a basis for further work. As it can be concluded, emotion recognition literature is not as clear as it could be expected. Although the prior art results are high enough to indicate potential and important developments in this field, a more thorough analysis proves that the proposed methodologies are not completely clear or reproducible, since a variety of details are kept unexplained, even when code is made available. Furthermore, the data settings used are also not completely fair or in conformity with the real conditions that would be encountered if those methodologies would have a real-life application.

In most of these approaches, namely the three replicated, data is divided into smaller segments and randomly split into train/test sets. This division method not only allows for signals from the same subjects to be in both sets but also segments from the same signal as well. Although the segments are not exactly the same, they convey identical patterns and characteristics that can be seen as a contamination of training and testing sets. In fact, in Chapter 6, this methodology is used by considering both different signals and subjects for training and testing sets, proving the impact that this fairer division has on results, and cross-database experiments. In conclusion, state-of-the-art shows most approaches are not yet ready to be applied on a daily basis in a functional and robust way.

Due to all of the reasons mentioned above, the replications presented were not always successful. However, concerning the last approach, from Sarkar and Etemad [169], the results obtained were more consistent and closer to the ones reported by the authors. Although there are some parameters and methods still poorly defined and data division is also biased, Sarkar and Etemad

[169] presented a more detailed description of the work developed, in addition to providing the code in [183]. Furthermore, Sarkar and Etemad [169] methodology uses a pre-training of the convolutional layers, with large amount of data, by considering signal transformations and their recognition. In addition, this approach trained the convolutional and fully-connected blocks of its network separately, which allowed for a more controlled training, highly benefiting emotion recognition. Thus, this approach was more reliable, presented a promising methodology and its results were in accordance with the quality of the method.

Furthermore, although the replications developed tried to be the most similar as possible to the methods proposed, there were some differences concerning libraries or programs used, since, for example, Sarkar and Etemad [169] developed the approach in TensorFlow and Dar *et al*. [170] even used Matlab instead of Python. These language differences may also have some impact on results, as previously mentioned.

In summary, the results obtained during these replications were able to identify some common and recurrent problems in the literature, understanding in a more complete and detailed manner what already exists and what is lacking in the approaches developed so far. Thus, after these replications, in Chapters 6 and 7, more trustworthy data division settings were considered as well as other deep learning architectures that could result in better and more reliable results.

Chapter 6

Self Supervised Learning Improvement for Emotion Recognition

6.1 Introduction

Considering the replicated work of Sarkar and Etemad [169], presented and analysed in Chapter 5, a self-supervised learning technique was developed. However, as it was discussed, the results obtained were not quite high as the ones presented in the literature, which could be due to the already mentioned difficulty of replicating some papers and the fact that the original methodology is developed in TensorFlow and then replicated to PyTorch.

However, this self-supervised technique presented itself as a good starting point for the final architecture proposed in this thesis, since it offered room for improvement. Thus, the idea was to come up with a stronger self-supervised task that could result in more robust convolutional layers for the emotion recognition problem. Some possibilities were considered, such as increasing the number of signal transformation tasks or even mix them, which means, allowing for the same signal to suffer more than one transformation and, thus, having more than one label (for example, an inverted and permuted signal). Although these alterations would, eventually, increase the difficulty of the problem and probably result in less high performances in signal transformation recognition, it would also improve the robustness of the network and, thus, the performance of the emotion recognition task.

In this way, the following sections present the alterations done to the self-supervised learning proposed by Sarkar and Etemad [169], as well as the results obtained. From all of the changes tried, the one that obtained better results was then selected for further improvements and additions, explained in Chapter 7. Furthermore, other validation tests were performed, such as cross-database, signal-independent and subject-independent settings, to understand how well the network would perform in these cases.

No other modifications were done regarding data pre-processing or network architectures, in both signal transformation recognition and emotion recognition tasks, keeping the same methods proposed by Sarkar and Etemad [169] and fully explained in Chapter 5.

6.2 Methodology

6.2.1 Signal Transformation Recognition

As explained in Chapter 5, the signal transformation recognition task was performed in order to optimize the convolutional weights, then used in emotion recognition. However, each original signal suffered 6 isolated transformations, resulting in 6 different signals. The same signal only had a label from 0 to 6, and never more than one transformation applied. Nonetheless, the addition of different transformations to the same signal could increase the network's ability to distinguish transformations from more complex signals.

Having this in mind, two main alterations were considered, namely applying more than one transformation at once and adding new and different signal transformations to the six already used.

6.2.1.1 Mixing different transformations

Since each signal only had one single transformation, the first addition considered was to apply one more transformation (signals with exactly two transformations, excluding the original signals which had none). Furthermore, to increase the unpredictability, signals could also have up to two transformations, which means that signals could have one or two transformations.

The same technique was used for three transformations, imposing that all of the signals that were not original would have three different transformations applied and, after that, up to three transformations (1, 2 or 3 transformations per signal). This choice was completely randomized, making the set more unpredictable and, thus, more difficult to generalize.

In addition, it was tested up to four and five transformations, however the results obtained seemed to indicate that more than three stopped to be beneficial in terms of emotion recognition performance, which led to the belief that imposing four or five transformations per signal would be too confusing for the model or it would result in a train too specific for signal transformation recognition, leading to less relevant convolutional weights for the emotion recognition task.

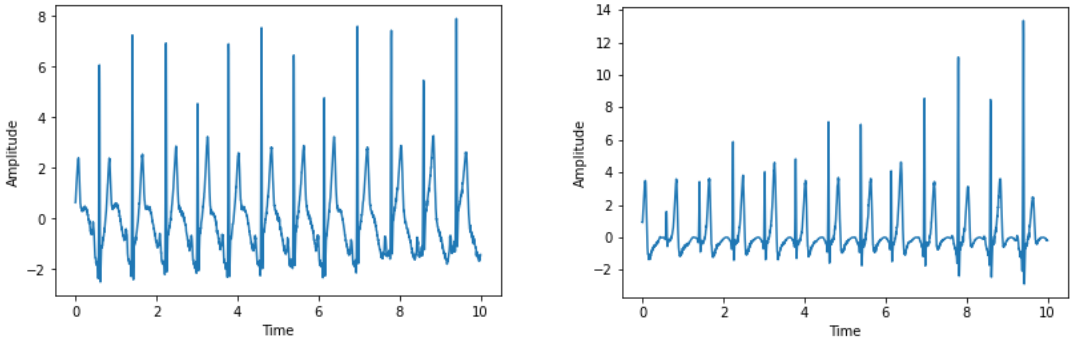
6.2.1.2 Addition of new transformations

Besides applying more than one transformation per signal, it became relevant to test other transformations that could also be important for the network to recognize relevant ECG patterns. In this way, two other transformations were considered, having into account the ones pointed out as relevant by Ribeiro Pinto *et al*. [184]:

- *Baseline Wander:* Simulates a periodic undulation on the signal, through the addition of a sinusoidal wave with frequency near 1 Hz. In this specific situation, it was used a sine wave with an amplitude of 1.

- *Magnitude Warping:* Rescales the signal in a non-uniform way, using a sinusoidal wave instead of a fixed factor, in such a way that some parts of the signal have their amplitude expanded and others shrunk. The sinusoidal wave used was the same for the *Baseline Wander* transformation, with an offset of 1.

These transformations were applied to the original signal, resulting in the signals seen in Figure 6.1. In Chapter 5, Figure 5.5, the other 6 transformations are also presented, giving a visual idea of the variations each one caused in the original ECG signal.



(a) Baseline Wander (b) Magnitude Warping

Figure 6.1: Addition of two signal transformations ((a) and (b)) for improving the pre-training.

Initially, these eight transformations were used similarly to the first six, by applying only one transformation per signal. Furthermore, these eight transformations were also combined with other pretraining changes presented in Section 6.2.1.1, namely the one that led to the best pretraining results by mixing more than one transformation per signal, which was the methodology of adding a maximum of three transformations per signal in a random way.

6.2.2 Emotion Recognition

In order to understand if the alterations done in the pretraining step were useful, the obtained convolutional weights were also frozen and transferred to the emotion recognition network. Thus, for each self-supervised learning modification previously described, it was tested its impact on emotion recognition, since this was the task to optimize.

As mentioned, data pre-processing, network architecture and all the hyperparameters were kept the same, which means that data was segmented into 10-second segments and randomly divided into train and test sets, in a 90%-10% proportion, respectively, considering, as usual, three random splits per lead to obtain more stable and reliable results. Although there is no overlap between segments, to avoid data leakage between the two settings, it is interesting to consider more challenging and real data divisions. In this way, using the optimized self-supervised pretraining, emotion recognition was also evaluated using signal and subject-independent settings, which forces training and testing sets to have signals from different subjects. Furthermore, cross-database was also applied, ensuring that a different database was used for testing, which resulted

in different signal acquisition and subjects for the test set. It is important that a method can generalize well in both settings, ensuring its application on a daily basis with possible different signals and subjects.

6.2.2.1 Signal-Independent Settings

Taking into account the approaches available in the literature, it can be understood that, usually, signal segments are used as inputs. In addition, those segments are randomly shuffled before being divided into train and test segments which results in information of the same signal to be present in both sets. In this way, it became relevant to test a division that would assure a more strict train/test split, with different signals in each set. This experiment proves to be relevant to understand if the network developed is signal-dependent or if it is able to predict emotion for completely unknown and unseen signals.

For this, data was prepared considering three signal-independent splits for each ECG lead, with the already used division of 90%-10% for train and tests sets. In this way, signals were firstly assigned to their set and then spliced into 10-second segments, avoiding signal leakage between sets. Each segment was pre-processed following the pre-processing steps already described in Chapter 5, Section 5.2.3.2.

6.2.2.2 Subject-Independent Settings

Subject-independence remains a challenge in most machine and deep learning problems using personal data. Although, in literature, there are a lot of approaches that do not even consider these type of more demanding sets, as seen in Chapter 4, this failure can have a high impact when it comes to real applications of these methods. In this way, subject-independent sets were prepared in order to test emotion recognition using the convolutional weights of the best pretraining obtained (up to 3 transformations), so as to understand the impact of more demanding and fair data division on the methodology developed. For this, considering 23 subjects from DREAMER database, 10% (corresponding, approximately, to 3 subjects) were randomly selected for the test set and the other 90% (20 subjects) for training. Concerning train/validation division, the validation set was randomly obtained from the train set, corresponding to 10% of training data, with no concern with subjects. For each ECG lead, as usual during this thesis, there were considered three different random splits.

6.2.2.3 Cross-Database Experiments

Cross-database consists of a technique in which a given database is used to train a model and a different one is used for testing, to evaluate how well a given model can generalize not only between different subjects but also between signals with different acquisition settings. Real-life applications need to be robust and perform well in most of the scenarios, thus, in this study, both MAHNOB-HCI and AMIGOS were used in different test settings. These cross-database experiments were developed using the optimized trained model and the two datasets were equally

prepared, following the pre-processing already explained in Chapter 5. In this case, instead of using only a percentage of the datasets, they were fully considered in the test set, using both ECG leads.

6.3 Results

6.3.1 Signal Transformation Recognition

As already mentioned, different techniques were used in order to improve the signal transformation recognition task. The possibility of obtaining even more optimized convolutional weights, able to extract relevant ECG features could result in improved performances for the emotion recognition task, when compared with the ones presented in Section 5.3.3.3. For this, two main variations were tried, namely (1) applying more than one transformation per signal and (2) adding signal transformations. This Section presents the results concerning the variations applied to the pre-training task.

6.3.1.1 Mixing different transformations

Concerning the addition of more than one transformation per signal, different methods were tried and explained in Section 6.2.1.1. The mean accuracy of all the transformations is presented in Table 6.1.

Table 6.1: Summary of the average accuracy results for the Signal Transformation Recognition task, when varying the number of transformations per signal.

	Accuracy (%)	
Nr Transformations per Signal	**Average**	**Standard Deviation**
Sarkar and Etemad [169]	99.26	0.74
Exact 2	98.77	0.77
Up to 2	98.52	2.13
Exact 3	98.44	1.14
Up to 3	97.29	4.38
Up to 4	97.05	4.40
Up to 5	92.07	9.65

As expected, the performance of the model decreases with the addition of simultaneous transformations per same signal. Furthermore, when all signals have the same amount of transformations, the network is able to perform better than when the number of transformations is random and changes from signal to signal, since the unpredictability of the data fed to the network increases in the second scenario. In this way, by analysing the results obtained, it can be concluded that the transformations done to the proposed pre-training by Sarkar and Etemad [169], difficult the signal transformation recognition, which could translate into more robust and optimal convolutional weights, able to extract features even more relevant for emotion recognition, thus improving this task.

6.3.1.2 Addition of new transformations

Besides adding more than one transformation to each signal, considering other signal transformations beyond the 6 proposed by Sarkar and Etemad [169] can also be interesting and highlight other specific features or patterns of the ECG signal. Thus, as explained in Section 6.2.1.2, baseline wander and magnitude warping transformations were applied to original ECG signals, increasing the number of transformations per signal and the overall size of the transformation dataset used in the signal transformation task.

Firstly, the pretraining was done following the same method of Sarkar and Etemad [169], with the two additional transformations applied to every original signal. Having into account that the higher performance for emotion recognition was obtained by adding up to three simultaneous transformations per signal, these two transformations were also considered. Table 6.2 presents the results of Signal Transformation Recognition for both methods mentioned, showing an average accuracy of 96.38 % for the recognition of nine types of signal (original and eight transformations) and 96.49% with the addition of up to three transformations per signal.

As expected, the increase in the number of different signal transformations, combined with the fact that each signal could have up to three distinct transformations, led to one of the lowest accuracies obtained, when compared with the results presented in Table 6.1. However, it can be noticed that considering up to five transformations per signal resulted in an even lower performance due to the demanding task of recognizing such a high number of transformations applied to the same signal.

Although other combinations could be tried using eight transformations, the results would probably be alike and not significantly alter the performance of the emotion recognition task. In this way, no other transformations were applied to the pretraining.

Table 6.2: Signal Transformation Recognition results considering 8 signal transformations.

Transformation	**Accuracy (%)**	
	8 transformations	**8 transformations + up 3**
Original	98.03	97.95
Noise Addition	97.99	98.16
Scaling	85.01	84.68
Temporal Inversion	93.92	94.09
Negation	94.81	95.49
Permutation	99.55	99.59
Time-warping	98.38	98.65
Baseline Wander	99.78	99.86
Magnitude-warping	99.95	99.96
Average	**96.38**	**96.49**
Std. Dev	**4.26**	**4.58**

Thus, considering all the different alterations done into the pretraining, concerning the number of transformations used and applied to the same signal, it becomes important to understand

the impact that on emotion recognition. Having this said, the following section presents the emotion classification accuracies for two classes of arousal and valence, for each alteration done and described in the pretraining data.

6.3.2 Emotion Recognition

As already mentioned, the variations done during the self-supervised learning task lead to different convolutional weights that could be more or less suited for emotion recognition. Thus, in order to validate and understand what alterations in the pretraining were, in fact, beneficial for emotion recognition, the correspondent convolutional weights of each pretraining were frozen and transferred into the emotion recognition network. Results for the right lead are presented in Figure 6.2 and for the left lead in Figure 6.3.

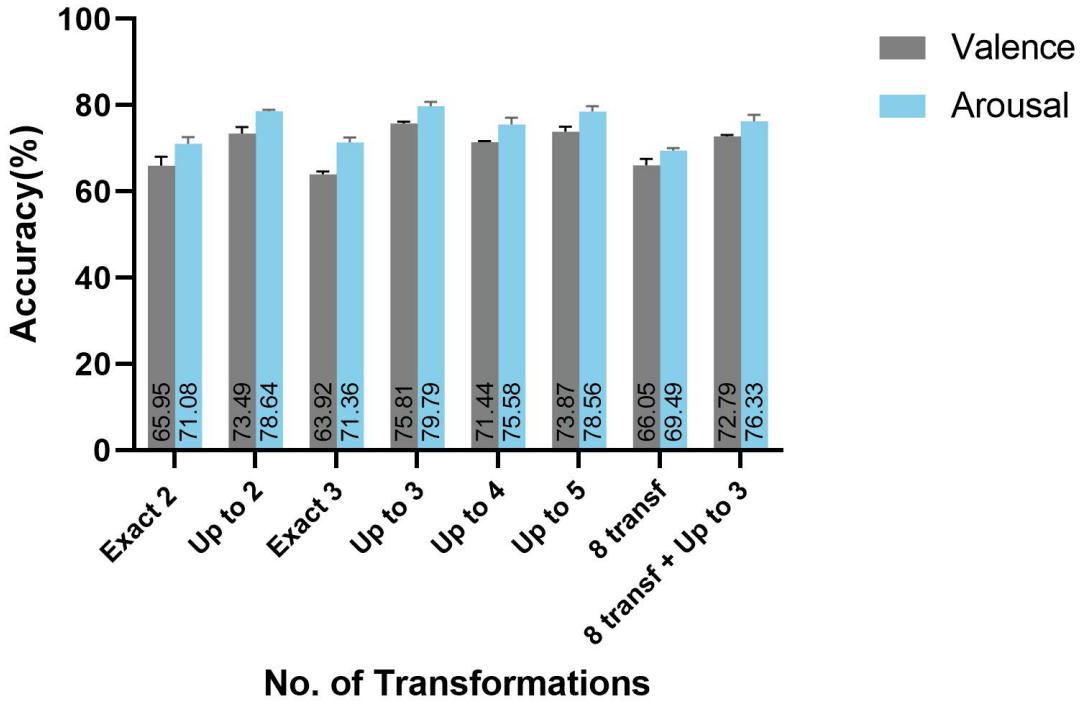


Figure 6.2: Emotion Recognition results for ECG **right lead** corresponding to each pre-training developed.

Through the analysis of Figure 6.2, it can be concluded that the best pre-training obtained was by randomly applying a maximum of three different transformations per signal, obtaining an average accuracy of 79.79% for arousal recognition and 75.81% concerning valence. Regarding left lead (Figure 6.3), the highest performance was also obtained by this pretraining, with similar accuracies of 78.05% for arousal and 75.19% for valence.

Furthermore, the pre-training that resulted only from the addition of two transformations, did not show any improvement when compared with the results obtained with six transformations,

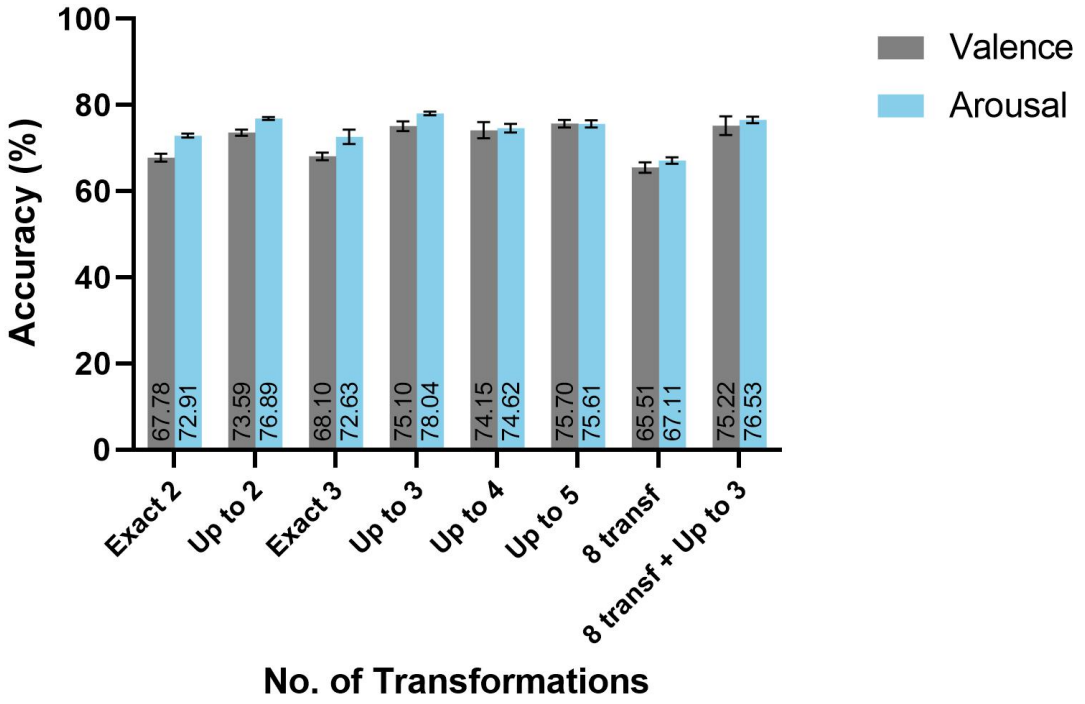


Figure 6.3: Emotion Recognition results for ECG **left lead** corresponding to each pre-training developed.

presented in Figure 5.8, which shows that the simultaneous application of more than one transformation per signal is more effective than adding the number of transformations while keeping only one transformation per signal.

Furthermore, the results obtained for each lead allow us to conclude that ensuring the same number of transformations per signal in the pre-training, namely two and three, led to poorly and less robust convolutional weights due to a lower signal variety, which results in lower emotion recognition performances. However, the pretraining with more random data, made of signals with a different number of transformations, showed to be more effective. The maximum usage of four and five transformations also resulted in higher accuracies compared with methodologies with a fixed number of transformations per signal. In this way, considering the results presented in Table 6.1, it is possible to establish a relationship between signal transformation recognition accuracies and emotion recognition ones, since the pre-training that achieved the higher performances in transformation recognition were also the ones that resulted in less relevant convolutional weights for emotion recognition, which demonstrates that increasing the difficulty of this self-supervised learning resulted into better and more efficient pretrainings for emotion classification. Finally, regarding the addition of two transformations, the results decreased in comparison with the best pretraining, although, in regard to the ECG left lead results, valence performed slightly better. This can indicate that this pretraining was also successful, however, signals probably became a little too demanding and led to a more specific training for signal transformation recognition,

translating into lesser appropriate weights for emotion recognition tasks.

By using up to three transformations per signal as the pre-training method, it was possible to achieve the results reported when using the features directly extracted from Sarkar and Etemad [169] convolutional layers and optimized weights (see Figure 5.8), which showed that the methods applied to improve pre-training were, in fact, successful.

Concerning this pretraining, the third random split for the right ECG lead was the one able to achieve the highest results, with an accuracy of 80.79% for arousal and 76.25% for valence. Thus, it is considered the optimized model and its convolutional and fully-connected weights are then further used in different experiments in Chapter 7 and in Section 6.3.2.3, for cross-database results. Nonetheless, the slightly higher performance of this random split can also be related to a possible easier test set, filled with cleaner or less challenged segments that allowed a better model performance.

6.3.2.1 Signal-Independent Settings

Considering more strict data division experiments, train and test sets were prepared in a signal-independent way, meaning that there were no segments belonging to the same signal in both test and train sets. As already explained, three signal-independent settings were prepared for both ECG leads and the average and standard deviation results are available in Table 6.3.

Table 6.3: Emotion Recognition results (%) for signal-independent settings.

	Lead			
	ECGr		**ECGl**	
	Acc (%)	**Std. Dev**	**Acc (%)**	**Std. Dev**
Arousal	54.89	4.32	55.83	2.98
Valence	58.32	4.13	50.47	1.39

The model performance for signal-independent settings highly decreased, proving its signal dependency, for both leads. Although for the right ECG lead, valence presented a mean accuracy 4% higher than arousal and, for the left lead, arousal was 5% more accurately predicted, all of the results obtained were low and these differences can just be due to the data division used. Since only three divisions were considered and the network has proved to be highly dependent on data splits, it can be concluded that results are overall low and emotion is equally badly recognized for both leads and levels of arousal and valence.

As it can be understood, this problem is mainly associated with the small emotion databases available, which do not allow for the model to be properly trained, by providing a high amount and variety of data. Due to this low variability, signals are not well generalized and the network learns too signal-specific information, performing badly when there are no segments of a given signal in the training set. In this way, there is the need for larger databases to be built so that the amount of data available for this kind of research increases, since no approaches can be real-life applied without assuring model robustness as well as signal and subject independence.

6.3.2.2 Subject-Independent Settings

As explained in Section 6.2.2.2, three subject-independent sets were prepared for each lead and then used for emotion recognition. In Table 6.4 the results obtained are presented, considering the average accuracy and the standard deviation obtained for each lead.

Table 6.4: Emotion Recognition results (%) for subject-independent settings.

	Lead			
	ECGr		**ECGl**	
	Accuracy	**Std. Dev.**	**Accuracy**	**Std. Dev.**
Arousal	54.24	2.48	56.72	2.42
Valence	54.33	6.13	56.33	0.75

The average results for both leads are quite similar between arousal and valence, around 54% for ECG right lead and 56% for the left lead. The model performance with these settings is considerably lower when compared with random segment divisions, using the same convolutional weights from the best pre-training. In Section 6.3.2, emotion recognition is around 79/80% accurate concerning arousal levels and 75/76% for valence classification, which proves the difficulty of dealing with subject-independent settings.

When the model has not yet known any signal from a given subject, it becomes highly difficult to recognize emotions, pointing the already mentioned subject dependence related with emotion and ECG signals, analyzed in detail in the Chapter 2, Section 2.2.2.

This problem, present in almost all machine and deep learning tasks, is still a major barrier for this kind of approaches to be feasible and usable in real conditions. For that reason, most state-of-the-art results are not completely trustworthy and only present more positive and somewhat misleading results regarding the performance and robustness of the developed methodologies.

In this way, it becomes highly important the development of specific techniques or methods to reduce subject dependence and obtain more robust approaches. However, this problem is emphasized and even more difficult to solve due to the small databases available, with scarce data to train a robust model able to correctly generalize for unknown subjects or different types of signals, as already mentioned.

6.3.2.3 Cross-Database Experiments

Finally, after considering signal and subject independence in the sets developed, cross-database experiments were conducted, ensuring that even the acquisition settings of signal collection were different. For a model to perform well in cross-database experiments, it has to be extremely robust and well trained, which usually depends on a high amount of data so that the architecture is prepared to deal with a large variety of signals. However, having into account the already presented results for signal and subject independence, which were considerably lower and showed a bad model performance, one could expect that these cross-database experiments should not lead to high performances, due to the increased difficulty of the sets when compared with the other two

previously mentioned. Using a different database will simultaneously ensure different signals, subjects, and acquisition settings. Table 6.5, presents the results for these experiments for both AMIGOS and MAHNOB-HCI, proving how demanding this task is.

Table 6.5: Cross-Database results for AMIGOS and MAHNOB-HCI using the optimized trained model.

		Accuracy (%)	
Database	**Lead**	**Arousal**	**Valence**
MAHNOB - HCI	ECGr	55.51	56.9
	ECGl	61.47	59.1
AMIGOS	ECGr	46.39	53.63
	ECGl	48.63	53.21

As it can be understood, both MAHNOB-HCI and AMIGOS obtained low accuracy results for both leads. However, using ECG signals from AMIGOS, the results were lower than with MAHNOB-HCI, which could indicate that AMIGOS may have more difficult or noisy signals for the optimized network to recognize emotions from.

In conclusion, the results show that the model is not sufficiently robust to perform well in completely different databases, indicating that more improvements are needed to achieve robust approaches that may generalize well for unseen data. For this to be achieved, larger datasets are needed, and specific techniques to prevent both signal and subject dependency, as mentioned in Sections 6.3.2.1 and 6.3.2.2.

6.4 Summary and Conclusions

During this chapter, the main goal was to improve self-supervised learning to obtain more robust convolutional weights for the emotion recognition network that could, subsequently, improve its performance. In this way, it can be considered that the goal was completely achieved, resulting in higher accuracies of both arousal and valence recognition, although never reaching the accuracies reported by Sarkar and Etemad [169].

Furthermore, this chapter also focused on experimenting and analysing more demanding data divisions, since these are not usually considered or evaluated in the literature approaches analysed in Chapter 4. Thus, three different sets were tested, namely signal-independent, subject-independent and cross-database. Having into account the results presented for these data divisions, the network proved to be not only subject but also signal-dependent. This problem may be due to the development of shallow networks, that could be in need of more complex architectures, however, the main reason is associated with the lack of signals and signal diversity, since the available databases are reduced. Thus, besides these databases having a reduced number of subjects, which contributes to subject-dependency, the number of signals per subject also seems insufficient to learn robust representations of emotion, which leads to poor performances when signals not present in the train set are afterwards used in test sets. Since the network is unable

to perform well in signal and subject-independent settings, it becomes clear that cross-database could also not achieve high accuracy results, since it is even more demanding.

Having this said, it can be concluded that there is still a long way to go when it comes to emotion recognition using artificial intelligence techniques, indicating that it remains a rather new and unadvanced field, with room for improvement. Therefore, a first important step towards improving it and develop better methodologies would be developing larger and more complete databases, where these types of more realistic data division can be performed and still assuring sufficient training data for robust learning.

Nonetheless, in Chapter 7, different aggregation techniques were added to the already developed network, to improve the currently satisfying results already achieved with the improvement of the self-supervised learning approach. The main goal was to predict emotion levels per signal and not for each segment individually, aggregating the segments belonging to the same signal through architectures such as MLPs and LSTMs, to extract more relevant signal features and obtaining better emotion recognition performances. Furthermore, the same demanding data divisions were further tested so as to understand if the methods added could lead to some improvement in these settings.

Chapter 7

Multiple Instance Learning

7.1 Introduction

Some of the most commonly used and available datasets for emotion recognition, analysed in Chapter 3, are only sparsely labelled, either by external raters or through self-reporting. Although subjects usually have a more clear idea of the emotion felt, it is not possible to continuously interrupt people to self-assess their emotion. In this way, emotion is not continuously labelled and self-assessment is usually done at the end of the emotion elicitation.

Having this in mind, although each ECG signal is divided into 10-second segments in the approach presented in Chapters 5 and 6, the emotion labels available in all the datasets used correspond to the self-assessment of each subject in relation to the entire ECG signal, since modelling and predicting emotions over time and obtaining continuous data labelling is quite costly and not always possible. Thus, each signal has only one correspondent emotion annotation and by using segments of 10 seconds, the same label is considered for all of the segments of a specific signal. This label generalization can, sometimes, hinder the performance and success of emotion recognition since the label is related to the entire signal and its variations, and not the segments considered. Labels represent the emotion felt in a given time window, being unable to consider continuous subtle affective changes that may occur over time and during the same emotion elicitation experiment. For example, if a subject is watching a video that is supposed to elicit happiness and that feeling is only felt at the beginning or in the end, the rest of the signal may not always convey this emotion. Furthermore, a subject can also be distracted during his/her elicitation, which can result in some useless signal segments collected during the experiment.

In this way, it became interesting to consider different aggregation techniques that would allow to predict a label for the entire signal and not for every 10-second segments.

Multiple Instance Learning (MIL) offers an improved solution for learning weakly supervised settings, considering the temporal ambiguity of the affective states. In these types of settings, each label only specifies the presence or absence of a specific affective state within a given period, not specifying or localising the exact part where it is expressed in the signal. Current MIL methods

have been used in different fields such as bioinformatics [185], medical image analysis [186], text processing [187] and emotion recognition [188].

MIL consists of a specific type of supervised learning where different segments or instances are grouped into sets, called bags, and labels are only used at the bag level, and not for each individual instance [189]. Usually, MIL has been used for binary classification problems, where if there is at least one positive instance in a bag, it is labelled positive and if it contains only negative instances it will have a negative label.

In this specific case, the idea is to group all the 10-second segments of a given signal in the same bag, obtain only one prediction per signal and compare it to the self-reported label in the dataset, instead of replicating the same label for each segment. The way these segments are grouped together in a bag can vary a lot and some of them were tried out, such as more simple and heuristic methods like mean, median and maximum, and others more complex, namely Multilayer Perceptrons and Long Short-Term Memory.

7.2 Methodology

Following the already developed and discussed method in Chapter 6, the aggregation techniques were explored in order to group the segments belonging to the same signal after the fully-connected layers. Thus, maintaining the architecture already developed, the best weights of both convolutional and fully-connected blocks, from the optimized model obtained in Chapter 6, were transferred and directly used, only training the aggregation segment of the network. For this reason, all signals used during this Chapter are right ECG lead signals since the best weights were obtained with data from the same lead. Nonetheless, after applying these optimized weights and developing experiments with different aggregation techniques, other types of data division were considered, such as signal and subject-independent settings, where only the pre-training convolutional weights were frozen and directly applied, meaning that both the fully-connected and aggregation structures were trained considering each data division. For this, the segments corresponding to the signals of each set were previously used to train the fully connected layers instead of always using the optimized weights. This setup assured that, for every fold, fully-connected layers had been trained with the correspondent train data, instead of using the same fully-connected weights for all folds. For these cases, only the two best aggregation methods were tested using these different data divisions, to analyse if the aggregation methods could somehow improve the model performances when more realistic data divisions were applied, contrasting with the low results obtained in Chapter 6 concerning the same data splits.

7.2.1 Data Preparation and Pre-processing

In order to correctly group the segments, only considering the ones belonging to the same signal, data preparation suffered some changes. Instead of immediately splitting the signals into segments, those were only pre-processed following the same steps of Sarkar and Etemad [169], namely: high IIR filter to remove baseline wander and user-specific z-score normalisation.

Thus, signals were entirely received by the network and split into 10-second segments before entering the convolutional layers. For this reason, batch size was equal to 1, wherefore the network received each signal at a time. However, since each signal has a different size, considering the video seen when it was recorded, the number of segments per signal also differed. There are significant differences, considering that the duration of the videos used in DREAMER varied from 65 seconds (resulting in 6 segments) to 393 seconds (resulting in 39 segments) [9] . Nonetheless, although MIL techniques can be used with a different number of instances per a small number could mean not enough relevant data for the prediction to be accurate and too many segments could be too confusing for the network.

Having this said, it was fixed a number of 60 segments per signal, assuring a significant amount of data per bag. For this, signals were segmented with the needed strides to result in 60 equally spaced instances. After going through the network developed in Chapter 6, with the optimum weights presented in Section 6.3.2, different aggregation forms were tested and are further described.

7.2.2 Aggregation Methods

7.2.2.1 Heuristic Methods

Since the output of the fully-connected layers are logits, which correspond to raw scores output by the last layer of a neural network before activation takes place, they already convey information of the probability of belonging to a given class. Logits are then fed to a softmax which outputs probabilities of belonging to each class, then converted to the predicted class of each segment input. In this way, simple methods of aggregation can be used and logically result in good predictions, considering that each instance is usually well predicted by the model presented in Chapter 6 (see Figure 7.1).

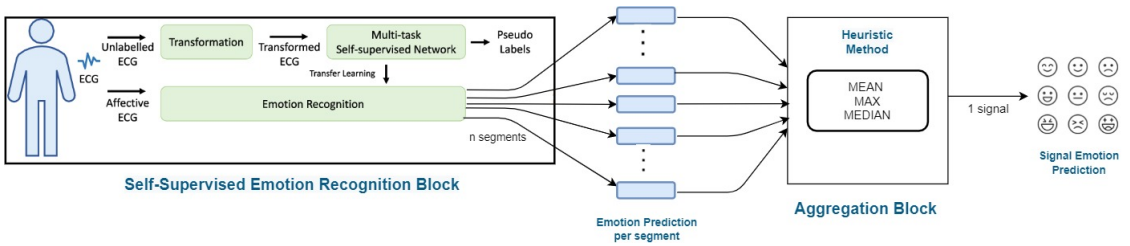


Figure 7.1: Emotion Recognition architecture using heuristic aggregation methods.

Thus, the first heuristic method tried was mean, calculating the mean of all logits of each bag. After, median was tried and, finally, the maximum was also applied. The maximum aggregation method chose the instance that offered a higher level of certainty about the class to which it belonged, as opposed to the mean which calculated the bag's average logit value and the median, which computed the median value for the same bag.

These heuristic methods were the first to be tried since they were simple and gave an initial idea concerning the possible performance of applying Multiple Instance Learning to the already

developed method. However, more complex ways of aggregating these instances could be able to ignore some of the segments and give more attention to others, with a higher relevance for the emotion recognition task. Unlike mean or median, instances are learnt in a more intelligent way, and not all instances have the same weight on decision-making for the final result. Having this said, the following sections present other aggregation techniques with more layers and complexity.

7.2.2.2 Multilayer Perceptron Methods

A Multilayer Perceptron (MLP) is a class of feedforward neural network, consisting of at least three layers of nodes: input layer, hidden layer and output layer. Each node, besides the input, uses a nonlinear activation function and MLP can have more than one layer of artificial neurons, between the input and the output. The data flows in the forward direction and the neurons are trained with the backpropagation learning. MLPs usually approximate any continuous function, solving problems that are not linearly separable. The general scheme of the architecture used can be seen in Figure 7.2.

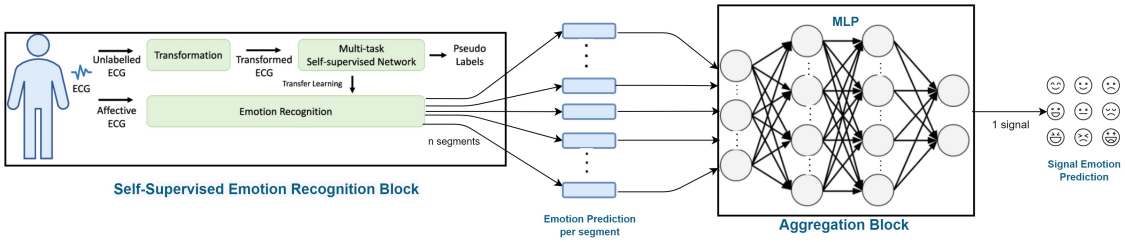


Figure 7.2: Emotion Recognition architecture using MLP aggregation methods.

In this way, MLPs with different parameters were empirically tested in order to obtain an optimized emotion recognition network. During the experiments, MLPs with 1 and 2 hidden layers were tested, using 100, 150, 200 and 512 hidden neurons. Furthermore, other parameters were changed, such as the dropout value in the configuration with 512 hidden nodes, using a dropout of 20% and 40%. In Section 7.3.1.1 the detailed hyperparameters used are described concerning all the architectures used.

7.2.2.3 Long Short-Term Memory and Bidirectional Long Short-Term Memory Methods

The Electrocardiogram is a time-series signal, from which temporal relations can be extracted and be useful for different machine and deep learning tasks, such as emotion recognition. Long short-term memory recurrent neural networks consist of an improvement over the regular recurrent neural networks, since these have a long-term dependency problem, as explained in Chapter 4, Section 4.4.1. Through the incorporation of gating functions into their functioning, LSTMs prevent this problem and maintain a hidden vector h and a memory vector m that control state updates and outputs [190].

In this way, LSTMs use three gates for input and output of the memory space: input gate, forget gate and output gate. These three gates determine how much of the input is reflected, forgotten

and whether or not to display that calculation up to the present [191]. Thus, the LSTM network implements temporal memory through the switch of the gates to prevent gate vanishing. Figure 7.3 represents a basic LSTM unit, where the external inputs are its previous cell state $c_{(t-1)}$, the previous hidden state $h_{(t-1)}$ and the current input vector $x_{(t)}$ [192]. Furthermore, different activation functions are used concerning the three gates and are also displayed in the Figure, presenting both sigmoid and tanh activation functions.

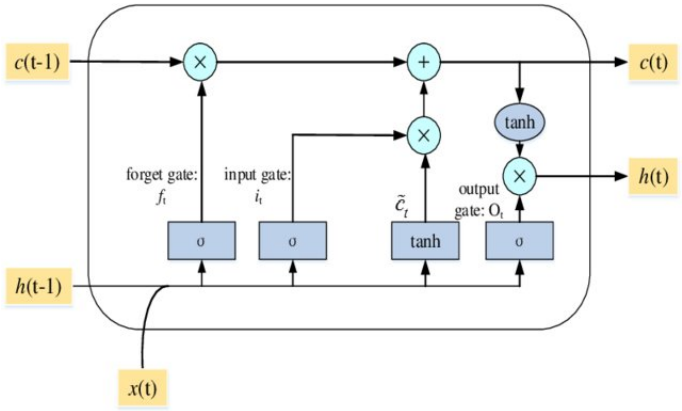


Figure 7.3: Structure of the LSTM unit, from [192].

In this way, LSTMs are useful for learning sequential temporal information, which may be of relevance concerning ECG data. Furthermore, as already discussed, emotion levels are continuous measures that may alter and suffer some changes through time. For this reason, the application of Long Short-Term Memory networks can be an asset in detecting these emotion variations. In addition, literature reported good results on emotion recognition with the application of these structures [29; 170], besides already trying to use them in Chapter 5, Section 5.3.2.

LSTMs also have their bidirectional variants, in which the data is fed to the algorithm from the beginning to the end, coming back to the beginning. In some cases, this kind of architectures are able to outperform feedforward neural networks and, thus, it is a structure that may also be applied for emotion recognition [93; 193].

Having this said, after the fully-connected layers, as aggregation structures, there were applied both an LSTM layer and a BiLSTM layer that received all segments from a given signal, predicting the emotion levels of arousal and valence for that specific signal, as it is presented in Figure 7.4 The results obtained with these structures are further analysed in Section 7.4.1.3

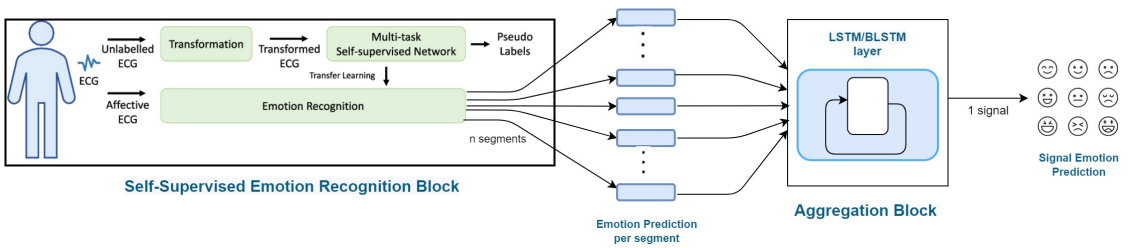


Figure 7.4: Emotion Recognition architecture using LSTM aggregation methods.

7.2.2.4 LSTMs and MLPs combination Methods

Both LSTM and BiLSTM layers were tried out in the previous section, predicting arousal and valence levels after receiving the signal's segments. However, in order to consider more complex aggregation methods, with a high number of hidden nodes in the LSTM or BiLSTM layer, it is possible to add a multilayer perceptron, that would receive the temporal information from the latter structures, further process it and finally obtain the desired emotion predictions, as it is described in Figure 7.5.

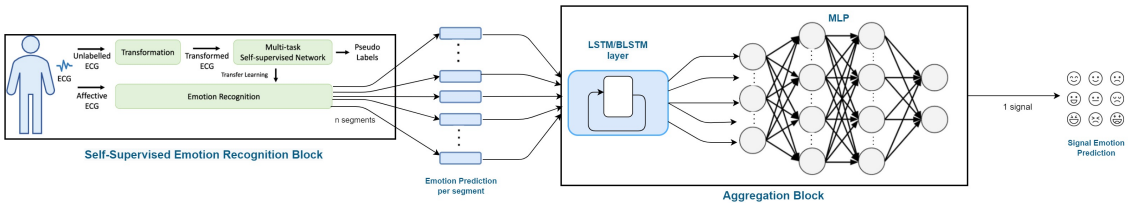


Figure 7.5: Emotion Recognition architecture using combined LSTM and MLP aggregation methods.

In this way, a method identical to the one described in Section 7.3.1.1 was followed, considering different LSTMs and MLPs combinations to be further tested. Both LSTMs and BiLSTMs were applied with a varying number of hidden nodes, combined with MLPs with different structures, such as 1 or 2 layers, varying hidden neurons and dropouts from 0% to 60%. The more detailed information regarding the hyperparameters used for each experiment are further described in Section 7.3.1.2.

7.3 Experimental Setup

7.3.1 Aggregation Hyperparameters

7.3.1.1 Multilayer Perceptron

In Table 7.1 it is possible to understand all the experiments developed for multilayer perceptron aggregation methods. Initially, some configurations were tried out with a random split in order to get an initial idea about the configurations that would be worth exploring. For this, after the fully-connected layers, it was added MLPs with only one hidden layer and 100, 150 and 200 neurons with no dropout applied. After that, a dropout of 20% was considered to the previously mentioned configurations, also considering 512 hidden neurons.

From these configurations, MLP with 200 neurons and no dropout, and MLPs with both 150 and 512 neurons with a dropout of 20% were the ones with better performances and selected to a more thorough testing. Thus, 10-fold cross-testing was considered, in which data was divided into 10 folds of train/test, assuring that all data would be, at some point, used in the test set using a 90%-10% division. Besides the configurations selected, others were considered, such as MLPs with 3 layers, 2 of them with 512 hidden neurons and a dropout of 20%, 40% and 60%. In addition,

a similar configuration was also tried, using a dropout of 20% in the MLP layers but changing the dropout of the fully connected layers from 20% to 40%. Finally, the last configuration applied a different number of hidden neurons considering arousal and valence. For arousal, an MLP with 2 layers, 200 hidden neurons and a dropout of 20% was considered, while, for valence, the number of hidden neurons was 512.

Finally, having into account all the architectures tried and the mean accuracy obtained by the 10 folds tested for each, three MLP configurations were selected as the most promising and further used for other experiments considering more demanding data division, namely, *(1)* the multilayer perceptron with 200 hidden nodes and no dropout, *(2)* the fully-connected layers with a dropout of 40% followed by an MLP with 2 hidden layers, 512 nodes and a dropout of 20% and, lastly, *(3)* a multilayer perceptron with 2 layers, a dropout of 40%, 200 hidden nodes for arousal and 512 for valence.

Table 7.1: **Scheme of the Multilayer Perceptron experiments**. The 3 chosen structures as the most promising, are marked in bold and underlined.

		***No. Layers**	**Hidden Neurons**	**Dropout**		
Random Split tested		2	100	0%	20%	-
			150	0%	20%	-
			200	0%	20%	-
		3	512	20%		
10-fold tested	**chosen from random split**	**2**	**200**	**0%**		
		2	150	20%		
			512	20%		
	others	**2**	****512 V/200 A**	**20%**		
		3	512	20%	40%	60%
			512	20%	**40% (FCN)**	

*Concerns both hidden and classification layers **V - valence A - arousal

7.3.1.2 LSTMs and MLPs combination Methods

Concerning the combined LSTMs and MLPs aggregation techniques, Table 7.2 presents all the structures developed and tested. Initially, it was considered some configurations that were tested with a random data division, so as to indicate the most promising aggregation methods. In this way, an LSTM layer was applied with 64 and 128 hidden nodes, varying the MLP. Concerning this addition, it was used an MLP with 128 hidden nodes, with 1 classification layer, that received the information from LSTM and predicted emotion levels. In addition, other structures were taken into account, such as MLPs with 2 layers, 128 hidden nodes and dropouts of 20% and 40%. BiLSTMs with 128 nodes were also considered, adding an MLP with 1 layer, predicting arousal and valence levels after receiving the last BiLSTM layer, and MLPs with 2 layers, 128 hidden nodes and a varying dropout of 20%, 40% and 60%.

Table 7.2: **Scheme of the combined LSTM and MLP experiments.**. The 3 chosen structures as the most promising, are marked in bold and underlined.

		(Bi)LSTM		MLP				
		Architecture	**Hidden Nodes**	***No. layers**	**Hidden Neurons**	**Dropout**		
Random Split tested		**LSTM**	64	1	-	-		
			128			-		
			128	2	128	20%	40%	-
		BiLSTM	128	1	-	-		
				2	128	20%	40%	60%
10-fold tested	**chosen from random split**	**LSTM**	**128**	**1**	**-**	**-**		
			128	2	128	20%		
		BiLSTM	**128**	**2**	**128**	**40%**	60%	-
	others	**LSTM**	**256**	**2**	**512**	**50%**		
		BiLSTM	64	2	128	20%		

*Concerns both hidden and classification layers

From the results of these initial tests, some of the configurations were further used applying a 10-fold cross-testing, such as LSTM structures with 128 hidden layers followed by MLPs with one and two layers, 128 hidden nodes and 20% dropout. A BiLSTM architecture was also selected for 10-fold cross-testing, with an MLP of 2 layers, 128 hidden nodes and a dropout of 40% and 60%. Furthermore, one more LSTM and BiLSTM structures were considered, as shown in Table 7.2, using 256 and 64 hidden nodes respectively, and MLPs with 2 layers. For the LSTM architecture, the added MLP had 512 hidden nodes and a dropout of 50% was applied. Concerning BiLSTM, the MLP consisted of 128 hidden nodes with a dropout of 20%.

It becomes important to mention that all of these configurations were tried out and empirically tested, not following any other paper or approach to choose the hyperparameters used.

Finally, considering the performances obtained by all of the configurations evaluated, three of them were selected as the best LSTM and MLP combinations achieved and were also used in further experiments with other data divisions. The three configurations chosen were *(1)* LSTM layer with 128 hidden nodes followed by the MLP with one layer, the same number of hidden nodes and a dropout of 20%, *(2)* the LSTM layer with 256 hidden nodes followed by an MLP with 2 layers, 512 hidden nodes and a dropout of 50% and *(3)* BiLSTM layer and MLP with 2 layers and dropout of 40%, both with 128 hidden nodes.

7.3.2 Signal-Independent Settings

Similarly to Chapter 6, more real and demanding data division was considered. As mentioned in Section 7.2, the optimized model was used for every aggregation method tried, which means that only the aggregation architecture was trained considering the 10 cross-testing folds. However, the optimized fully-connected weights were obtained using segments of signals that were probably present in every test fold, since it resulted from a random segment split, which leads to considerably higher and not completely trustworthy results.

In order to assure that the fully-connected weights resulted from training with only the training data of each fold, each signal of the 10 folds was divided into 10-second segments and further used to train the fully-connected layers. This method, highly alike the one used in Chapter 6, Section

6.2.2.1, assured that the segments belonging to training and testing sets belonged to different signals and that the data used to train the aggregation structure was the same involved in the training of the fully-connected layers.

As mentioned, only the two best aggregation methods were tested using these signal-independent settings.

7.3.3 Subject-Independent Settings

After the signal-independent settings, subject-independent train/test sets were also developed and tested concerning the two best aggregation methods. For this, the data pre-processing described in Section 7.2.1 was followed, only changing the data division. Instead of considering 10 folds, 3 data divisions per lead were considered, assuring subject-independence between the sets and following the usual 90%-10% train/test split, which resulted in signals of 3 subjects to belong to the test set and the others to the train set. Like in signal-independent settings, the signals belonging to each set were segmented into 10-second segments to train the fully-connected layers and then the weights obtained were used in the correspondent data split for the subject-independent experiments considered.

7.3.4 Cross-Database Experiments

As the last data division modification, cross-database experiments were also conducted, considering all the signals from both AMIGOS and MAHNOB-HCI database as testing sets and these signals were pre-processed as explained in Section 7.2.1. Since each aggregation method was trained using 10-fold cross-testing, the training model used for these cross-database experiments was the one that resulted in a higher emotion recognition rate for the two most promising aggregation methods.

7.4 Results

In this section, the results from the methods previously explained are presented and analysed. However, having into account the high number of architectures tested, there are only going to be presented the three best methodologies concerning the use of MLPs and the combination of LSTMs and MLPs. Concerning the other methodologies, Appendix A can be accessed, presenting all the results obtained for the structures tested with the 10-fold cross-testing. As already mentioned, for the more demanding data division experiments, only the best method of each was used so as to evaluate the impact of data division on the ability of the network to perform emotion recognition.

Furthermore, since this Chapter is focused on applying aggregation methodologies, it is important to directly compare them with the results that would have been obtained for that given data division without the aggregation. Having this said, for all the MIL techniques experimented, the emotion recognition performances are evaluated considering the prediction obtained by the

segments, exactly after the fully-connected layers and before entering the aggregation block, and the previsions obtained when the entire signal is considered, i. e. after the aggregation method.

7.4.1 Aggregation Methods

7.4.1.1 Heuristic Methods

Concerning the heuristic methods, namely mean, median and maximum, the results obtained were considerably high, as it can be seen in Table 7.3. The maximum aggregation technique was the one that showed the lowest performance when compared with the other two methods, although it still obtained around 80% accuracy for both valence and arousal. Both mean and median were able to achieve a mean accuracy of 86%, outperforming the results reported by Sarkar and Etemad [169], which indicates that using Multiple Instance Learning methods can be beneficial to improve the performance of emotion recognition networks. However, the standard deviation associated with each method was also high, indicating that results vary significantly according to the data fold considered, which is, once more, an indication of the high dependence of emotion recognition to the data division, an inherent behaviour to the usage of small databases.

Table 7.3: Emotion Recognition results concerning heuristic aggregation methods.

	Arousal		**Valence**	
	Accuracy (%)		**Accuracy (%)**	
Heuristic Method	Average Acc.	Std Deviation	Average Acc.	Std Deviation
Mean	86.74	6.85	85.72	4.62
Maximum	80.53	5.27	79.49	3.66
Median	86.23	6.90	86.03	5.17

Furthermore, as mentioned, emotion recognition results without the aggregation methods are also considered, calculating the mean performance of the segments without applying any aggregation method. In Figure 7.6 the accuracy of these four methods is compared in terms of valence and arousal, proving that an individual emotion prediction level per signal can result in higher model performances than predicting emotion for every 10-second segments. Nonetheless, as mentioned, the maximum aggregation method was the one that performed poorly, obtaining similar results to the ones reported when no aggregation is performed. Concerning valence, the maximum was able to outperform the use of instances by 2%, however, for arousal, the results obtained were lower than with segments.

On the other hand, both mean and median improved the accuracy results in around 6% for arousal and 8% and 9% for valence, respectively.

These initial results were promising and highly motivating for more complex aggregation methods to be tried, in order to further improve emotion classification.

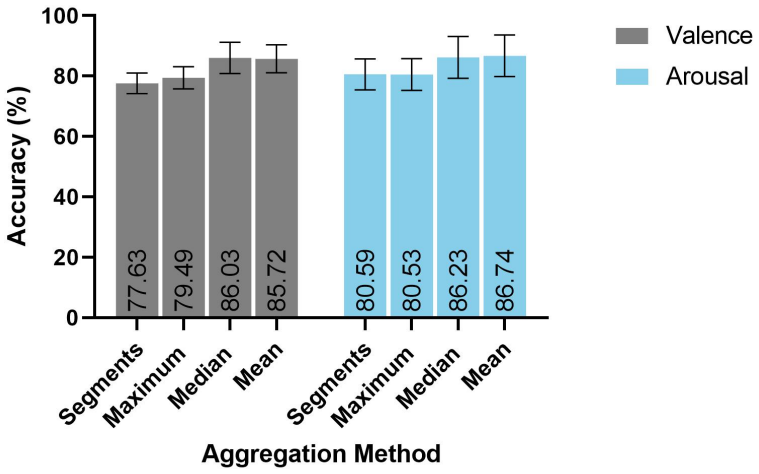


Figure 7.6: **Emotion Recognition results for heuristic aggregation methods** compared with the performance obtained when no aggregation technique is used.

7.4.1.2 Multilayer perceptron Methods

After the use of more simplistic and direct segment-aggregation techniques, more complex structures were developed. Concerning multilayer perceptrons, as explained in Section 7.2.2.2, different architectures were considered until choosing the three most promising for emotion recognition.

In this way, Figure 7.7 presents the emotion recognition performances of these three architectures, concerning both arousal and valence levels. By applying an MLP with 200 hidden nodes and a dropout of 20%, it was possible to achieve the better general performance, obtaining 82.88% accuracy for arousal recognition and 82.33% in terms of valence classification.

Furthermore, the architecture developed where the dropout of the fully-connected layers was increased to 40%, followed by an MLP with 2 layers, 512 hidden nodes and a dropout of 20%, achieved the highest performance concerning arousal recognition, with an accuracy of 83.83%. Concerning valence, there was a slight decrease to 81.37%. Nonetheless, it can be considered a promising architecture with high performances in emotion recognition.

Finally, the highest performance in terms of valence classification was obtained by applying a 2-layer MLP with 512 hidden neurons for valence and 200 for arousal and a dropout of 20%. With this architecture, accuracies of 83.28% and 81.96% were obtained for valence and arousal, respectively.

Furthermore, the accuracy comparison when using segment-prediction, without applying any aggregation techniques was also considered. Although one would expect these accuracies to remain constant for all architectures, since they go through the same model structures (convolutional and fully-connected blocks), with predefined and frozen weights, there must be some randomness in the data, which lead to slight accuracy variations for the prevision obtained after the fully connected layers for each one of the architectures considered.

Having this into account, Figure 7.8 presents the comparison for arousal and valence between

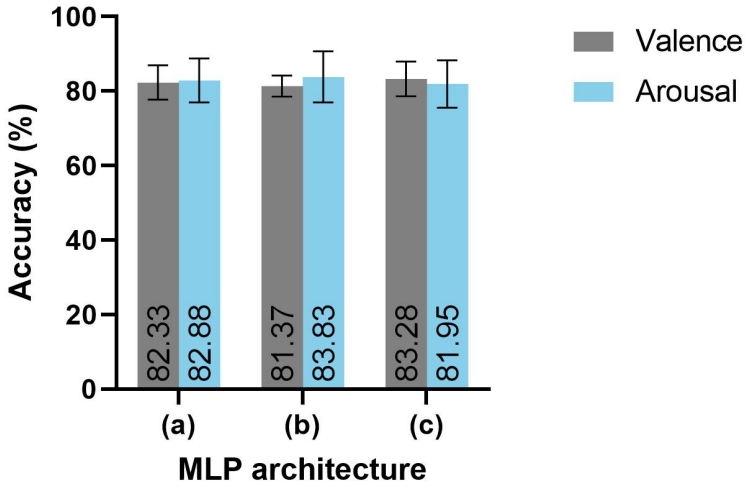
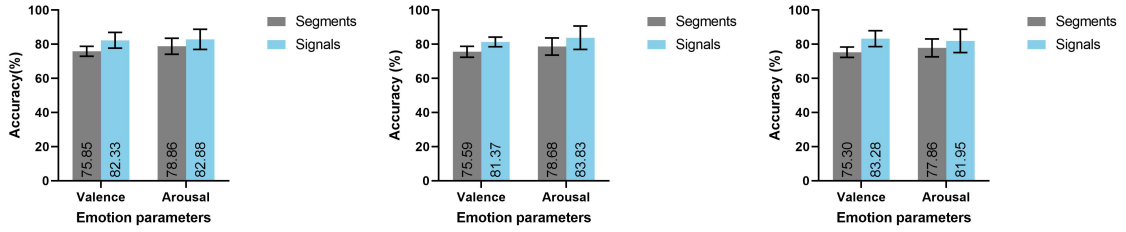


Figure 7.7: **Emotion Recognition results for MLP aggregation methods.** **(a)** MLP, 2 layers, 200 hidden neurons, dropout 20% - best general results for valence and arousal.; **(b)** FCN dropout 40% + MLP, 2 layers, 512 hidden neurons, dropout 20% - best arousal results.; **(c)** MLP, 2 layers, 512 (valence) and 200 (arousal) hidden neurons, dropout 20% - best valence results.



(a) (b) (c)

Figure 7.8: **Emotion Recognition results comparing MLP aggregation methods with no aggregation techniques.** (a) MLP, 2 layers, 200 hidden neurons, dropout 20%.; (b) FCN dropout 40% + MLP, 2 layers, 512 hidden neurons, dropout 20% ; (c) MLP, 2 layers, 512 (valence) and 200 (arousal) hidden neurons, dropout 20%.

the three best methodologies and the correspondent performances when the prediction was done after the fully-connected layers, avoiding the aggregation methods.

As it is presented in this Figure, the results obtained using signal segments, without any aggregation, are lower than when the same segments are aggregated per signal, which illustrates that Multiple Instance Learning can be beneficial by forcing emotions to be predicted per signal and not for every segment. Both MLP aggregation and no aggregation results are quite similar concerning the three methodologies, since the accuracies obtained by segments is around 75% and 78% for valence and arousal, respectively, while for the MLP aggregation methods, both arousal and valence are recognized with an accuracy of about 82%. This fact may indicate that the improvement concerning valence recognition is slightly higher when aggregation techniques are used. The most significant valence improvement is obtained with the last architecture presented (c), MLP with

2 layers, 512 (valence) and 200 (arousal) hidden neurons and a dropout 20%, presenting a 6% increase. Taking arousal into account, the performance improvement is mildly lower, however for the (b) architecture, there is an accuracy increase of 5%.

7.4.1.3 Long Short-Term Memory Methods

Considering the use of both LSTM and BiLSTM layers as aggregation methods, the results obtained were not considerably higher, for which they were not further tested using the 10-fold cross testing. Table 7.4 presents the mean accuracy results obtained for both LSTM and BiLSTM when using two random splits.

Table 7.4: Emotion recognition results for LSTM and BiLSTM architectures.

	LSTM		**BiLSTM**	
	Acc (%)	**Std. Dev**	**Acc (%)**	**Std. Dev**
Arousal	86.51	3.58	84.53	5.96
Valence	80.95	2.38	73.81	2.38

Although the accuracies presented are considerably high, in comparison with the other architectures tested with these same random splits, LSTM and BiLSTM showed lower performances, which made them less promising techniques to be further tested using the 10-fold cross-testing. Furthermore, these random splits seemed to have test sets with easily learned signals which led to usually significant higher accuracy results than with most of the folds further used and their mean. In this way, no more tests were performed considering only an LSTM or BiLSTM layer as the aggregation method.

7.4.1.4 LSTMs and MLPs combination Methods

In addition to the separate use of MLPs and LSTMs as aggregation techniques, both were combined, offering an even more complex methodology for the aggregation of every signal segment.

In this way, as mentioned, different architectures were tested so as to evaluate the results that could be obtained with this type of approach. In Figure 7.9, the results of the three best methods from the combination of LSTM and MLP layers are illustrated.

The overall best results obtained were around 78.5% for both valence and arousal, using an LSTM layer with 128 hidden nodes followed by an MLP with 1 layer and no hidden layers, receiving the data from LSTM and predicting valence and arousal levels.

Concerning arousal, applying an LSTM with 256 hidden nodes followed by an MLP with 1 layer, 512 hidden neurons and a dropout of 50% assured its best recognition, obtaining an accuracy of 78.82%. However, the other two methods also reported an arousal accuracy of around 78%.

The last architecture considered was the one with the second overall better results, consisting of a BiLSTM layer with 128 nodes, and an MLP layer with 128 hidden neurons and a dropout of 40%. For both these two last approaches, valence recognition achieved accuracies around 77%.

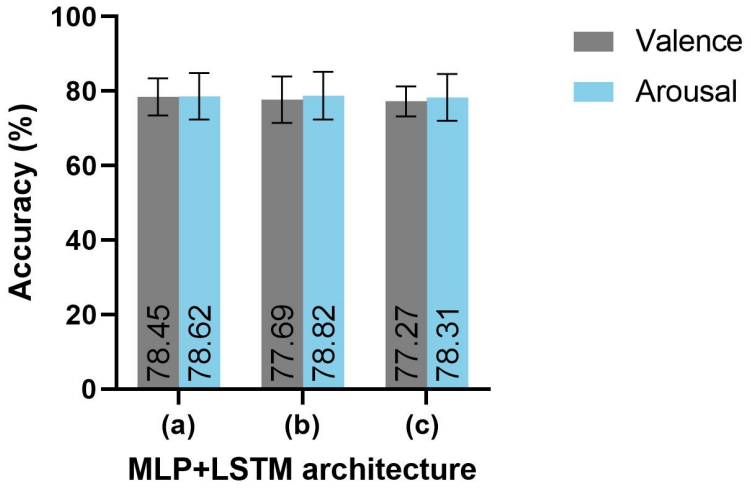


Figure 7.9: **Emotion Recognition results for LSTM and MLP combined aggregation methods.** (a) LSTM layer, 128 hidden nodes + MLP, 1 layer - best overall results; (b) LSTM layer, 256 hidden nodes + MLP, 2 layers, 512 hidden neurons, dropout 50 - best arousal ; (c) BiLSTM, 128 hidden nodes + MLP, 2 layers, 128 hidden neurons, dropout 20% - best 2nd overall results.

In this way, as it can be concluded, the results for the three architectures presented in this section are quite similar, although the methodologies are not that alike, which may indicate a stabilization on the maximum results that can be obtained by combining these two methodologies.

On the other hand, Figure 7.10 provides information regarding the instance-based approach and the accuracies obtained without each aggregation method. As it also happened for both heuristic and MLP methods, all the aggregation techniques based on the combination of LSTM and MLP structures were able to outperform the results obtained without segment aggregation.

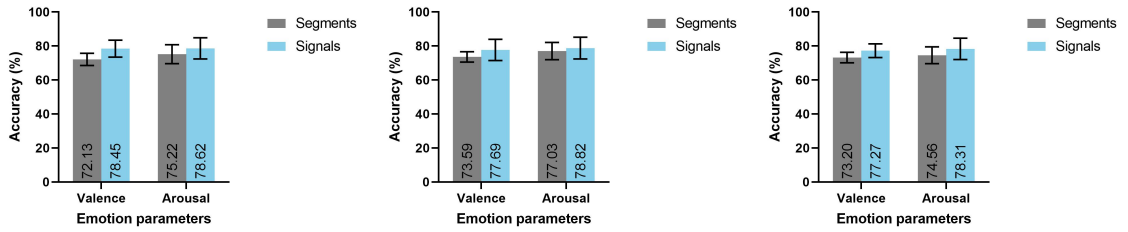


Figure 7.10: **Emotion Recognition results comparing LSTM+MLP aggregation methods and no aggregation techniques.** (a) LSTM layer, 128 hidden nodes + MLP, 1 layer; (b) LSTM layer, 256 hidden nodes + MLP, 2 layers, 512 hidden neurons, dropout 50% ; (c) BiLSTM, 128 hidden nodes + MLP, 2 layers, 128 hidden neurons, dropout 20%.

Like the other aggregation techniques used, the combination of LSTM and MLP also assured an increase of the model performance when compared with the emotion recognition obtained after the fully connected layers and with no aggregation network segment considered. The most

significant accuracy increase reported in Figure 7.10 was for valence, when using with the overall better methodology, consisting of an LSTM layer of 128 hidden nodes followed by an MLP that received the LSTM information and predicted emotion levels. For this, valence accuracy was 6% higher.

Concerning arousal, the improvement of applying aggregation was not so notorious, illustrating a maximum increase of 4% when applying the BiLSTM layer in combination with a 2-layer MLP. However, having into account the results analysed, aggregation techniques show to be efficient, outperforming the methodology of predicting emotion for each segment.

7.4.2 Signal-Independent Settings

As explained in Section 7.3.2, signal-independent settings were considered, to evaluate if the network was also signal-dependent, as it was verified in Chapter 6, Section 6.3.2.1 or if the Multiple Instance Learning techniques applied could, somehow, reduce this problem.

In this way, fully connected weights were trained with the same training data of each fold, not using the optimized weights. In Figure 7.11 the results are presented for the best MLP method and compared with the results obtained using instances and their emotion prediction right after the fully-connected layers.

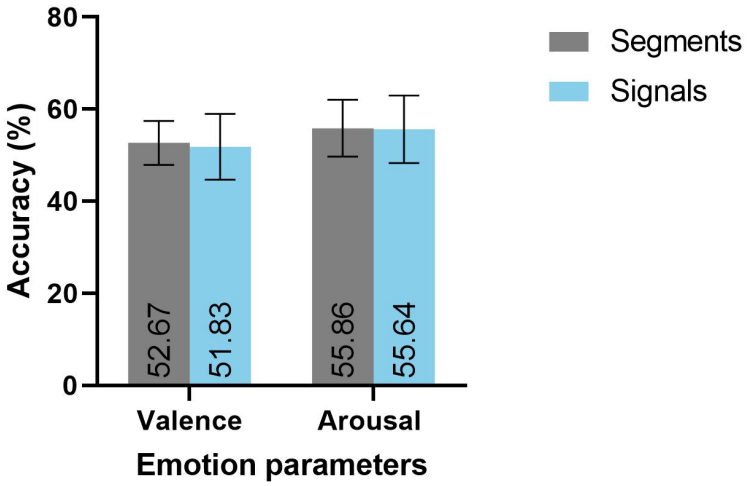


Figure 7.11: Emotion Recognition results for the best MLP method for signal-independent settings

As one can conclude, results are significantly lower than the ones obtained when using the optimized weights, which indicates that the network is unable to perform well under signal-independent settings. Although signal-independent folds are considered to train every aggregation method during this Chapter, since the optimized fully-connected weights that are used were previously trained using a random split of segments, the emotion predictions that result from these layers are not obtained under signal-independent settings and use segments from a lot of different signals. In this way, since the results obtained are higher with this methodology, aggregation methods can then improve them, as seen in Section 7.4.1. However, when fully-connected layers

are trained using this type of more strict data division, with no data leakage between train and test sets, the results suffer a significant decrease, since the network has never "seen" segments from the signals to classify.

In Figure 7.11, the results obtained are quite similar, with or without aggregation methods, proving that, if the initial prediction is not correctly done, aggregation is unable to improve the overall classification. In this particular case, the aggregation methods resulted in slightly lower results and arousal presents a 3% higher accuracy than valence.

Considering Figure 7.12, arousal also presents better results when compared with valence. However, both of these emotion parameters are poorly recognized, obtaining accuracies mildly above 50%, which is random, concerning the binary classification used for these parameters.

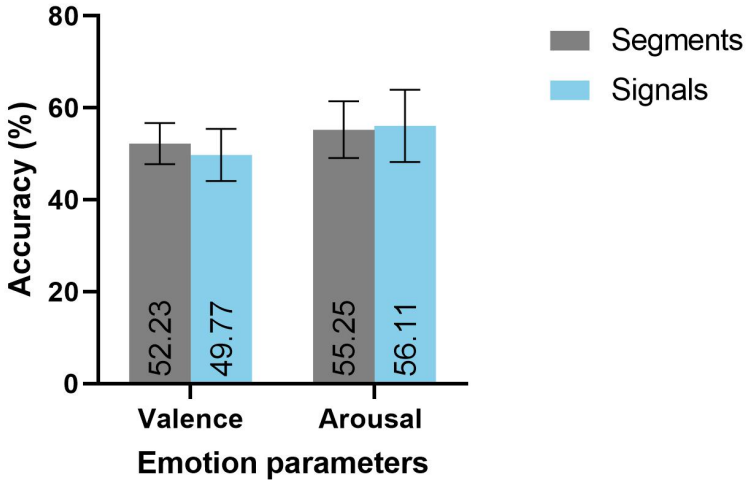


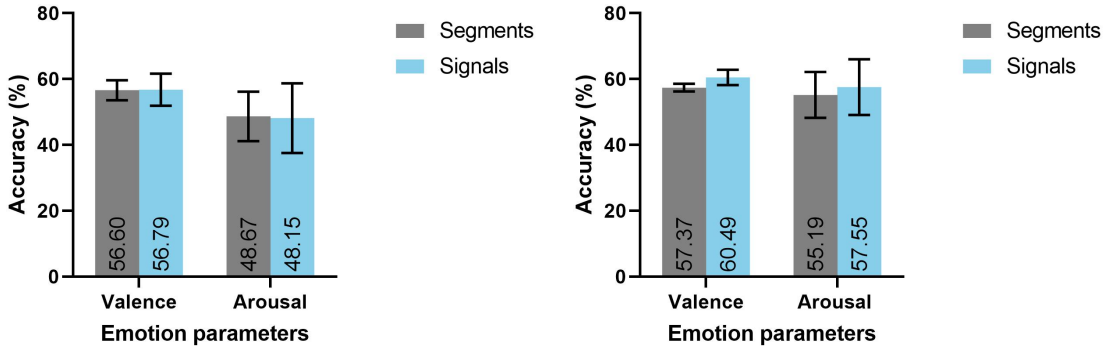
Figure 7.12: Emotion Recognition results for the best LSTM+MLP method for signal-independent settings

Having this said, although there are slight differences between the results obtained for these two methods, the conclusions to be taken are similar, pointing out the lack of ability to obtain good performances using these signal-independent settings. The reasons for this to occur were already discussed in Chapter 6 and do not differ when aggregation methods are added to the network. In this way, it is expected that identical low results are obtained for both subject-independent settings and cross-database experiments, indicating the need for better databases, with a high variety of signals so that the emotion recognition field can develop more robust and complete methodologies.

7.4.3 Subject-Independent Settings

As mentioned in Section 7.3.3, three random data splits were considered for each ECG lead. Furthermore, in Figure 7.13 and 7.14 instance results are also presented and compared with the ones resulting from the best aggregation technique for MLP and LSTM+MLP methods, respectively.

As it can be seen, considering subject-independent settings, the results highly decrease in comparison with the ones presented in Figure 7.7 and obtained with random signal divisions. In



(a) (b)

Figure 7.13: Emotion Recognition results comparing MLP aggregation and no aggregation techniques for subjects independent settings using (a) right and (b) left ECG leads.

this way, as it was already concluded in Chapter 6, the architecture and methodology developed is highly subject-dependent. This problem, as mentioned, is strongly associated with the small databases available that don't allow a robust model training to perform well in both signal and subject-independent settings.

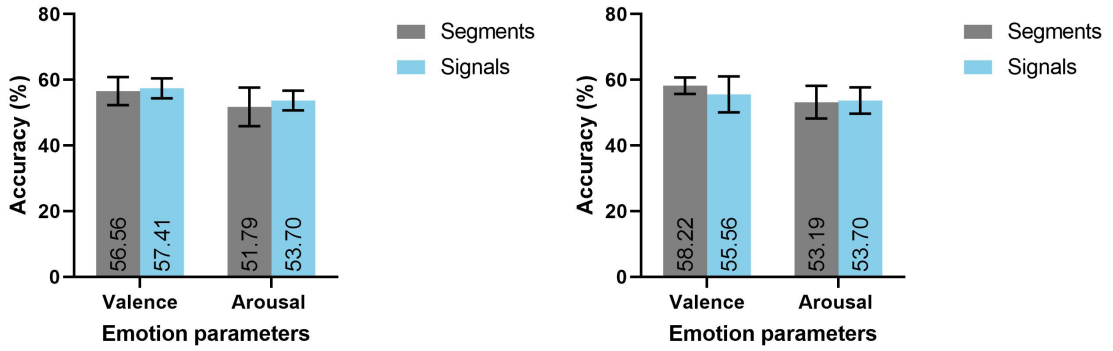
Similarly to what is concluded in Section 7.4.2, it is expected that the addition of aggregation methods will not be sufficient to improve the results in this kind of data divisions. Figure 7.13 demonstrates that the accuracy results are already low concerning the emotion prediction after the fully connected layers. Having this into account, since the segments were not correctly recognized, with relevant patterns detected and important features extracted, the signal information that is fed to the aggregation method is inappropriate for a correct emotion recognition. Thus, it becomes irrelevant how complex or well developed the aggregation network structure might be, if the segments are beforehand poorly predicted.

Concerning both leads of Figure 7.13, it is notorious that the signals from the left lead were better recognized, specially concerning arousal. However, since these results are obtained with only three random splits and standard deviations, in particular for the arousal recognition rates, are quite high, such difference between valence and arousal can be due to the data randomness associated with this type of division, since it was already concluded how data-dependent this methodology is.

Furthermore, concerning Figure 7.14 the results presented lead to similar conclusions. The emotion recognition accuracies concerning segments are, upfront, poor, which could only result in low performances concerning aggregation methodologies.

In addition, the results obtained are quite alike between both leads. Nonetheless, it can be also noticed that valence recognition outperforms arousal for both leads. Nonetheless, as mentioned, due to the small number of different sets considered and the high values of standard deviation associated with some accuracies, one can only conclude that the model is unable to perform well in subject-independent settings, and not infer other conclusions or dependencies between the results

and the impact of the lead or the emotion parameter concerned.



(a) (b)

Figure 7.14: Emotion Recognition results comparing combined LSTM and MLP aggregation and no aggregation techniques for subjects independent settings using (a) right and (b) left ECG leads.

7.4.4 Cross-Database Experiments

The last experiments conducted using aggregation methods were cross-database tests using MAHNOB-HCI and AMIGOS. For these, it was only considered the best approach obtained with MLP and LSTM + MLP aggregation techniques.

For each one of these methods, it was used the model trained by the fold that resulted in accuracies closer to the average results for the 10-fold cross-testing. Table 7.5 presents the performances obtained for both MAHNOB-HCI and AMIGOS databases for the best MLP method, while Table 7.6 illustrates the results obtained for the best LSTM+MLP method.

Table 7.5: Cross-Database results for AMIGOS and MAHNOB-HCI concerning the best MLP method.

		Accuracy (%)	
Database	Lead	**Arousal**	**Valence**
MAHNOB-HCI	ECGr	61.93	63.84
	ECGl	59.55	63.72
AMIGOS	ECGr	44.51	52.51
	ECGl	45.47	50.95

As it can be understood, as expected, the model does not perform well when using different databases as test sets, since the signals considered were not already seen, were acquired in a different environment and belong to unknown subjects, which increases the difficulty of the model to correctly recognize emotions in these settings. However, results were not worse than the ones obtained for signal and subject-independent settings, which indicates that the network dependency starts at the signal level, and, due to this, it will never be able to perform well when other conditions are also altered, besides the signals, such as the subjects or acquisition settings.

Table 7.6: Cross-Database results for AMIGOS and MAHNOB-HCI concerning the best LSTM+MLP method.

		Accuracy (%)	
Database	Lead	**Arousal**	**Valence**
MAHNOB-HCI	ECGr	53.46	40.81
	ECGl	60.02	41.53
AMIGOS	ECGr	50.60	50.84
	ECGl	48.81	46.66

Having this said, as it happened in Chapter 6, Section 6.3.2.3, better results were obtained using MAHNOB-HCI, which is more notorious concerning the MLP method, where with MAHNOB-HCI results were around 60%, while the average accuracies obtained with AMIGOS were lower than 50%. Nonetheless, for the LSTM+MLP architecture, both MAHNOB-HCI and AMIGOS presented similar low results. Concerning MAHNOB-HCI, although, for left lead, a 60.02% accuracy was obtained for arousal recognition, valence accuracies were around 41%.

Having this said, as it was already understood by signal and subject-independent settings results, aggregation is not able to improve the model performances when compared with the use of segments and emotion prediction after the fully connected layers when more demanding data division is considered. The ways of solving this problem consist of better databases, which can further allow for more complex and optimized networks to be developed, increasing emotion recognition quality and reliability.

7.5 Summary and Conclusions

The present Chapter focused on the application of Multiple Instance Learning techniques to the already improved methodology presented in Chapter 6, motivated by other approaches, already mentioned, where this technique was applied and led to more consistent and promising results.

Multiple Instance Learning (MIL) consists of an improved technique to be applied in weakly supervised settings, which is the case, concerning the emotion databases available in literature and described in Chapter 3. Emotions are continuous states, which indicates the need of a continuous labelling method. However, modeling and predicting emotion through time for a continuous data labeling is not usually possible or done. For that reason, the datasets available not only have a reduced number of signals but also one unique label per signal, indicating the emotion felt during a given emotion elicitation technique, such as an emotional video.

As it is known, most approaches use emotion segments instead of the entire signal, which forces them to consider the same label for all the segments. Having this said, there as a high motivation for developing MIL techniques and apply them to the already developed network, aggregating segments that belong to the same signal and predicting emotions per signal and not for every segment.

As explained, different MIL structures were considered, such as simple heuristic methods, such as mean, maximum and median, and more complex architectures, namely MLPs, LSTMs

and the combination of both. Having into account the results obtained for all the MIL approaches tested, it was concluded that aggregation methods may indeed improve emotion recognition. Having this said, this addition to the network was successful and led to higher accuracy levels to be achieved.

Nonetheless, it would be expected that more complex aggregation methods would be able to obtain higher accuracies than the heuristic ways of aggregation. Applying mean, maximum and median to the segments requires no other addition to the network nor the training of some network architecture. Thus, other methodologies such as MLPs and LSTMs, that would be trained, could become more robust and selectively ignore or highlight some segments over others.

However, the results did not corroborate this theory and the higher performances obtained were through the application of mean and median. These results can be due to the fact that signal-independent folds are used to train the aggregation structures, which becomes more demanding for the model, obtaining lower accuracies. Furthermore, it is possible that MLP and LSTM structures may not be focusing on the most correct patterns for emotion recognition, thus worsening the results when compared to a simple mean or median computation. Furthermore, concerning LSTM, only the last layer is further used and passed on to the following structures. Although that layer already presents information from all the others, it can also not be the most correct information to be used concerning emotion recognition. All these facts and others can be possible reasons of why heuristic and more simple ways of aggregation lead to higher emotion recognition performances.

In addition, as mentioned, it would also not be expected that the accuracy obtained without using any aggregation method would change concerning the MIL technique applied, since those results are obtained with the data that only go through the convolutional and fully-connected layers, which had their weights frozen. However, some data randomness can be associated or layers might not be completely well frozen, which leads to their weights to be altered during the aggregation model training. Although it is considered that this layer freezing was well applied, these suppositions are mentioned, since it remains unknown the reason for this segment accuracy changing with the aggregation method applied.

Finally, the already identified weaknesses of this network and most approaches available in the literature, were not mitigated or diminished through the application of MIL techniques. Thus, signal and subject-dependence continued to be verified and to highly decrease the model's performance.

As mentioned during both this and Chapter 6, such problems may be reduced by the use of larger datasets, since ensuring a greater variety and amount of signals brings the opportunity to strongly train models and prepare them for different signals. In this way, models would be more prepared to correctly identify emotions for unknown signals or even subjects.

Having this said, the present Chapter demonstrates to be relevant since it allowed for other techniques to be applied and successfully improve results. Furthermore, it also highlighted this field's already detected problem of inadequate evaluation.

Chapter 8

Conclusion and Future Work

The main goal of this dissertation was to develop a deep learning architecture for emotion recognition using ECG signals, having into account more realistic data division scenarios and assessing signal and subject-dependence problems. To achieve this, there was the need to explore the fundamentals behind the electrocardiogram and emotions, as well as understanding their natural functioning and behaviour. Furthermore, after getting acquainted with the basic theoretical information regarding these aspects, several prior art approaches and techniques were explored and analysed, so as to understand the state-of-the-art already developed, and the scientific evolution that had already happened in this field, as opposed to what could still be lacking.

Having this said, interesting approaches were analysed, with promising and motivating results, that indicated the potential of applying both machine and deep learning techniques for the development of emotion recognition methodologies. In addition, although deep learning approaches were considerably less explored in literature than hand-crafted methods, the existing studies using deep learning architectures reported results able to outperform machine learning approaches, which demonstrated the potential of applying deep neural networks in this field.

However, some limitations were encountered, namely concerning data division methods, presenting random train/test splits. In addition, most deep learning approaches used ECG segments instead of the entire signals, leading to train and test sets with segments belonging to the same signal. This unrealistic way of evaluating model performances translate into some misleading results, since it considers sets with data belonging to the same subject and from the same signals. In this way, it became relevant to replicate some approaches, so as to find a strong baseline for the work developed in this dissertation, as well as testing these promising methods with more strict and demanding data settings.

Having this said, different approaches were explored and replicated, having into account different methodologies or deep learning architectures. However, the lack of important information and the absence of clarity concerning all the steps of the proposed methodologies resulted in low emotion recognition performances obtained during the replications. Considering the mentioned drawbacks and limitations unveiled by the replication stage, the self-learning method proposed by Sarkar and Etemad [169] revealed itself to be the most promising, and was used as baseline for the

rest of the work conducted.

Thus, different improvements were applied to this approach, starting with the self-supervised learning task. The techniques applied for its improvement were successful, which led to an increase in both arousal and valence recognition accuracies, although never reaching the results reported by Sarkar and Etemad [169]. Furthermore, different data divisions were applied to this improved emotion recognition network, resulting in considerably lower performances than when applying random data splits, which indicated both signal and subject-dependency.

Nonetheless, so as to further improve the network developed, Multiple Instance Learning techniques were considered, with the main goal of overcoming some limitations regarding the sparsely labelled datasets available. By applying different aggregation methodologies through the use of Multilayer Perceptron and Long Short-Term Memory structures, emotion was predicted per annotated signal, instead of predicting arousal and valence levels for all the segments belonging to the same signal. With these methods, results considerably improved when compared with the emotion predictions per segment. Furthermore, so as to obtain more reliable results, 10-fold cross-testing was used for signal-independent settings, in order to train these aggregation structures. Having this said, MIL showed to be efficient and to surpass some problems that result from small databases, with scarce annotations. However, these aggregation methods were not sufficient to reduce the signal and subject-dependence problems already previously encountered. For all the realistic settings considered, such as signal and subject-independent settings and cross-database experiments, the model performance suffered a sharp decline, proving the remaining limitations in methodologies that are developed and improved using random data divisions.

These results put in perspective the overall approaches that reported promising model performances with random data splits, since this dependency prevents them from truly have an application into the real world. Thus, besides trying to find a promising methodology for emotion recognition, this thesis focused on understanding the current problems and systematic mistakes that are being done in literature, which falsely demonstrate a more evolved state of the art by not actually considering realistic data divisions.

However, the mentioned dependencies can be easily justified not by the use of shallow networks or the application of incomplete methodologies, but due to the databases available, which are considerably small. Not having sufficient subjects nor sufficient data per subject, results in models that are not well trained, and are unable to successfully recognize emotions for unseen data, even with complex architectures that would probably allow for accurate results to be obtained. As it is known, deep learning approaches require a large amount of data and, at this point, there are no databases available with physiological signals for emotion recognition that may ensure the amount of data needed to obtain robust models and with real application potential.

In this way, besides reporting these systematic and recurrent problems, this dissertation still improved an available literature approach, considering, as it was intended, different deep learning architectures, such as Long Short-Term Memory layers and Multilayer Perceptrons. Nonetheless, other methods could also be considered, such as hand-crafted methodologies or even multimodal approaches. However, due to the short time to develop the present dissertation, there is a lot of

future work that can be done so as to improve and scientifically evolve this field.

Concerning the limitations encountered, there's the need to develop larger databases and, if possible, with stronger supervised methodologies, presenting a variety of signals per subject and a high number of subjects. Only with sufficient data and signal variety, it will be possible to develop techniques that are robust enough to recognize emotions without any signal, subject or database dependency. Furthermore, along with databases enlargement, different techniques may be considered and used to reduce some of the limitations of low data availability, such as unsupervised learning, one-shot learning, weakly-supervised learning and, as used in this dissertation, self-supervised learning methodologies. All of these approaches are suitable and often bring improvements in situations with little data available, as it was seen in this dissertation.

Furthermore, other future improvements may focus on the development of subject-independent techniques since it was one of the goals of this dissertation due to the reduced number of these methodologies in this field. However, since the problem is still at the signal level, there's the need to first solve this signal dependency and then focus on specific techniques to ensure good performances with subject-independent settings.

Having this said, it can be considered that the dissertation goals were achieved and, hopefully, it may be a scientific work with interesting discussions and future work suggestions concerning the emotion recognition and affective computing field.

Appendix A

Aggregation Methods

Table A.1: Emotion Recognition results for the other different MLP aggregation methods tested using 10-fold cross testing.

Method			Arousal		Valence	
No. Layers	**Hidden neurons**	**Dropout**	**Acc (%)**	**Std. Dev.**	**Acc (%)**	**Std. Dev.**
2	150	20%	81.68	6.85	79.69	3.73
2	512	20%	82.82	6.83	82.10	4.44
3	512	20%	82.9	7.00	82.10	3.88
3	512	40%	82.70	6.91	81.88	2.70
3	512	60%	82.66	6.78	83.06	3.88

Table A.2: Emotion Recognition results for the other different MLP + LSTM aggregation methods tested using 10-fold cross testing.

Method								
(Bi)LSTM		MLP			Arousal		Valence	
Architecture	**Hidden Nodes**	**No. Layers**	**Hidden neurons**	**Dropout**	**Acc (%)**	**Std. Dev.**	**Acc (%)**	**Std. Dev.**
LSTM	128	2	128	20%	76.67	7.99	75.14	2.11
BiLSTM	128	2	128	60%	78.15	7.60	76.54	4.49
BiLSTM	64	2	128	20%	74.69	7.37	74.81	5.89

References

[1] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas. Human emotion recognition: Review of sensors and methods. *Sensors (Switzerland)*, 20(3), 2020. doi: 10.3390/s20030592.

[2] M. Malfaz and M. A. Salichs. A new architecture for autonomous robots based on emotions. *IFAC Proceedings Volumes (IFAC-PapersOnline)*, 37(8):805–809, 2004. doi: 10.1016/s1474-6670(17)32079-7.

[3] S. Sosnowski, A. Bittermann, K. Kühnlenz, and M. Buss. Design and evaluation of emotion-display eddie. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3113–3118, 2006. doi: 10.1109/IROS.2006.282330.

[4] M. A. Delkhoon and F. Lotfizadeh. An Investigation on the Effect of Gender on Emotional Responses and Purchasing Intention Due to Advertisements. *UCT Journal of Social Sciences and Humanities Research*, 2(01):6–11, 2014.

[5] S. Scotti, M. Mauri, R. Barbieri, B. Jawad, S. Cerutti, L. Mainardi, E. Brown, and M. Villamira. Automatic quantitative evaluation of emotions in e-learning applications. *Conference proceedings : Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, 1: 1359–62, Feb. 2006. doi: 10.1109/IEMBS.2006.260601.

[6] B. Woolf, T. Dragon, I. Arroyo, D. Cooper, W. Burleson, and K. Muldner. Recognizing and responding to student affect. In *Human-Computer Interaction. Ambient, Ubiquitous and Intelligent Interaction*, pages 713–722, Jul. 2009. doi: 10.1007/978-3-642-02580-8_78.

[7] G. N. Yannakakis and J. Hallam. Real-time game adaptation for optimizing player satisfaction. *IEEE Transactions on Computational Intelligence and AI in Games*, 1(2):121–133, 2009. doi: 10.1109/TCIAIG.2009.2024533.

[8] M. El Ayadi, M. S. Kamel, and F. Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011. ISSN 00313203. doi: 10.1016/j.patcog.2010.09.020.

[9] S. Katsigiannis and N. Ramzan. DREAMER: A Database for Emotion Recognition Through EEG and ECG Signals from Wireless Low-cost Off-the-Shelf Devices. *IEEE Journal of Biomedical and Health Informatics*, 22(1):98–107, 2015. ISSN 21682194. doi: 10.1109/JBHI.2017.2688239.

[10] P. Bota, C. Wang, A. Fred, and H. Silva. Emotion assessment using feature fusion and decision fusion classification based on physiological data: Are we there yet? *Sensors (Switzerland)*, 20(17):1–17, 2020. ISSN 14248220. doi: 10.3390/s20174723.

[11] F. Agrafioti. *ECG in Biometric Recognition: Time Dependency and Application Challenges*. PhD thesis, University of Toronto, 2011.

[12] C. Godin, F. Prost-Boucle, A. Campagne, S. Charbonnier, S. Bonnet, and A. Vidal. Selection of the most relevant physiological features for classifying emotion. In *Proceedings of the 2nd International Conference on Physiological Computing Systems*, PhyCS 2015, page 17–25, Setubal, PRT, 2015. SCITEPRESS - Science and Technology Publications, Lda. ISBN 9789897580857. doi: 10.5220/0005238600170025.

[13] B. Reuderink, C. Mühl, and M. Poel. Valence, arousal and dominance in the EEG during game play. *International Journal of Autonomous and Adaptive Communications Systems*, 6(1):45, 2013. doi: 10.1504/ijaacs.2013.050691.

[14] J. Miranda, M. Khomami Abadi, N. Sebe, and I. Patras. Amigos: A dataset for mood, personality and affect research on individuals and groups. *IEEE Transactions on Affective Computing*, PP, Feb. 2017. doi: 10.1109/TAFFC.2018.2884461.

[15] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012. ISSN 19493045. doi: 10.1109/T-AFFC.2011.25.

[16] M. K. Abadi, R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe. DECAF: MEG-Based Multimodal Database for Decoding Affective Physiological Responses. *IEEE Transactions on Affective Computing*, 6(3):209–222, 2015. ISSN 19493045. doi: 10.1109/TAFFC.2015.2392932.

[17] S. Koelstra, C. Mühl, M. Soleymani, J. S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. DEAP: A database for emotion analysis; Using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012. ISSN 19493045. doi: 10.1109/T-AFFC.2011.15.

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

[19] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou. End-to-End Multimodal Emotion Recognition Using Deep Neural Networks. *IEEE Journal on Selected Topics in Signal Processing*, 11(8):1301–1309, 2017. ISSN 19324553. doi: 10.1109/JSTSP.2017.2764438.

[20] N. Fragopanagos and J. G. Taylor. Emotion recognition in human-computer interaction. *Neural Networks*, 18(4):389–405, 2005. ISSN 08936080. doi: 10.1016/j.neunet.2005.03.006.

[21] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak. Emotion recognition using facial expressions. *Procedia Computer Science*, 108:1175–1184, 2017. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2017.05.025. International Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland.

[22] K. Kaulard, D. W. Cunningham, H. H. Bülthoff, and C. Wallraven. The MPI facial expression database - a validated database of emotional and conversational facial expressions. *PLoS ONE*, 7(3), 2012. ISSN 19326203. doi: 10.1371/journal.pone.0032321.

[23] F. Agrafioti, D. Hatzinakos, and A. K. Anderson. ECG pattern analysis for emotion detection. *IEEE Transactions on Affective Computing*, 3(1):102–115, 2012. ISSN 19493045. doi: 10.1109/T-AFFC.2011.28.

[24] S. Tivatansakul and M. Ohkura. Emotion Recognition using ECG Signals with Local Pattern Description Methods. *International Journal of Affective Engineering*, 15(2):51–61, 2016. ISSN 2187-5413. doi: 10.5057/ijae.ijae-d-15-00036.

[25] G. Keren, T. Kirschstein, E. Marchi, F. Ringeval, and B. Schuller. End-to-end learning for dimensional emotion recognition from physiological signals. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 985–990, 2017. doi: 10.1109/ICME.2017.8019533.

[26] A. Goshvarpour, A. Abbasi, and A. Goshvarpour. An accurate emotion recognition system using ECG and GSR signals and matching pursuit method. *Biomedical Journal*, 40(6):355–368, 2017. ISSN 23194170. doi: 10.1016/j.bj.2017.11.001.

[27] Y. L. Hsu, J. S. Wang, W. C. Chiang, and C. H. Hung. Automatic ECG-Based Emotion Recognition in Music Listening. *IEEE Transactions on Affective Computing*, 11(1):85–99, 2020. ISSN 19493045. doi: 10.1109/TAFFC.2017.2781732.

[28] N. S. Suhaimi, J. Mountstephens, and J. Teo. EEG-Based Emotion Recognition: A State-of-the-Art Review of Current Trends and Opportunities. *Computational Intelligence and Neuroscience*, 2020, 2020. ISSN 16875273. doi: 10.1155/2020/8875426.

[29] S. Siddharth, T.-P. Jung, and T. Sejnowski. Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing. *IEEE Transactions on Affective Computing*, PP:1–1, May 2019. doi: 10.1109/TAFFC.2019.2916015.

[30] H. Ferdinando, T. Seppanen, and E. Alasaarela. Comparing features from ecg pattern and hrv analysis for emotion recognition system. In *2016 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–6, 2016. doi: 10.1109/CIBCB.2016.7758108.

[31] C. C. Kordsachia, I. Labuschagne, S. C. Andrews, and J. C. Stout. Diminished facial EMG responses to disgusting scenes and happy and fearful faces in Huntington's disease. *Cortex*, 106:185–199, 2018. ISSN 19738102. doi: 10.1016/j.cortex.2018.05.019.

[32] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu, and X. Yang. A review of emotion recognition using physiological signals. *Sensors (Switzerland)*, 18(7), 2018. ISSN 14248220. doi: 10.3390/s18072074.

[33] J. Ribeiro Pinto, J. S. Cardoso, and A. Lourenco. Evolution, current challenges, and future possibilities in ECG Biometrics. *IEEE Access*, 6:34746–34776, 2018. ISSN 21693536. doi: 10.1109/ACCESS.2018.2849870.

[34] J. Chen, B. Hu, Y. Wang, P. Moore, Y. Dai, L. Feng, and Z. Ding. Subject-independent emotion recognition based on physiological signals: A three-stage decision method. *BMC Medical Informatics and Decision Making*, 17(Suppl 3), 2017. ISSN 14726947. doi: 10.1186/s12911-017-0562-x.

[35] Emotion | definition of emotion by oxford dictionary on lexico.com also meaning of emotion. https://www.lexico.com/definition/emotion. (Accessed on 02/02/2021).

[36] C. L. Lisetti. Affective computing. *Pattern Analysis and Applications*, 1(1):71–73, 1998. ISSN 1433-7541. doi: 10.1007/bf01238028.

[37] M. Strauss, C. Reynolds, S. Hughes, K. Park, G. McDarby, R. Picard, J. Tao, and T. Tan. Affective Computing and Intelligent Interaction. In *First International Conference, ACII 2005*, volume 3784, pages 699–706, Oct. 2005. ISBN 978-3-540-29621-8. doi: 10.1007/11573548.

[38] J. Kaur and J. R. Saini. Emotion Detection and Sentiment Analysis in Text Corpus: A Differential Study with Informal and Formal Writing Styles. *International Journal of Computer Applications*, 101(9):1–9, 2014. ISSN 0975-8887. doi: 10.5120/17712-8078.

[39] C. Darwin, J. Murray, and A. Street. *Expression of the Emotions in Man and Animals*. John Murray, London, Albemarle Street, 1872. doi: 10.103710001-000.

[40] P. Ekman. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169–200, 1992. doi: 10.1080/02699939208411068.

[41] R. Plutchik. A psychoevolutionary theory of emotions. *Social Science Information*, 21 (4-5):529–553, 1982. doi: 10.1177/053901882021004003.

[42] Plutchik's wheel of emotion - system concepts. https://www.system-concepts.com/insights/emotion-ai-part-1/plutchiks-wheel-of-emotion/, Jan. 2020. (Accessed on 01/11/2021).

[43] C. Izard. Basic emotions, natural kinds, emotion schemas, and a new paradigm. *Perspectives on psychological science*, 2:260, Sep. 2007. doi: 10.1111/j.1745-6916.2007.00044.x.

[44] J. Russell, A. Weiss, and G. Mendelsohn. Affect grid: A single-item scale of pleasure and arousal. *Journal of Personality and Social Psychology*, 57:493–502, Sep. 1989. doi: 10.1037/0022-3514.57.3.493.

[45] P. J. Lang. The Emotion Probe: Studies of Motivation and Attention. *American Psychologist*, 50(5):372–385, 1995. doi: 10.1037/0003-066X.50.5.372.

[46] A. Mehrabian. Comparison of the PAD and PANAS as Models for Describing Emotions and for Differentiating Anxiety from Depression. *Journal of Psychopathology and Behavioral Assessment*, 19(4):331–357, 1997. doi: 10.1007/BF02229025.

[47] O. Sourina and Y. Liu. Eeg-enabled affective human-computer interfaces. In C. Stephanidis and M. Antona, editors, *Universal Access in Human-Computer Interaction. Design and Development Methods for Universal Access*, pages 536–547, Cham, 2014. Springer International Publishing. ISBN 978-3-319-07437-5. doi: 10.1007/978-3-319-07437-5_51.

[48] A. Mehrabian. Pleasure-arousal.dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292, 1996. doi: 10.1007/BF02686918.

[49] J. Fontaine, K. Scherer, E. Roesch, and P. Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18:1050–7, Jan. 2007. doi: 10.1111/j.1467-9280.2007.02024.x.

[50] E. Cambria, A. Livingstone, and A. Hussain. The hourglass model revisited. *IEEE Intelligent Systems*, 35(5):96–102, 2020. doi: 10.1109/MIS.2020.2992799.

[51] W. L. Zheng, J. Y. Zhu, and B. L. Lu. Identifying stable patterns over time for emotion recognition from eeg. *IEEE Transactions on Affective Computing*, 10(3):417–429, 2019. ISSN 19493045. doi: 10.1109/TAFFC.2017.2712143.

[52] X. Guo. Study of emotion recognition based on electrocardiogram and RBF neural network. *Procedia Engineering*, 15:2408–2412, 2011. ISSN 18777058. doi: 10.1016/j.proeng.2011.08.452.

[53] M. Liu, D. Fan, X. Zhang, and Gong. Human emotion recognition based on galvanic skin response signal feature selection and svm. In *2016 International Conference on Smart City and Systems Engineering (ICSCSE)*, pages 157–160, 2016. doi: 10.1109/ICSCSE.2016.0051.

[54] Z. Yin, M. Zhao, Y. Wang, J. Yang, and J. Zhang. Recognition of emotions using multimodal physiological signals and an ensemble deep learning model. *Computer Methods and Programs in Biomedicine*, 140:93–110, 2017. ISSN 18727565. doi: 10.1016/j.cmpb.2016.12.005.

[55] B. Cheng and G. Liu. Emotion recognition from surface emg signal using wavelet transform and neural network. In *2008 2nd International Conference on Bioinformatics and Biomedical Engineering*, pages 1363–1366, 2008. doi: 10.1109/ICBBE.2008.670.

[56] M. S. Lee, Y. K. Lee, M. T. Lim, and T. K. Kang. Emotion recognition using convolutional neural network with selected statistical photoplethysmogram features. *Applied Sciences (Switzerland)*, 10(10), 2020. ISSN 20763417. doi: 10.3390/app10103501.

[57] J. Cohn and F. De la Torre. Automated face analysis for affective computing. *Handbook of affective computing*, pages 131–150, Jan. 2014. doi: 10.1093/oxfordhb/9780199942237.013.020.

[58] P. Ekman and W. Friesen. Facial action coding system: Investigator's guide. *Consulting Psychologists Press*, 1978.

[59] S. Bashyal and G. K. Venayagamoorthy. Recognition of facial expressions using Gabor wavelets and learning vector quantization. *Engineering Applications of Artificial Intelligence*, 21(7):1056–1064, 2008. ISSN 09521976. doi: 10.1016/j.engappai.2007.11.010.

[60] B. Ko. A brief review of facial emotion recognition based on visual information. *Sensors*, 18:401, Jan. 2018. doi: 10.3390/s18020401.

[61] H. Zhang, A. Jolfaei, and M. Alazab. A face emotion recognition method using convolutional neural network and image edge computing. *IEEE Access*, PP:1–1, Oct. 2019. doi: 10.1109/ACCESS.2019.2949741.

[62] V. Mishra. Real-time Facial Expression Recognition using Convolutional Neural Networks. *International Journal for Research in Applied Science and Engineering Technology*, 8(6):1389–1394, 2020. doi: 10.22214/ijraset.2020.6225.

[63] S. Li and W. Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, pages 1–1, 2020. doi: 10.1109/TAFFC.2020.2981446.

[64] W. Mellouk and H. Wahida. Facial emotion recognition using deep learning: review and insights. *Procedia Computer Science*, 175:689–694, Jan. 2020. doi: 10.1016/j.procs.2020.07.101.

[65] A. Dawel, L. Wright, J. Irons, R. Dumbleton, R. Palermo, R. O'Kearney, and E. McKone. Perceived emotion genuineness: normative ratings for popular facial expression stimuli and the development of perceived-as-genuine and perceived-as-fake sets. *Behavior Research Methods*, 49(4):1539–1562, 2017. ISSN 15543528. doi: 10.3758/s13428-016-0813-2.

[66] J. Nicholson, K. Takahashi, and R. Nakatsu. Emotion recognition in speech using neural networks. In *ICONIP'99. ANZIIS'99 ANNES'99 ACNN'99. 6th International Conference on Neural Information Processing. Proceedings (Cat. No.99EX378)*, volume 2, pages 495–501 vol.2, 1999. doi: 10.1109/ICONIP.1999.845644.

[67] R. Banse and K. Scherer. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70:614–36, Apr. 1996. doi: 10.1037/0022-3514.70.3.614.

[68] M. El, M. S. Kamel, and F. Karray. Survey on speech emotion recognition : Features , classification schemes , and databases. *Pattern Recognition*, 44(3):572–587, 2011. ISSN 0031-3203. doi: 10.1016/j.patcog.2010.09.020.

[69] T. Nwe, S. Foo, and L. De Silva. Speech emotion recognition using hidden markov models. *Speech Communication*, 41:603–623, Nov. 2003. doi: 10.1016/S0167-6393(03)00099-2.

[70] B. Schuller, G. Rigoll, and Lang. Hidden markov model-based speech emotion recognition. In *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, volume 1, pages I–401, 2003. doi: 10.1109/ICME.2003.1220939.

[71] V. Kamble, R. Deshmukh, A. Karwankar, V. Ratnaparkhe, and S. Annadate. Emotion recognition for instantaneous marathi spoken words. In *Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA)*, Nov. 2014. ISBN 978-3-319-12011-9. doi: 10.1007/978-3-319-12012-6_37.

[72] A. Mehrabian. *Nonverbal Communication*. Routledge: London, Uk, 2017.

[73] J. F. Iaccino. *Left brain-right brain differences: Inquiries, evidence, and new approaches*. Lawrence Erlbaum Associates, Inc, 1993.

[74] P. Molchanov, S. Gupta, and K. Kim. Hand gesture recognition with 3d convolutional neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1–7, 2015. doi: 10.1109/CVPRW.2015.7301342.

[75] T. Sapiński, D. Kamińska, A. Pelikant, and G. Anbarjafari. Emotion recognition from skeletal movements. *Entropy*, 21:646, Jun. 2019. doi: 10.3390/e21070646.

[76] C. Corneanu, F. Noroozi, D. Kaminska, T. Sapinski, S. Escalera, and G. Anbarjafari. Survey on emotional body gesture recognition. *IEEE Transactions on Affective Computing*, pages 1–1, Oct. 5555. doi: 10.1109/TAFFC.2018.2874986.

[77] N. Dalal, B. Triggs, and C. Schmid. Human Detection Using Oriented Histograms of Flow and Appearance. In A. Leonardis, H. Bischof, and A. Pinz, editors, *European Conference on Computer Vision (ECCV '06)*, volume 3952 of *Lecture Notes in Computer Science (LNCS)*, pages 428–441, Graz, Austria, May 2006. Springer-Verlag. doi: 10.1007/11744047_33.

[78] P. Felzenszwalb, D. Mcallester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. doi: 10.1109/CVPR.2008.4587597.

[79] T. Starner and A. Pentland. Real-time american sign language recognition from video using hidden markov models. In *Proceedings of International Symposium on Computer Vision - ISCV*, pages 265–270, 1995. doi: 10.1109/ISCV.1995.477012.

[80] A. D. Wilson, S. Member, and I. C. Society. Parametric hidden markov models for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9):884–900, 1999. doi: 10.1109/34.790429.

[81] O. Amft, C. Lombriser, T. Stiefmeier, and G. Tröster. Recognition of user activity sequences using distributed event detection. *Second European Conference on Smart Sensing and Context (EuroSSC)*, pages 126–141, Oct. 2007. doi: 10.1007/978-3-540-75696-5_8.

[82] H. Gunes and M. Piccardi. Affect recognition from face and body: early fusion vs. late fusion. In *2005 IEEE International Conference on Systems, Man and Cybernetics*, volume 4, pages 3437–3443 Vol. 4, 2005. doi: 10.1109/ICSMC.2005.1571679.

[83] S. Saha, S. Datta, A. Konar, and R. Janarthanan. A study on emotion recognition from body gestures using kinect sensor. In *2014 International Conference on Communication and Signal Processing*, pages 056–060, 2014. doi: 10.1109/ICCSP.2014.6949798.

[84] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza. Emotion recognition in context. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1960–1968, 2017. doi: 10.1109/CVPR.2017.212.

[85] G. Caridakis, G. Castellano, L. Kessous, A. Raouzaiou, L. Malatesta, S. Asteriadis, and K. Karpouzis. Multimodal emotion recognition from expressive faces, body gestures and speech. In *Artificial Intelligence and Innovations 2007: from Theory to Applications*, IFIP Advances in Information and Communication Technology, pages 375–388, United States, 2007. Springer. ISBN 9780387741604. doi: 10.1007/978-0-387-74161-1_41.

[86] L. Kessous, G. Castellano, and G. Caridakis. Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis. *Journal on Multimodal User Interfaces*, 3:33–48, Mar. 2009. doi: 10.1007/s12193-009-0025-5.

[87] S. D. Kreibig. Autonomic nervous system activity in emotion : A review. *Biological Psychology*, 84(3):14–41, 2019. ISSN 0301-0511. doi: 10.1016/j.biopsycho.2010.03.010.

[88] E. Vanman. An empirical study of machine learning techniques for affect recognition in human-robot interaction. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2662–2667, 2005. doi: 10.1109/IROS.2005.1545344.

[89] G. Vloed and J. Berentsen. Measuring emotional wellbeing with a non-intrusive bed sensor. In *IFIP Conference on Human-Computer Interaction (INTERACT)*, volume 5727, pages 908–911, Aug. 2009. doi: 10.1007/978-3-642-03658-3_108.

[90] A. Fridlund and C. E. Izard. *Electromyographic studies of facial expressions of emotions and patterns of emotions*, pages 243—-286. Social psychophysiology: A sourcebook, 1983.

[91] K. H. Kim, S. W. Bang, and S. R. Kim. Emotion recognition system using short-term monitoring of physiological signals. *Medical and Biological Engineering and Computing*, 42(3):419–427, 2004. ISSN 01400118. doi: 10.1007/BF02344719.

[92] G. Valenza, L. Citi, A. Lanatá, E. P. Scilingo, and R. Barbieri. Revealing real-time emotional responses: A personalized assessment based on heartbeat dynamics. *Scientific Reports*, 4:1–13, 2014. ISSN 20452322. doi: 10.1038/srep04998.

[93] R. Harper and J. Southern. A bayesian deep learning framework for end-to-end prediction of emotion from heartbeat. *IEEE Transactions on Affective Computing*, pages 1–1, 2020. doi: 10.1109/TAFFC.2020.2981610.

[94] M. Murugappan and S. Murugappan. Human emotion recognition through short time electroencephalogram (eeg) signals using fast fourier transform (fft). In *2013 IEEE 9th International Colloquium on Signal Processing and its Applications*, pages 289–294, 2013. doi: 10.1109/CSPA.2013.6530058.

[95] J. Kim and E. Andre. Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2067–2083, 2008. doi: 10.1109/TPAMI.2008.26.

[96] Y. Yaslan and C. E. Faculty. Emotion recognition via random forest and galvanic skin response: Comparison of time based feature sets, window sizes and wavelet approaches. In *2016 Medical Technologies National Congress*, pages 1–4, 2016. doi: 10.1109/TIPTEKNO.2016.7863130.

[97] L. Bi and F. xin'an. Emotion recognition from eeg based on bayesian networks. *Energy Procedia*, 11:278–285, Jan. 2011. doi: 10.1016/j.egypro.2011.10.242.

[98] J. L. Qiu, W. Liu, and B. L. Lu. Multi-view emotion recognition using deep canonical correlation analysis. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11305 LNCS(September): 221–231, 2018. ISSN 16113349. doi: 10.1007/978-3-030-04221-9_20.

[99] J. Pinto. Continuous Biometric Identification on the Steering Wheel. Master's thesis, University of Porto, 2017.

[100] P. Tate. *Seeley's Principles of Anatomy and Physiology*. The name of the publisher, McGraw-Hill, New York, NY, 2 edition, 2009. ISBN 978–0–07–337819–0.

[101] E. N. Marieb and K. Hoehn. *Human Anatomy & Physiology*. Pearson, Glenview, IL, 9 edition, 2013. ISBN 978-0-321-74326-8.

[102] V. C. Scanlon and T. Sanders. *Essentials of Anatomy and Physiology*. F. A. Davis Company, Philadelphia, PA, 5 edition, 2007. ISBN 978-0-8036-1546-5.

[103] A. Kennedy, D. D. Finlay, D. Guldenring, R. Bond, K. Moran, and J. McLaughlin. The Cardiac Conduction System: Generation and Conduction of the Cardiac Impulse. *Critical Care Nursing Clinics of North America*, 28(3):269–279, 2016. ISSN 08995885. doi: 10.1016/j.cnc.2016.04.001.

[104] M. Sampson and A. McGrath. Understanding the ECG. Part 1: Anatomy and physiology. *British Journal of Cardiac Nursing*, 10(11):548–554, 2015. ISSN 1749-6403. doi: 10.12968/bjca.2015.10.11.548.

[105] J. G. Betts, P. DeSaix, E. Johnson, J. E. Johnson, O. Korol, D. H. Kruse, B. Poe, J. A. Wise, M. Womble, and K. A. Young. *Anatomy and physiology*, chapter 19. OpenStax College, Rice University, Houston, Texas, 9 edition, 2017. ISBN 781947172043 1947172042.

[106] M. Elgendi, B. Eskofier, S. Dokos, and D. Abbott. Revisiting QRS detection methodologies for portable, wearable, battery-operated, and wireless ECG systems. *PLoS ONE*, 9(1), 2014. ISSN 19326203. doi: 10.1371/journal.pone.0084018.

[107] A. Oosterom, R. Hoekema, and G. Uijen. Geometrical factors affecting the interindividual variability of the ECG and the VCG. *Journal of electrocardiology*, 33 Suppl:219–227, Feb. 2000. doi: 10.1054/jelc.2000.20356.

[108] R. Hoekema, G. J. Uijen, and A. Van Oosterom. Geometrical aspects of the interindividual variability of multilead ECG recordings. *IEEE Transactions on Biomedical Engineering*, 48(5):551–559, 2001. ISSN 00189294. doi: 10.1109/10.918594.

[109] F. Agrafioti, F. M. Bui, and D. Hatzinakos. Secure telemedicine: Biometrics for remote and continuous patient verification. *Journal of Computer Networks and Communications*, 2012, 2012. ISSN 20907141. doi: 10.1155/2012/924791.

[110] B. J. A. Schijvenaars. *Intra-individual Variability of the Electrocardiogram computerized EeG analysis*. PhD thesis, University of Rotterdam, 2000.

[111] F. Agrafioti, J. Gao, and D. Hatzinakos. Heart Biometrics: Theory, Methods and Applications. *Biometrics*, 2011. doi: 10.5772/18113.

[112] J. Catalano. *Guide to ECG Analysis*. J.B. Lippincott, 1993. ISBN 9780781729307.

[113] J. Sohn, S. Yang, J. Lee, Y. Ku, and H. C. Kim. Reconstruction of 12-lead electrocardiogram from a three-lead patch-type device using a LSTM network. *Sensors (Switzerland)*, 20(11): 1–13, 2020. ISSN 14248220. doi: 10.3390/s20113278.

[114] 12-lead ECG placement guide with illustrations. https://www.cablesandsensors.com/pages/12-lead-ecg-placement-guide-with-illustrations. (Accessed on 01/17/2021).

[115] R. Macleod and B. Birchler. ECG Measurement and Analysis. http://www.sci.utah.edu/~macleod/bioen/be6000/labnotes/ecg/l3-ecg.pdf, 2009.

[116] E. Frank. An Accurate, Clinically Practical Systen For Spatial Vectorcardiography. *Advances in internal medicine*, 6:91–131, 1954. doi: 10.1097/00000441-196205000-00018.

[117] A. Burns, B. R. Greene, M. J. McGrath, T. J. O'Shea, B. Kuris, S. M. Ayer, F. Stroiescu, and V. Cionca. SHIMMER^{TM} - A wireless sensor platform for noninvasive biomedical research. *IEEE Sensors Journal*, 10(9):1527–1534, 2010. ISSN 1530437X. doi: 10.1109/JSEN.2010.2045498.

[118] Z. Zhang and D. Wei. A new ECG identification method using Bayes' theorem. In *TENCON 2006 - 2006 IEEE Region 10 Conference*, pages 1–4, 2006. doi: 10.1109/TENCON.2006.344146.

[119] R. Subramanian, J. Wache, M. K. Abadi, R. L. Vieriu, S. Winkler, and N. Sebe. Ascertain: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, 9(2):147–160, 2018. ISSN 19493045. doi: 10.1109/TAFFC.2016.2625250.

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

[121] I. M. Revina and W. R. S. Emmanuel. A Survey on Human Face Expression Recognition Techniques. *Journal of King Saud University - Computer and Information Sciences*, 2018. ISSN 1319-1578. doi: 10.1016/j.jksuci.2018.09.002.

[122] J. Pan and W. J. Tompkins. A Real-Time QRS Detection Algorithm. *IEEE Transactions on Biomedical Engineering*, BME-32(3):230–236, 1985. doi: 10.1109/TBME.1985.325532.

[123] M. Murugappan, S. Murugappan, and S. Z. Bong. Frequency band analysis of electrocardiogram (ecg) signals for human emotional state classification using discrete wavelet transform (dwt). *Journal of physical therapy science*, 25:753–759, Jul. 2013. doi: 10.1589/jpts.25.753.

[124] S. Jerritta, M. Murugappan, K. Wan, and S. Yaacob. Emotion recognition from facial emg signals using higher order statistics and principal component analysis. *Journal of the Chinese Institute of Engineers*, 37(3):385–394, 2014. doi: 10.1080/02533839.2013.799946.

[125] C. Izard. Emotion theory and research: Highlights, unanswered questions, and emerging issues. *Annual review of psychology*, 60:1–25, Aug. 2008. doi: 10.1146/annurev.psych.60.110707.163539.

[126] S. Chen, L. Zhang, F. Jiang, W. Chen, J. Miao, and H. Chen. Emotion Recognition Based on Multiple Physiological Signals. *Zhongguo yi liao qi xie za zhi = Chinese journal of medical instrumentation*, 44(4):283–287, 2020. ISSN 16717104. doi: 10.3969/j.issn.1671-7104.2020.04.001.

[127] E. Alickovic and Z. Babic. The effect of denoising on classification of ECG signals. *2015 25th International Conference on Information, Communication and Automation Technologies, ICAT 2015 - Proceedings*, 2015. doi: 10.1109/ICAT.2015.7340540.

[128] J. A. Domínguez-Jiménez, K. C. Campo-Landines, J. C. Martínez-Santos, E. J. Delahoz, and S. H. Contreras-Ortiz. A machine learning model for emotion recognition from physiological signals. *Biomedical Signal Processing and Control*, 55:101646, 2020. ISSN 17468108. doi: 10.1016/j.bspc.2019.101646.

[129] K. N. Minhad, S. H. M. Ali, and M. B. I. Reaz. A design framework for human emotion recognition using electrocardiogram and skin conductance response signals. *Journal of Engineering Science and Technology*, 12(11):3102–3119, 2017. ISSN 18234690.

[130] J. Mathew, P. Patra, D. Pradhan, and A. Kuttyamma. Preface. *Communications in Computer and Information Science*, 305:V–VI, Jan. 2012. doi: 10.1007/978-3-642-32112-2_13.

[131] S. Z. Bong, M. Murugappan, and S. Yaacob. Analysis of electrocardiogram (ECG) signals for human emotional stress classification. *Communications in Computer and Information Science*, 330 CCIS(September 2013):198–205, 2012. ISSN 18650929. doi: 10.1007/978-3-642-35197-6_22.

[132] A. Mikuckas, I. Mikuckiene, A. Venckauskas, E. Kazanavicius, R. Lukas, and I. Plauska. Emotion recognition in human computer interaction systems. *Elektronika ir Elektrotechnika*, 20(10):51–56, 2014. doi: 10.5755/j01.eee.20.10.8878.

[133] C. Kappeler-Setz. *Multimodal Emotion and Stress Recognition*. PhD thesis, ETH Zurich, 2012.

[134] P. J. Bota, C. Wang, A. L. Fred, and H. Placido Da Silva. A Review, Current Challenges, and Future Possibilities on Emotion Recognition Using Machine Learning and Physiological Signals. *IEEE Access*, 7:140990–141020, 2019. doi: 10.1109/ACCESS.2019.2944001.

[135] P. Hamilton. Open source ecg analysis. *Comput Cardiol*, 29:101 – 104, 10 2002. doi: 10.1109/CIC.2002.1166717.

[136] W. Zong, T. Heldt, G. B. Moody, and R. G. Mark. An open-source algorithm to detect onset of arterial blood pressure pulses. *Computers in Cardiology*, 30(October):259–262, 2003. doi: 10.1109/cic.2003.1291140.

[137] I. I. Christov. Real time electrocardiogram QRS detection using combined adaptive threshold. *BioMedical Engineering Online*, 3:1–9, 2004. ISSN 1475925X. doi: 10.1186/1475-925X-3-28.

[138] H.-w. Guo, Y.-s. Huang, C. Lin, and J. Chien. 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)*, pages 274–277, 2016. doi: 10.1109/BIBE.2016.40.

[139] S. Sriramprakash, P. V. D, and O. V. R. Murthy. Stress Detection in Working People. *Procedia Computer Science*, 115:359–366, 2017. ISSN 1877-0509. doi: 10.1016/j.procs.2017.09.090.

[140] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams. Deep affect recognition from r-r intervals. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, UbiComp '17, page 754–762, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450351904. doi: 10.1145/3123024.3125608.

[141] Y. Xu and G.-Y. Liu. A method of emotion recognition based on ecg signal. In *2009 International Conference on Computational Intelligence and Natural Computing*, volume 1, pages 202–205, 2009. doi: 10.1109/CINC.2009.102.

[142] L. Xun and G. Zheng. Ecg signal feature selection for emotion recognition. *TELKOMNIKA Indonesian Journal of Electrical Engineering*, 11, Jan. 2013. doi: 10.11591/telkomnika.v11i3.2215.

[143] H. Uyarel, E. Okmen, N. Cobanoğlu, A. Karabulut, and N. Cam. Effects of anxiety on QT dispersion in healthy young men. *Acta Cardiologica*, 61(1):83–87, 2006. ISSN 00015385. doi: 10.2143/AC.61.1.2005144.

[144] H. T. M. Ping Gong. Emotion recognition based on the multiple physiological signals. In *2016 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, pages 140–143, 2016. doi: 10.1109/RCAR.2016.7784015.

[145] S. Walter, S. Gruss, K. Limbrecht-Ecklundt, H. C. Traue, P. Werner, A. Al-Hamadi, N. Diniz, G. M. da Silva, and A. O. Andrade. Automatic pain quantification using autonomic parameters. *Psychology and Neuroscience*, 7(3):363–380, 2014. ISSN 19833288. doi: 10.3922/j.psns.2014.041.

[146] W. Wei, Q. Jia, Y. Feng, and G. Chen. Emotion Recognition Based on Weighted Fusion Strategy of Multichannel Physiological Signals. *Computational Intelligence and Neuroscience*, 2018(1), 2018. ISSN 16875273. doi: 10.1155/2018/5296523.

[147] Y.-l. Hsu, J.-s. Wang, and W.-c. Chiang. Automatic ecg-based emotion recognition in music listening. *IEEE Transactions on Affective Computing*, 11(1):85–99, 2020. doi: 10.1109/TAFFC.2017.2781732.

[148] N. Hjortskov, D. Rissén, A. K. Blangsted, N. Fallentin, U. Lundberg, and K. Søgaard. The effect of mental stress on heart rate variability and blood pressure during computer work. *European Journal of Applied Physiology*, 92(1-2):84–89, 2004. doi: 10.1007/s00421-004-1055-z.

[149] S. Boonnithi and S. Phongsuphap. Comparison of heart rate variability measures for mental stress detection. *Computing in Cardiology*, 38(January 2011):85–88, 2011. ISSN 23258861.

[150] J. Marín-Morales, J. L. Higuera-Trujillo, A. Greco, J. Guixeres, C. Llinares, E. P. Scilingo, M. Alcañiz, and G. Valenza. Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors. *Scientific Reports*, 8(1):1–15, 2018. ISSN 20452322. doi: 10.1038/s41598-018-32063-4.

[151] M. Ménard, P. Richard, H. Hamdi, B. Daucé, and T. Yamaguchi. Emotion recognition based on heart rate and skin conductance. *PhyCS 2015 - 2nd International Conference on Physiological Computing Systems, Proceedings*, pages 26–32, Jan. 2015. doi: 10.5220/0005241100260032.

[152] J. Taelman, S. Vandeput, I. Gligorijević, A. Spaepen, and S. Huffel. Time-frequency heart rate variability characteristics of young adults during physical, mental and combined stress in laboratory environment. *Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, 2011:1973–6, Aug. 2011. doi: 10.1109/IEMBS.2011.6090556.

[153] M. M, W. Khairunizam, S. Yaacob, and J. Selvaraj. Electrocardiogram-based emotion recognition system using empirical mode decomposition and discrete fourier transform. *Expert Systems*, 31, Mar. 2013. doi: 10.1111/exsy.12014.

[154] H. Ferdinando, T. Seppänen, and E. Alasaarela. Enhancing emotion recognition from ECG signals using supervised dimensionality reduction. *ICPRAM 2017 - Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods*, 2017-January (January):112–128, 2017. doi: 10.5220/0006147801120118.

[155] J. Selvaraj, M. Murugappan, K. Wan, and S. Yaacob. Classification of emotional states from electrocardiogram signals: A non-linear approach based on hurst. *BioMedical Engineering Online*, 12(1):1–18, 2013. ISSN 1475925X. doi: 10.1186/1475-925X-12-44.

[156] C. Schubert, M. Lambertz, R. A. Nelesen, W. Bardwell, J. B. Choi, and J. E. Dimsdale. Effects of stress on heart rate complexity-A comparison between short-term and chronic stress. *Biological Psychology*, 80(3):325–332, 2009. ISSN 03010511. doi: 10.1016/j.biopsycho.2008.11.005.

[157] C. Xiefeng, Y. Wang, S. Dai, P. Zhao, and Q. Liu. Heart sound signals can be used for emotion recognition. *Scientific Reports*, pages 1–11, 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-42826-2.

[158] H. Abdi and L. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:433 – 459, 07 2010. doi: 10.1002/wics.101.

[159] M. R. B. Clarke, R. O. Duda, and P. E. Hart. Pattern Classification and Scene Analysis. *Journal of the Royal Statistical Society. Series A (General)*, 137(3):442, 1974. ISSN 00359238. doi: 10.2307/2344977.

[160] B. Schölkopf, A. Smola, and K. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10(5):1299–1319, 1998. doi: 10.1162/089976698300017467.

[161] A. Globerson and S. Roweis. Metric learning by collapsing classes. *Advances in Neural Information Processing Systems*, pages 451–458, 2005. ISSN 10495258. doi: 10.5555/2976248.2976305.

[162] T. Dissanayake, Y. Rajapaksha, R. Ragel, and I. Nawinne. An ensemble learning approach for electrocardiogram sensor based human emotion recognition. *Sensors (Switzerland)*, 19 (20):1–24, 2019. ISSN 14248220. doi: 10.3390/s19204495.

[163] D. Bzdok, M. Krzywinski, and N. Altman. Machine learning: supervised methods. *Nature Methods*, 15(1):5–6, 2018. ISSN 1548-7091. doi: 10.1038/nmeth.4551.

[164] A. Goel and S. Mahajan. Comparison: KNN & SVM Algorithm. *International Journal for Research in Applied Science & Engineering Technology (IJRASET)*, 887(December): 2321–9653, 2017. ISSN 2321-9653.

[165] S. Patel and A. Patel. Deep Leaning Architectures and its Applications A Survey. *International Journal of Computer Sciences and Engineering*, 6(6):1177–1183, 2018. doi: 10.26438/ijcse/v6i6.11771183.

[166] J. F. Kolen and S. C. Kremer. Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies. *A Field Guide to Dynamical Recurrent Networks*, Nov. 2010. doi: 10.1109/9780470544037.ch14.

[167] L. Santamaria-Granados, M. Munoz-Organero, G. Ramirez-Gonzalez, E. Abdulhay, and N. Arunkumar. Using Deep Convolutional Neural Network for Emotion Detection on a Physiological Signals Dataset (AMIGOS). *IEEE Access*, 7:57–67, 2019. ISSN 21693536. doi: 10.1109/ACCESS.2018.2883213.

[168] P. Sarkar and A. Etemad. Self-Supervised Learning for ECG-Based Emotion Recognition. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2020-May:3217–3221, 2020. ISSN 15206149. doi: 10.1109/ICASSP40776.2020.9053985.

[169] P. Sarkar and A. Etemad. Self-supervised ECG representation learning for emotion recognition. *arXiv*, pages 1–13, 2020. ISSN 23318422. doi: 10.1109/TAFFC.2020.3014842.

[170] M. N. Dar, M. U. Akram, S. G. Khawaja, and A. N. Pujari. Cnn and lstm-based emotion charting using physiological signals. *Sensors (Switzerland)*, 20(16):1–26, 2020. ISSN 14248220. doi: 10.3390/s20164551.

[171] L. Chao, J. Tao, M. Yang, Y. Li, and Z. Wen. Long short term memory recurrent neural network based multimodal dimensional emotion recognition. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, pages 65–72, Oct. 2015. doi: 10.1145/2808196.2811634.

[172] M. Kächele, P. Thiam, G. Palm, F. Schwenker, and M. Schels. Ensemble methods for continuous affect recognition: Multi-modality, temporality, and challenges. *AVEC 2015 - Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, co-Located with MM 2015*, pages 9–16, 2015. doi: 10.1145/2808196.2811637.

[173] S. Chen and Q. Jin. Multi-modal dimensional emotion recognition using recurrent neural networks. *AVEC 2015 - Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, co-Located with MM 2015*, pages 49–56, 2015. doi: 10.1145/2808196.2811638.

[174] J. Lin, S. Pan, C. Lee, and S. Oviatt. An explainable deep fusion network for affect recognition using physiological signals. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 2069–2072, Nov. 2019. ISBN 978-1-4503-6976-3. doi: 10.1145/3357384.3358160.

[175] M. Gjoreski. Deep Ensembles for Inter-Domain Arousal Recognition. In *AffComp@IJCAI*, pages 52–65, 2018.

[176] P. Kawde and G. K. Verma. Multimodal affect recognition in V-A-D space using deep learning. In *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, pages 890–895, 2017. doi: 10.1109/SmartTechCon.2017.8358500.

[177] Siddharth, T. P. Jung, and T. J. Sejnowski. Multi-modal Approach for Affective Computing. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2018-July:291–294, 2018. ISSN 1557170X. doi: 10.1109/EMBC.2018.8512320.

[178] L. Chao, B. Zhongtian, L. Linhao, and Z. Zhao. Exploring temporal representations by leveraging attention-based bidirectional LSTM-RNNs for multi-modal emotion recognition. *Information Processing and Management*, 57(3):102185, 2020. ISSN 03064573. doi: 10.1016/j.ipm.2019.102185.

[179] S. Oh, J. Y. Lee, and D. K. Kim. The design of CNN architectures for optimal six basic emotion classification using multiple physiological signals. *Sensors (Switzerland)*, 20(3): 1–17, 2020. ISSN 14248220. doi: 10.3390/s20030866.

[180] A. Saeed, T. Ozcelebi, and J. Lukkien. Multi-task self-supervised learning for human activity detection. *Proceedings of the Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(2), Jun. 2019. doi: 10.1145/3328932.

[181] L. Jing, X. Yang, J. Liu, and Y. Tian. Self-supervised spatiotemporal feature learning via video rotation prediction, 2019.

[182] M. Kocabas, S. Karagoz, and E. Akbas. Self-supervised learning of 3d human pose using multi-view geometry, 2019.

[183] P. Sarkar. load_model/saved_model · master · pritam sarkar (17ps21) / self-supervised ecg representation learning · gitlab. https://code.engineering.queensu.ca/17ps21/SSL-ECG/-/tree/master/load_model/saved_model. (Accessed on 06/07/2021).

[184] J. Ribeiro Pinto, J. Cardoso, and A. Lourenço. *Deep Neural Networks For Biometric Identification Based On Non-Intrusive ECG Acquisitions*, pages 217–234. CRC Press, 11 2019. ISBN 9780815393641. doi: 10.1201/9781351013437-11.

[185] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997. ISSN 00043702. doi: 10.1016/s0004-3702(96)00034-3.

[186] T. Tong, R. Wolz, Q. Gao, R. Guerrero, J. V. Hajnal, and D. Rueckert. Multiple instance learning for classification of dementia in brain mri. *Medical Image Analysis*, 18(5):808–818, 2014. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2014.04.006.

[187] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2003.

[188] L. Romeo, A. Cavallo, L. Pepa, N. Berthouze, and M. Pontil. Multiple instance learning for emotion recognition using physiological signals. *IEEE Transactions on Affective Computing*, pages 1–1, 2019. doi: 10.1109/TAFFC.2019.2954118.

[189] S. Duffner and C. Garcia. Multiple Instance Learning for Training Neural Networks under Label Noise. In *International Joint Conference on Neural Networks (IJCNN)*, Glasgow, United Kingdom, July 2020. Virtual conference.

[190] F. Karim, S. Majumdar, H. Darabi, and S. Chen. Lstm fully convolutional networks for time series classification. *IEEE Access*, 6:1662–1669, 2018. doi: 10.1109/ACCESS.2017.2779939.

[191] J. Yang, X. Huang, H. Wu, and X. Yang. Eeg-based emotion classification based on bidirectional long short-term memory network. *Procedia Computer Science*, 174:491–504, 2020. ISSN 1877-0509. doi: https://doi.org/10.1016/j.procs.2020.06.117. 2019 International Conference on Identification, Information and Knowledge in the Internet of Things.

[192] X. Yuan, L. Li, and Y. Wang. Nonlinear dynamic soft sensor modeling with supervised long short-term memory network. *IEEE Transactions on Industrial Informatics*, 16(5):3168–3176, 2020. doi: 10.1109/TII.2019.2902129.

[193] A. Zeyer, P. Doetsch, P. Voigtlaender, R. Schlüter, and H. Ney. A comprehensive study of deep bidirectional lstm rnns for acoustic modeling in speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2462–2466, 2017. doi: 10.1109/ICASSP.2017.7952599.