Fear Classification using Affective Computing with Physiological Information and Smart-Wearables

by

Jose Ángel Miranda Calero

A dissertation submitted by in partial fulfillment of the requirements for the degree of Doctor of Philosophy in

Electrical Engineering, Electronics and Automation

Universidad Carlos III de Madrid

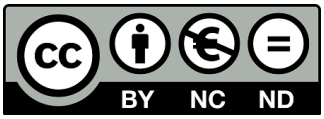
PhD Advisor(s):

Celia López-Ongil
Marta Portela García

Supervisor:

Celia López-Ongil

April, 2022



To my cats.

Acknowledgements

En el año 2015 decidí embarcarme de nuevo en el ámbito académico. Antes de hacerlo, la primera llamada que realicé para coger impulso fue a David Puerta, mi amigo, mi maestro, y mi mentor académico. Sin ese empujón, no estaría donde estoy hoy. En su día me enseñaste que el rigor y la pasión pueden ir de la mano. Gracias por darme esa y otras muchas lecciones tan valiosas que me han servido, me sirven y servirán.

Esta tesis no sería posible ni habría llegado a buen puerto sin la supervisón de Celia López y Marta Portela, mis supervisoras. La primera me ha enseñado a ser creativo a la vez que ingeniero, me ha guiado en momentos fáciles y en momentos no tan fáciles, me ha permitido equivocarme, y me ha regalado mucho de su tiempo. Gracias por ser mi mentora científica. La segunda, la cual siento como hermana mayor científica y ejemplo a seguir académicamente, me ha enseñado a no rendirme, me ha dado ánimos en los momentos difíciles, y también me ha regalado mucho de su tiempo. De ambas he aprendido lo mucho que aún me queda por aprender.

Agradecer también a todas las personas integrantes del Departamento de Tecnología Electrónica por su apoyo durante estos años. En especial a los Técnicos de Laboratorio con los que he tenido el placer de coincidir y trabajar. Sin vuestro trabajo y ayuda, estaríamos vendidos.

El equipo UC3M4Safety y el proyecto EMPATIA-CM. Dos cosas que han formado parte de esta tesís y viceversa. Gracias a todas las personas involucradas en ambas, habéis dado sentido a esta trabajo.

Gracias a toda mi familia. A mi padre y a mi madre, soy el reflejo de toda vuestra paciencia, perseverancia y educación que habéis mantenido a lo largo de vuestra vida y nos habéis sabido trasmitir a mi hermana y a mi.

Finalmente, muchas gracias a mi pareja, Patricia. Este trabajo también es tuyo. Te ha tocado la peor parte, pero ya hemos llegado. Gracias por ser paciente y por enseñarme a ser paciente. Gracias por estar conmigo y aguantar carros y carretas. Espero poder ser capaz de devolverte algún día, aunque sea una pequeña parte, el tiempo invertido.

Jose Angel Miranda
Marzo 2022

"iustum et tenacem propositi virum
non civium ardor prava iubentium,
non vultus instantis tyranni
mente quatit solida neque Auster,
dux inquieti turbidus Hadriae,
nec fulminantis magna manus Iovis:
si fractus illabatur orbis,
impavidum ferient ruinae"
Horacio, Carmina III, 3, 1-8.

Published and Submitted Contents

- **Papers in Journals**
  - J. A. Miranda, M. F. Canabal, L. Gutiérrez-Martín, J. M. Lanza-Gutierrez, M. Portela-García, C. López-Ongil, "Fear Recognition for Women Using a Reduced Set of Physiological Signals," Sensors, 2021, 21(5), 1587.
    This document is partially included in Chapter 4.
  - J. A. Miranda, M. F. Canabal, L. Gutiérrez-Martín, J. M. Lanza-Gutierrez, C. López-Ongil, "Edge computing design space exploration for heart rate monitoring," VLSI Integration, 2022.
    This document is partially included in Chapter 5.
  - J. A. Miranda, E. Rituerto-González, M. F. Canabal, A. R. Bárcenas, J. M. Lanza-Gutiérrez, C. Pelaez-Moreno, and C. López-Ongil, "Bindi: Affective internet of things to combat gender-based violence," IEEE Internet of Things, 2022, manuscript submitted for publication.
    This document is partially included in Chapter 6.
  - J. A. Miranda, A. P. Montoro, C. López-Ongil, and J. Andreu-Pérez, "FT2F-SQA: Few-shot type-2 fuzzy-based subject-invariant ppg quality assessment for extreme edge physiological monitoring," IEEE TIM, 2022, manuscript submitted for publication.
    This document is partially included in Chapter 5.
  - M. F. Canabal, J. A. Miranda, A. P. Montoro, I. P. Garcilópez, S. P. Álvarez, E. G. Ares, and C. López-Ongil, "Design and validation of an efficient and adjustable GSR sensor for emotion monitoring," IEEE Sen-

sors, 2022, manuscript in progress.

- **Papers in Conference Proceedings**
  - J. A. Miranda, A. Vaskova, M. Portela-García, M. García-Valderas and C. López-Ongil, "On-line testing of sensor networks: A case study," IEEE 23rd International Symposium on On-Line Testing and Robust System Design (IOLTS), 2017, pp. 201-202, doi: 10.1109/IOLTS.2017.8046218.
  - J. A. Miranda, M. F. Canabal, J. M. Lanza-Gutiérrez, M. Portela-García, C. López-Ongil, Teresa Riesgo Alcaide, "Meaningful Data Treatment from Multiple Physiological Sensors in a Cyber-Physical System," DCIS 2017: XXXII Conference on Design of Circuits and Integrated Systems, 22nd-24th November 2017, Barcelona (Spain). pp. 100-104.
  - J. A. Miranda, M. F. Canabal, M. Portela García, C. Lopez-Ongil, "Embedded emotion recognition: Autonomous multimodal affective internet of things," Proceedings of the cyber-physical systems workshop, 2018, (Vol. 2208, pp. 22-29).
  - J. A. Miranda, R. Marino, J.M Lanza-Gutierrez, Teresa Riesgo, M. Garcia-Valderas, C. Lopez-Ongil, "Embedded emotion recognition within cyber-physical systems using physiological signals," In 2018 Conference on design of circuits and integrated systems (DCIS) (pp. 1-6). IEEE.
  - J. A. Miranda, M. F. Canabal, J. M. Lanza-Gutiérrez, M. P. García and C. López-Ongil, "Toward Fear Detection using Affect Recognition," 2019 XXXIV Conference on Design of Circuits and Integrated Systems (DCIS), 2019, pp. 1-4, doi:10.1109/DCIS201949030.2019.8959852.
  - E. Rituerto-González, J. A. Miranda, M. F. Canabal, J. M. Lanza-Gutiérrez, C. Peláez-Moreno, C. López-Ongil, "A hybrid data fusion architecture for bindi: A wearable solution to combat gender-based violence," In International Conference on Multimedia Communications, Services and Security (pp. 223-237). Springer, Cham., 2020.
  - J. A. Miranda, M. F. Canabal, L. Gutiérrez-Martín,J. M. Lanza-Gutiérrez, C. López-Ongil, "A Design Space Exploration for Heart Rate Variability in a Wearable Smart Device," In 2020 XXXV Conference on Design of Circuits and Integrated Systems (DCIS) (pp. 1-6). IEEE.

  - M. F. Canabal, J. A. Miranda, J. M. Lanza-Gutiérrez, A. P. Garcilópez, C. López-Ongil, "Electrodermal Activity Smart Sensor Integration in a Wearable Affective Computing System," In 2020 XXXV Conference on Design of Circuits and Integrated Systems (DCIS) (pp. 1-6). IEEE.
  - J. A. Miranda, A. P. Montoro, C. López-Ongil, and J. Andreu-Pérez, "Towards Type-2 Fuzzy-Based PPG Quality Assessment for Physiological Monitoring," IEEE WCCI 2022, manuscript submitted for publication.
- **Other publications of the author**
  - J. A. Miranda, E. Rituerto-González, L. Gutiérrez-Martín, C. Luis-Mingueza, M. F. Canabal, A. R. Bárcenas, C. López-Ongil, "WEMAC: Women and Emotion Multi-modal Affective Computing dataset," arXiv preprint arXiv:2203.00456, 2022.
  - M. Á. Blanco Ruiz, L. Gutiérrez Martín, J. Á. Miranda Calero, M. F. Canabal Benito, E. Romero Perales, C. Sainz de Baranda Andujar, R. San Segundo Manuel, D. Larrabeiti López, C. Peláez-Moreno, and C. López Ongil, "UC3M4Safety Database - List of Audiovisual Stimuli Annotations," 2021, [Online]. Available: `https://doi.org/10.21950/CXAAHR`
  - C. S. de Baranda Andújar, M. B. Ruiz, J. Á. Miranda, L. Gutierrez-Martín, M. F. Canabal, R. San Segundo, C. López-Ongil, "Perspectiva de género y social en las STEM: La construcción de sistemas inteligentes para detección de emociones,", Sociología y tecnociencia: Revista digital de sociología del sistema tecnocientífico, 11(1), 83-115, 2021.
  - E. Rituerto-González, J. A. Miranda, M. F. Canabal, J. M. Lanza-Gutiérrez, C. Peláez-Moreno, C. López-Ongil, "A Hybrid Data Fusion Architecture for BINDI: A Wearable Solution to Combat Gender-Based Violence," In: Dziech A., Mees W., Czyżewski A. (eds) Multimedia Communications, Services and Security. MCSS 2020. Communications in Computer and Information Science, vol 1284. Springer, Cham. `https://doi.org/10.1007/978-3-030-59000-0_17`.
  - Andres Russu, Antonio J. de Castro, Francisco Cortes, Celia López-Ongil, Marta Portela, Ernesto Garcia, José A. Miranda, Manuel F. Canabal, Ignacio Arruego, Javier Martinez-Oter, Fernando López, "A light com-

pact and rugged IR sensor for space applications," Proc. SPIE 11129, Infrared Sensors, Devices, and Applications IX, 1112907 (9 September 2019); https://doi.org/10.1117/12.2529846.

Other research merits

- **Awards**
  - Best poster award, XXXIV Conference on Design of Circuits and Integrated Systems (DCIS), 2019.
  - YERUN Research Mobility Award 2021.
- **Patents**
  - Sistema y método para for determinar un estado emocional de un usuario (System and method for determining a user's emotional state), Modelo de Utilidad ES1269890 (Expedition date: 20/09/2021). Owned by UC3M / UPM.
- **Participation in research projects**
  - Ciencia y Tecnología para la caracterización in situ de la atmósfera de Marte. Desarrollo del Instrumento Dust Sensor para la misión EXOMARS'18 de ESA/IKI. Fases A/B y C/D. ESP2015-67624-R, MINECO.
  - Comunidad de Madrid, protEcción integral de las víctimas de violencia de género Mediante comPutación AfecTIva multimodAl (EMPATÍA-CM), Y2018/TCS-5046.
- **Supervised and/or Co-Supervised Bachelor Thesis**
  - Ismael Granados Neira, "Desarrollo e implementación de drivers para aplicación IoT," Internship coordinator, University Carlos III de Madrid, 2016.
  - Adrian Neira Robledo, "Integración de cámaras IR en una red de sensores inalámbrica para recintos críticos," Bachelor Thesis, University Carlos III de Madrid, 2017.
  - George Sebastian Roma, "Investigación, comparación e implementación

de RTOSs para plataformas orientadas a IoT," Bachelor Thesis, University Carlos III de Madrid, 2017.

- Oscar Escobar Muñoz, "Diseño Hardware de un Sistema Eficiente para la Medida del Consumo de Energía en Nodos Inalámbricos de Redes de Sensores," Bachelor Thesis, University Carlos III de Madrid, 2017.
- Javier Plaza Arenas, "Implementación de técnicas de Machine Learning para la detección de emociones en FPGA," Bachelor Thesis, University Carlos III de Madrid, 2018.
- A. Aranzana Sánchez, "Implementación de Técnicas de Extracción de Características para el Reconocimiento de Emociones Usando Sensores Fotopletismográficos," Bachelor Thesis, University Carlos III de Madrid, 2020.
- W. Allay Bakhtaoui, "Implementación de Técnicas de Extracción de Características para el Reconocimiento de Emociones Usando Sensores para la Conductividad de la Piel," Bachelor Thesis, University Carlos III de Madrid, 2020.
- Pedro Ruiz Perez, "Estudio e implementación embebida de filtros digitales con señales fotopletismográficas," Bachelor Thesis, University Carlos III de Madrid, 2020.
- L. Velasco Gonzalez, "Diseño e implementación de un sistema de recuperación fisiológica para experimentos de reconocimiento de emociones," Bachelor Thesis, University Carlos III de Madrid, 2021.
- Alexandru Stoica Stoica, " Estudio e implementación de técnicas de procesado orientadas a la eliminación de artefactos con señales fotopletismográficas para bindi," Bachelor Thesis, University Carlos III de Madrid, 2021.
- Juan Marcos Torero, "Diseño y desarrollo de pendiente inteligente con micrófono y sensor de pulso para su integración en el sistema BINDI," Bachelor Thesis, University Carlos III de Madrid, 2021.
- Mario Iañez Diaz, "Diseño e integración de Front-End para ECG de dos electrodos en Bindi," Bachelor Thesis, University Carlos III de Madrid, 2021.

- Andrés Lucas Rodriguez, "Diseño e implementación de un sensor textil inteligente para Bindi", Bachelor Thesis, University Carlos III de Madrid, 2021.

- **Supervised and/or Co-Supervised Master Thesis**
  - Jesus Jose Garcia de Cuerva Camacho, "Eliminación de artefactos de movimiento en señales fotopletismográficas para sistemas portables orientados a la detección de emociones," Master Thesis. University Carlos III de Madrid, 2019.
  - F. Adrián Hernández Gant, "Diseño de modelos de aprendizaje para detección de miedo en Bindi," Master Thesis. University Carlos III de Madrid, 2021.
  - Marta Subirán Adrados, "Design Space Exploration for the Multi-modal Data Fusion Architectures of Bindi," Master Thesis. University Carlos III de Madrid, 2021.

Biography

Jose A. Miranda Calero received the BSc degree in Industrial Electronics and Automation Engineering in 2013, and the MSc in Electronic Systems and Applications Engineering in 2016 with honors, both from Universidad Carlos III of Madrid. He is currently a PhD candidate at the Microelectronic Design and Applications (DMA) research group, which belongs to University Carlos III de Madrid. From 2013 to 2015, he has worked as an Embedded Software Engineer in different countries within Europe for public and private sectors. His research field comprises a wireless sensor, networks, wearable design, development and integration for safety applications, affective computing implementation into edge computing devices, and hardware acceleration. Regarding the development of wearable technology for safety applications, his main contribution relates to the design of Bindi, which is a new autonomous, smart, inconspicuous, connected, edge-computing-based, and wearable solution able to detect and alert when a user is under a violent situation employing emotion recognition using the physiological and auditory signals of the user. Bindi is being developed within the UC3M4Safety Team. Moreover, he has also participated in other projects related to space applications, such as DS-EXOMARS20.

Contents

List of Figures

List of Tables

List of Abbreviations

SoC	System-On-Chip
PPG	Photoplethysmography
ECG	Electrocardiogram
EDA	Electrodermal Activity
GSR	Galvanic Skin Response
SKT	Skin Temperature
VR	Virtual Reality
PA	Pleasure-Arousal model
PAD	pleasure-arousal-dominance model
PTSD	Post-Traumatic Stress Disorder
VR	Virtual Reality
SAM	Self-Assessment Manikins
ANS	Autonomous Nervous System
SNS	Sympathetic Nervous System
PNS	Parasympathetic Nervous System
BVP	Blood Volume Pulse
LED	Light Emitting Diode
BPM	Beats Per Minute
MAR	Motion Artifact Removal
SQA	Signal Quality Assessment
SCL	Skin Conductance Level
SCR	Skin Conductance Response
ERSCR	Event-Related Skin Conductance Response
NSSCR	Nonspecific Skin Conductance Response
SMNA	Sudomotor Nerve Activity
RLSD	Regularized Least-Squares Detrending
PSD	Power Spectral Density
DFT	Discrete Fourier Transform
FFT	Fast Fourier Transform
RP	Recurrence Plot
SFS	Sequential Forward Selection
SBE	Sequential Backward Elimination
SVM	Support Vector Machine
KNN	K-Nearest Neighbours

RF Random Forests
ENS Ensemble methods
PCA Principal Component Analysis
SMBO Sequential Model-Based Optimisation
CV Cross-Validation
LOO Leave-One-Out
LOTO Leave-One-Trial-Out
LOSO Leave-One-Subject-Out
LASO Leave-hAlf-Subject-Out
TEAP Toolbox for Emotional feAture extraction from Physiological signals
HRV Heart Rate Variability
FIR Finite Impulse Response
IIR Infinite Impulse Response
AGC Automatic Gain Control
ACC Accuracy
DSE Design Space Exploration
IBI Inter-Beat-Interval
MSE Multi-Scale Entropy
SMOTE Synthetic Minority Over-sampling TEchnique
AUC Area Under the Curve
RBF Radial Basis Function
mrMR Maximum Relevance — Minimum Redundancy
MDI Mutual Information Difference
ANOVA Analysis of Variance
DFA Detrended Fluctuation Analysis
BLE Bluetooth Low Energy
DTW Dynamic Time Warping
IoT Internet of Things
SQA Signal Quality Assessment
GPS Global Positioning System
SoC System on Chip
PCB Printed Circuit Board
COTS Commercial-Off-The-Shell
SDK Software-Development-Kit
HAL Hardware Abstraction Layer

Abstract

Among the 17 Sustainable Development Goals proposed within the 2030 Agenda and adopted by all of the United Nations member states, the fifth SDG is a call for action to effectively turn gender equality into a fundamental human right and an essential foundation for a better world. It includes the eradication of all types of violence against women. Focusing on the technological perspective, the range of available solutions intended to prevent this social problem is very limited. Moreover, most of the solutions are based on a panic button approach, leaving aside the usage and integration of current state-of-the-art technologies, such as the Internet of Things (IoT), affective computing, cyber-physical systems, and smart-sensors. Thus, the main purpose of this research is to provide new insight into the design and development of tools to prevent and combat Gender-based Violence risky situations and, even, aggressions, from a technological perspective, but without leaving aside the different sociological considerations directly related to the problem. To achieve such an objective, we rely on the application of affective computing from a realist point of view, i.e. targeting the generation of systems and tools capable of being implemented and used nowadays or within an achievable time-frame. This pragmatic vision is channelled through: 1) an exhaustive study of the existing technological tools and mechanisms oriented to the fight Gender-based Violence, 2) the proposal of a new smart-wearable system intended to deal with some of the current technological encountered limitations, 3) a novel fear-related emotion classification approach to disentangle the relation between emotions and physiology, and 4) the definition and release of a new multi-modal dataset for emotion recognition in women.

Firstly, different fear classification systems using a reduced set of physiological sig-

nals are explored and designed. This is done by employing open datasets together with the combination of time, frequency and non-linear domain techniques. This design process is encompassed by trade-offs between both physiological considerations and embedded capabilities. The latter is of paramount importance due to the edge-computing focus of this research. Two results are highlighted in this first task, the designed fear classification system that employed the DEAP dataset data and achieved an AUC of 81.60% and a Gmean of 81.55% on average for a subject-independent approach, and only two physiological signals; and the designed fear classification system that employed the MAHNOB dataset data achieving an AUC of 86.00% and a Gmean of 73.78% on average for a subject-independent approach, only three physiological signals, and a Leave-One-Subject-Out configuration. A detailed comparison with other emotion recognition systems proposed in the literature is presented, which proves that the obtained metrics are in line with the state-of-the-art.

Secondly, Bindi is presented. This is an end-to-end autonomous multimodal system leveraging affective IoT throughout auditory and physiological commercial off-the-shelf smart-sensors, hierarchical multisensorial fusion, and secured server architecture to combat Gender-based Violence by automatically detecting risky situations based on a multimodal intelligence engine and then triggering a protection protocol. Specifically, this research is focused onto the hardware and software design of one of the two edge-computing devices within Bindi. This is a bracelet integrating three physiological sensors, actuators, power monitoring integrated chips, and a System-On-Chip with wireless capabilities. Within this context, different embedded design space explorations are presented: embedded filtering evaluation, online physiological signal quality assessment, feature extraction, and power consumption analysis. The reported results in all these processes are successfully validated and, for some of them, even compared against physiological standard measurement equipment. Amongst the different obtained results regarding the embedded design and implementation within the bracelet of Bindi, it should be highlighted that its low power consumption provides a battery life to be approximately 40 hours when using a 500 mAh battery.

Finally, the particularities of our use case and the scarcity of open multimodal

datasets dealing with emotional immersive technology, labelling methodology considering the gender perspective, balanced stimuli distribution regarding the target emotions, and recovery processes based on the physiological signals of the volunteers to quantify and isolate the emotional activation between stimuli, led us to the definition and elaboration of Women and Emotion Multi-modal Affective Computing (WEMAC) dataset. This is a multimodal dataset in which 104 women who never experienced Gender-based Violence that performed different emotion-related stimuli visualisations in a laboratory environment. The previous fear binary classification systems were improved and applied to this novel multimodal dataset. For instance, the proposed multimodal fear recognition system using this dataset reports up to 60.20% and 67.59% for ACC and F1-score, respectively. These values represent a competitive result in comparison with the state-of-the-art that deal with similar multi-modal use cases.

In general, this PhD thesis has opened a new research line within the research group under which it has been developed. Moreover, this work has established a solid base from which to expand knowledge and continue research targeting the generation of both mechanisms to help vulnerable groups and socially oriented technology.

Chapter 1

Introduction

1.1 Context and motivation

Gender-based Violence constitutes a violation of human rights and fundamental freedoms recognised by the 1993 United Nations Declaration on the Elimination of Violence against Women [11]. This declaration provides a clear and complete definition of what this type of violence means, which is stated in its first article by considering any act of violence, whether it is physical, sexual, or psychological, based on belonging to the female gender. In 2020, the European Commission expanded such definition and stated that this violence includes the one against women, men and children [12]. Regarding the specific numbers, from 2000 to 2018, more than one in four (27%) ever-partnered women aged between 15 and 49 years had experienced physical or sexual, or both, intimate partner violence since the age of 15 years [13]. This problem is not new, in fact, in the European Union, the first principle of equal treatment for men and women was introduced in 1975 into the Treaty of Rome [14]. However, it is in 2007 through the Treaty of Lisbon [15] that the European Community included this principle among the values and objectives of the Union. Since then, different territories within Europe have taken these steps as a base building block for their Gender-based Violence laws. Regardless of such efforts, there was still a need for a set of community standards or rules to be applicable that targeted this problem. Thus, the Council of Europe Convention on preventing and combating violence against women and domestic violence, also known as the Istanbul Convention, was approved in 2011 and entered into force later in 2014 [16]. This convention established a common framework or instrument from which differ-

ent standards on prevention, protection, prosecution and provision of services to respond to the needs of victims and those at risk are set out. To date, all members have signed the convention and 35 out of 47 of them have ratified it, although in July 2021, Turkey became officially the first and only country to withdraw from it. Note that this country was among the initial precursors of this agreement. Moreover, the Istanbul Convention created a monitoring mechanism responsible for controlling, reporting, and evaluating legislative and other measures taken by the ratifier states. However, the implementation of all the recommendations by the convention is not always a straightforward task, since it depends on the resources of each state. For that reason, different European Union funding programs were launched to ease the implementation of these actions (DAPHNE, PROGRESS, REC), but always taking a mutual learning approach by leveraging the message within and outside the European Community, as it is conceived as a worldwide problem [17]. Along with these agreements, conventions, and funding programmes, different pacts and organisations were also created, such as the European Pact for Gender Equality (2011-2020) and the European Institute for Gender Equality. These actions have been accompanied by European regulations, which aim to safeguard the rights of victims from a legal point of view (EU 606/2013, 2012/29/EU).

Focusing on Spain, where this research has been developed, it must be highlighted the unanimous approval in 2004 of Organic Law 1/2004 through which this country became a fundamental reference in the world for the way of facing this problem. Specifically, it is a comprehensive law against Gender-based Violence, which also considers this type of violence to be that wielded on persons dependent on a woman when they are abused to cause harm to her. Moreover, Spain was one of the first countries to sign the Istanbul Convention in 2011, to be later ratified in 2014. Another key date in the national road map was the ratification of the National Agreement against Gender-based Violence by the different Groups in the National Parliament, the Regional Governments and Local Entities in December 2017. As for many other countries in the European Community, Spain is divided into autonomous regions which, apart from what refers to national legislative application, have their own regional laws and regulations. For instance, the Autonomous Community of Madrid possess a dilated experience regarding Gender-based Violence policies and le-

gal actions. One of the greatest achievements of this community related to this topic was the approval of the first regional Organisation Act No. 5/2005 of 20 December on comprehensive protection measures against Gender-based Violence. Moreover, this community also created a specific regional institutional body to assess and evaluate the integration of such policies, this was conceived as the regional Gender-based Violence observatory in 2003 (decree 256/2003, 27 November). Lastly, in 2016 they introduced a comprehensive strategy for preventing and combating Gender-based Violence, which consisted of a series of measures and actions to be developed from that year till 2021 and reflected the commitment of the Madrid government to fight toward the eradication of this problem. Despite all these national and regional measures, the Government office against Gender-based Violence accounted a total of 1117 women killed since 2003 by October 2021, Fig. 1-1.

Based on the above facts, we can conclude that Gender-based Violence is an emergency problem that leads society to deal with it by using different perspectives and adopting a multidisciplinary approach. For instance, from a sociological point of view, education and information awareness regarding the prevention and combat of violence against women is essential. Moreover, the technological perspective is also a fundamental aspect related to the development of new emerging technology that eases the creation of new platforms for preventing and responding to gender violence [18]. In fact, these and other perspectives as legal, psychological, and medical, amongst others, are linked and work together toward day-to-day solutions to combat this problem. This multidisciplinary claim is strongly supported by a wide range of professionals that work closely and personally with victims (law enforcement agents, judges, and psychologists) [19]. However, they identify two main drawbacks of the current public instruments by agreeing that more and better-organised efforts should be invested into the prevention mechanisms or tools and into the training for the professionals who deal directly with the victims towards avoiding more harmful re-victimisation.

Despite the institutional effort, developing solutions by applying a multidisciplinary focus to create safer communities is a challenging task. However, all the mentioned perspectives need to be engaged and cooperate in a narrower way to combat Gender-based Violence efficiently. Due to the digital transformation that some countries in

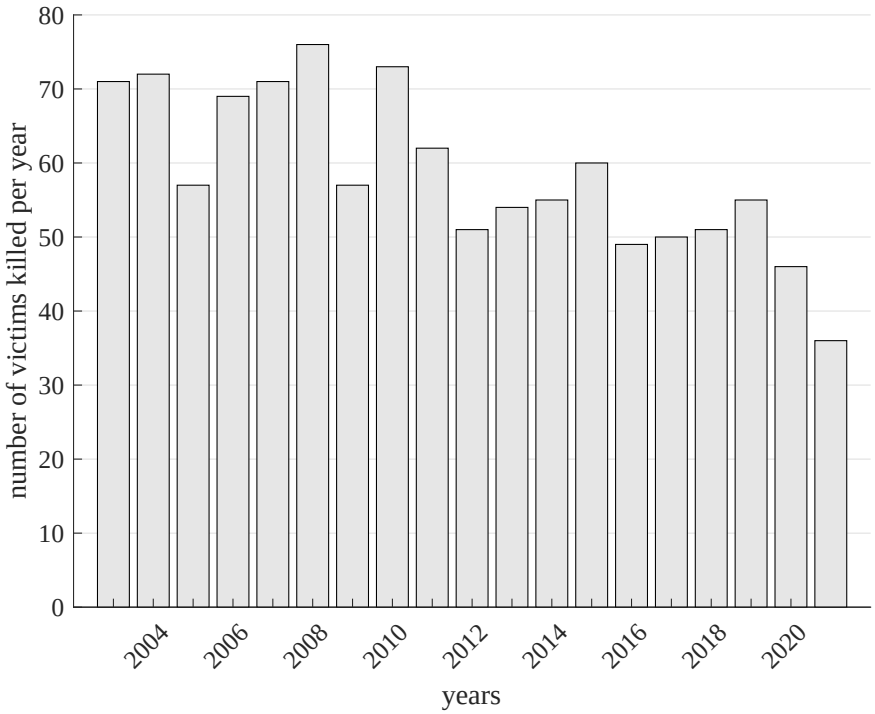


Figure 1-1: Total Gender-based Violence victims killed from 2003 to October 2021. Data provided by [1].

Europe are undergoing, they are trying to take advantage of the development of new technologies to provide services to communities, from which, some of them are intended to deal with this problem in question. For instance, in Spain, different services have been launched, such as VioGen [20], ATENPRO [21], and COMETA [7]. First, VioGen allows to estimate the risk level faced by a Gender-based Violence victim and determines the adequate type and degree of protection for her. This risk level is updated continuously according to her legal and social situation. This tool is the result of intensive research by the Spanish Home Affairs Department with various Spanish university research groups with experts in psychology, criminology, and sociology. Second, ATENPRO is a service that provides a direct and 24-7 hotline to the Spanish law enforcement agencies through a panic button. Specifically, the victim is given a mobile device that allows continuous communication at any moment and circumstance. Such communication is handle by a specialised telephonic assistance centre, where specifically trained attendants give an adequate response to handle this type of situations in real-time. Finally, COMETA is a system conceived as a set of telematic control devices adopted when a restraining order is issued against the aggressor. In this case, both the victim and the aggressor are given a geolocation device with basic voice and data telecommunication capabilities to communicate with

the control centre. The aggressor must also wear a lightweight bracelet-like radio-frequency device that connects to the geolocation devices. Although COMETA offers a technological solution for combating Gender-based Violence, its limited battery life and outdated technology present a high false-positive rate [22, 23], apart from the risk of harassment for the victims.

In addition to the public government organisations efforts, the private sector is also encouraged worldwide to bring a solution to the Gender-based Violence problem. In fact, different private initiatives continually come out with ideas to prevent and avoid this problem. For instance, the XPrize Foundation launched in 2018 a $1 million worldwide competition to challenge teams around the world to leverage technology for empowering women to respond to sexual aggression. The goal of this competition was to develop a technological solution capable of triggering emergency alerts autonomously, transmitting information to a network of community responders, and being as affordable as possible, all within 90 seconds. The seven finalists used wearable technology with the latest wireless communication protocols linked with different responders or even with law enforcement agencies. Only one of them included affective computing capabilities to their devices to seamlessly track emotional threat levels by using cardiac physiological information. Monitoring such information is proven as a solid emotional indicator [24].

Considering all the reviewed information and targeting the generation of new prevention and combating mechanisms, a new autonomous, smart, inconspicuous, connected, edge-computing, and wearable-ready tool able to detect and alert when a user is under a Gender-based Violence situation, might be exploited. On this basis, the research work described in this document is focused on providing a smart technological solution to help dealing with the stated problem. This system will be hereinafter referred to as BINDI along the entire document, and it has been developed by the UC3M4Safety group at University Carlos III of Madrid. Specifically, this research is focused on the design, development, and implementation of one of three devices that make up the system, which is a smart bracelet using embedded affective computing based on physiological monitoring for detecting fear-related emotional states. The other two devices are a smart pendant and a smartphone application. The former captures audio-on-demand, while the latter performs phys-

iological and physical data fusion and handles the emergency alarms to be sent to a trusted responders network or even to the law enforcement agencies. The nature of the problem to be addressed made this work be derived by a multidisciplinary approach, gathering knowledge from Gender Studies, Electronics, Telematics, Physiology, Speech and Audio Technologies, and Affective Computing.

1.2 Scope of this dissertation

This research aims to provide new insight into the development of tools to prevent and avoid Gender-based Violence risky situations and, even, aggressions, from a technological perspective, but without leaving aside the different sociological considerations related to the problem.

From a theoretical point of view, this work proposes a new way of using physiological signals and emotion recognition to provide autonomous, wearable, and inconspicuous solutions to protect vulnerable people. The goal in that aspect is to disentangle the relationship between physiological signals and fear-related emotions, providing alternatives to emotion recognition classification systems already proposed in the literature, new physiological monitoring wearable-ready system architectures, new sensor integration and embedded implementation into wearable devices, new techniques to mitigate physiological motion artifacts noise, and performing an analytical study of the entire proposed solution.

From a practical point of view, different fear binary emotion recognition systems were provided based on openly available databases that contain non-acted evoked emotions for a set of volunteers. Moreover, a new wearable hardware solution for gender violence detection was developed and implemented based on the ARM Cortex-M® processor family, using three of the most inconspicuous physiological sensors, Photoplethysmography (PPG), EDA, and SKT, and low power wireless communications. This device forms part of the BINDI system, which has been developed along with the UC3M4Safety group. The requirements for the complete system are the lowest power consumption possible, an inconspicuous and wearable integration of all devices and components, and the lowest computational time for the different digital processing architectures to achieve the fastest response time possible. Finally, a new database using immersive stimuli and specific stimuli oriented to the

Gender-based Violence use case was generated. The latter is particularly relevant as the generated database is unique in the literature.

Specifically, the goals of this research are the following:

- Proposing a new approach to detect fear-related emotions making use of the different emotional theories and physiological affective indicators.
- Deriving new wearable-ready fear-related detection systems by using open databases that have used physiological signals for emotion recognition.
- Dealing with the physiological behaviour (quasi-stationary, non-stationary and non-linearity) and proposing new digital processing techniques to take it into account for rapid-inference fear-related systems.
- Analysing and studying different integration constraints to be considered in wearable affective computing systems.
- Designing a new wearable hardware solution to deploy the fear-related detection system architectures proposed. At this point, we would face the hardware and embedded software implementation toward an inconspicuous, autonomous, low power, wireless and connected solution.
- Comparing the obtained results with similar published architectures and commercial solutions for gender violence prevention.
- Generating a new database focused on the specific targeted Gender-based Violence use case that gathers physiological, physical and emotional responses to immersive stimuli.

1.3 Document outline

The document is divided into three parts. The first part reviews the relationship between emotions and physiological signals by researching the different emotional theories and physiological affective indicators. The general framework of the databases used in the literature, which deal with emotion recognition by using physiological signals, are also analysed. In the second part, we present the application of the theory reviewed in the first part to the proposal and analysis of a new fear-related emotion recognition system. The third part presents both hardware and embedded software results of the edge-computing system developed for fear binary recognition. Finally, the new database for emotion recognition focused on fear de-

tection is also presented in the last part. Moreover, wearable integration constraints and physiological dynamics to be considered are given and analysed along the entire document.

Thus, the outline of the document is the following:

Part I

Chapter 2 describes basic and advanced notions needed to get a good understanding of the topics of this research. Specifically, emotional theories, human emotion classification methodologies, tools for emotion elicitation, and physiological affective indicators quantification are studied. All of these topics are supported by references from the state-of-the-art that will help understand the original content provided in the following Chapters.

Chapter 3 concentrates on providing an in-depth analysis regarding the structure and experimental procedures used for the generation of databases designed for emotion recognition. Moreover, each part of the whole data processing chain for the affective computing system design task using such databases is also detailed and explained.

Part II

Chapter 4 deals with one of the main purposes of the dissertation. This is the design and validation of novel fear recognition systems based on a reduced set of physiological signals. Different public available databases are selected to design two main fear binary emotion recognition systems. The encountered limitations of such databases are spotted and taken into consideration for the work presented in Chapter 6. Moreover, the reported results on this Chapter are compared against the current state-of-the-art.

Part III

Chapter 5 details the design and integration process for a new wearable hardware solution to deploy parts of the fear-related detection system architectures proposed in Chapter 4. On the one hand, this new wearable solution is contextualised by analysing current technology being applied towards Gender-based Violence prevention and combat. On the other hand, the design and integration challenges are comprehensively detailed and explained for both hardware and software perspectives.

Chapter 6 elaborates upon one of the main contributions of this research. This is the generation of a novel multi-modal dataset, WEMAC consisting of experiments performed in a laboratory environment with only women volunteers. Moreover, the different affective computing architectures proposed and presented in Chapter 4 are employed using the data gathered throughout this dataset. Finally, a multi-modal approximation by means of physiological and speech data fusion is also reported to provide a first baseline to be considered for future works.

Part IV

Chapter 7 concludes this research and provides some suggestions regarding the possible extension of this work in the near future.

Part I

Human emotions, physiological signals and affective computing

Chapter 2

Emotion classification and physiological quantification

This Chapter is based on four essential topics, which are needed for the development of this research: emotional theories, human emotion classification methodologies, tools for emotion elicitation, and physiological affective indicators quantification. First off, the main emotional theories are chronologically analysed and evaluated targeting the specific use case of this research. This is done by linking their emotional order of activation with the effects regarding the elaboration of affective computing systems. The different human emotions classification methodologies are also presented and analysed by reviewing their advantages and disadvantages and exploiting their relationship to the previous emotional theories. This is followed by a comprehensive analysis into the effects in emotion modulation by intrapersonal factors such as personality traits, cognition, attention and gender bias, which delves into providing personal contextualisation within the human emotion classification methodologies. Subsequently, a new pragmatic approach to merge those human emotion classification methods is presented to be later applied into the generated models towards narrowing the identification or recognition of fear emotion. Secondly, the different tools to evoke emotions are compared. Finally, a reduced set of physiological signals and their relation with emotions and emotional models are presented and analysed. Although it has been proven that many other physiological signals ensure specific emotion-related characteristics, they cannot be acquired by inconspicuous sensors to be used daily. Thus, the emotional-related information

that each of the most inconspicuous physiological sensors can provide, as well as their current wearable development and integration status and current challenges, is thoroughly studied.

2.1 Assumptions and Definitions

Before going into details about the different topics to be tackled within this Chapter, some assumptions and definitions must be provided.

- First of all, emotions are a compound of behavioural reactions, subjective cognitive processes, and physiological changes, mostly triggered by emotional stimuli [25].
- Emotional stimulus is referred to as any type of material or process through which specific emotion is elicited to a person. They derive into specific emotional responses.
- Emotional responses can be quantified or measured using subjective self-reports, physical and/or physiological information, and any type of data coming from the person being under the emotion elicitation and gathered during such process.
- Emotion recognition databases are those that use emotional stimuli under a specific presentation or interaction method to gather different emotional responses. All that information can be further used to train intelligent affective computing systems.
- The affective computing systems use all the elements above to generate a trained emotion recognition system.

2.2 Emotional Theories

Despite the emotional theory considered, it is agreed that emotions intervene directly in the adjustment of our response to an external stimulus. However, there is not a common agreement regarding the order in which the compounds of emotions are triggered upon the reception of such stimulus. Different emotional theories along the history have been postulated trying to tackle this process, some based on and some refuting the previous ones or predecessors. For instance, the emotional theories of Darwin [26], James-Lange [27], and Cannon-Bard [28], which are respectively

preceding one another, were the three leading theories of emotion before 1950 and each of them follow a different activation order. Darwin was the first one to try figuring out the origin of emotions and their triggering mechanisms. Although his work did not consider the physiological information within the emotional reaction process, he proposed three different principles which were profoundly linked to the cognitive process of emotions. Amongst those principles, the first one (principle of serviceable habits), which is based on the association between different actions and specific states of mind, is highlighted as it served as a building block for some of the following cognitive-emotional theories. The James-Lange theory states that emotions become conscious to the person once all the physiological information has been processed within the neocortex, Fig.2-1. On the other hand, the Cannon-Bard theory refutes James-Lange's claiming that physiological response occurs at the same time than emotion reaction and independently of one another.

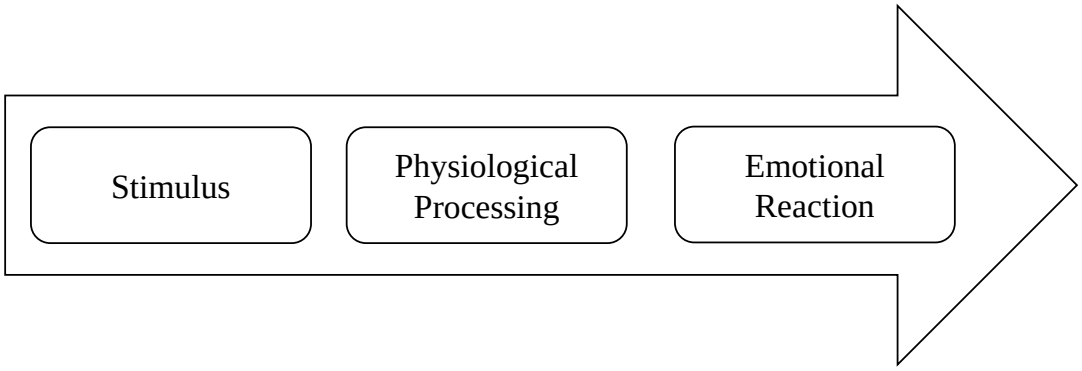


Figure 2-1: Order of activation for the James-Lange emotional theory.

These three theories were highly criticised, as the first two have a lack of empirical evidence and the one from Bard imposes the complete independence of physiological and emotional reactions. Thus, this led to the birth of cognitive emotional theories, in which the context of the situation and our previous experience also directly affects that behavioural response. For instance, the appraisal theory of emotions, mainly developed by Magda Arnold and Richard Lazarus [29], amongst others, was one of the first cognitive emotional theories. It is based on the assumption that emotions are directly determined by our appraisals or evaluations of stimuli, which can cause specific simultaneous physiological and emotional reactions in different people, Fig.2-2. One of the most controversial claims of this theory states that emotions could be originated directly from our own appraisals without the need for physiological

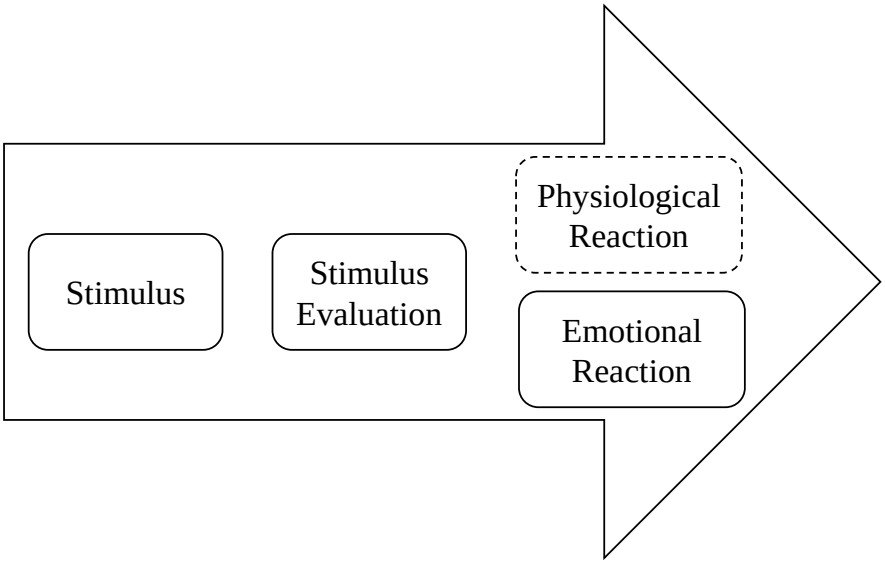


Figure 2-2: Order of activation for the appraisal theory of emotions.

arousal. This fact implies the possible uncorrelated physiological response with respect to the specific external stimulus. Along with the appearance of the appraisal theory of emotions, Albert Ellis introduced the rational-emotive theory [30], which includes the appraisal or evaluation process by claiming that emotions are directly affected by our thoughts or beliefs but it does not neglect the physiological response to a stimulus. Moreover, the latter is preceded by the emotional reaction in contrast to the other theories, Fig. 2-3. Although cognitive theories are widely accepted, much variation is observed within them. Nowadays, the complete role of cognition over emotions is still an open question [31].

Notwithstanding the enormous effort to disentangle the emotion origin paradigm throughout the years, there is still neither an agreed definition for emotion nor an order of activation regarding the different elements involved within. Specifically, in our case, the theory to build this research work on is the rational-emotive theory. This theory allows for the emotional quantification through physiological monitoring and admits the thoughts and beliefs repercussions over the felt emotion. The later is indeed a key factor when dealing with the development of emotion recognition tools to prevent and avoid gender-based violence situations. The life experiences of every gender-based violence victim are different, and the need of not just considering the actual inter-differences between the victims but also the individual intra-differences along the time is essential to provide a better and smarter socially and technologically integrated solution.

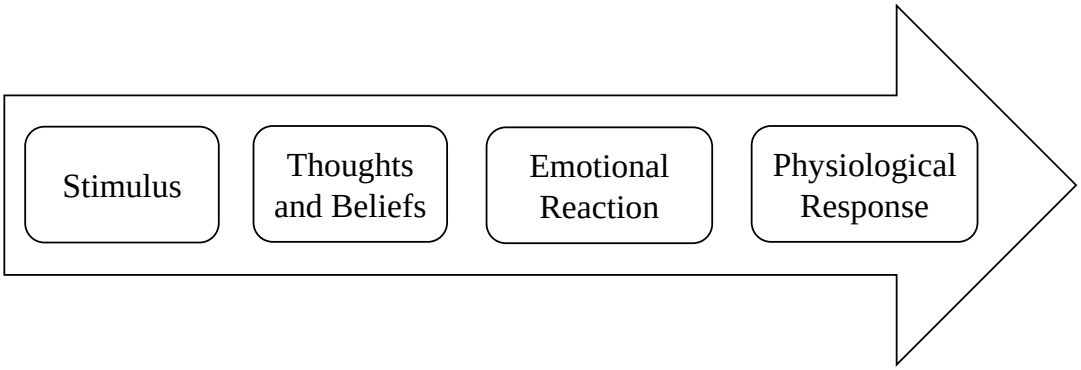


Figure 2-3: Order of activation for the rational-emotive theory.

2.3 Human Emotions Classifications

Within the context of unequivocally categorising and identifying emotions, the literature presents a vast effort to come up with standardised labelling or modelling regarding them [32]. In the affective computing scientific community, this is translated into the identification of the experimented emotion by a person through different data such as facial gestures, voice, posture, and physiological signals. On this bases, mappings between affective states and patterns or variations observed in these modalities have been proposed by different authors. Note that the term modality refers to sources of information. Thus, emotion detection could be defined as a pattern recognition problem [33]. However, as in any recognition problem, the system needs to be taught. The most common learning process in emotion recognition is by using a supervised classifier, which is defined as an oversee learning process through tagged information. This methodology requires to have different samples of information labelled or assigned to the correct emotion. The information gathered from all modalities will be distributed based on the labels or classes assigned to each sample acquired. Therefore, having accurate labels or emotional models will strongly affect the performance of the system designed.

In the next sections, the main human emotion classification methodologies, as well as their advantages and disadvantages, are reviewed. Moreover, some key factors when dealing with emotional experience and personality and how they can affect the emotion labelling process are summarised and analysed. Lastly, the fear-related emotions using the different human emotion classifications methods are connected, which provides a new approach to deal with fear recognition under gender-based violence situations.

2.3.1 Discrete classifications of human emotions

As early as the 19th century, Darwin proposed that emotions were discrete or categorical, i.e. they can be divided into modules such as fear, disgust, anger, and so forth [26]. Although, he did not provide any specification regarding the exact number of those emotions. Since then, different psychologists and physiologists has used the same or similar categorical approach to deal with emotions. This approach is based on the concept of basic emotions, which are universally recognisable and cross-cultural. Thus, it is claimed that each of these basic emotions has different physiological patterns associated, as well as different effects in voice, facial gestures, posture, etc. However, throughout the years, there have been different psychological and physiological theorists who provide lists of primary emotions, each of them based on different criteria to define which are basic emotions and which are not. For instance, Ekman and Friesen [34, 35] provided data to reaffirm the theory exposed by Darwin, and they based most of their early research for basic emotions into unequivocally facial expressions across cultures. Based on this criterion, they claimed the existence of six basic emotions: anger, disgust, fear, surprise, sadness, and joy. In 1972, they travelled to Papua New Guinea and met the Fori tribe. They presented different pictures to the tribe, who were able to identify the six different emotions. Afterwards, they showed images of facial expressions of the people from the Fori tribe, with the same emotions, to people of other nationalities and cultures. They concluded that emotions were correctly interpreted and claimed that emotions are universally recognisable by facial expressions.

Chronologically right after Darwin, some of the main contributions to this human emotion classification approach are summarised in Table 2.1. Among these authors, Robert Plutchik is known for having created the wheel of emotions [37], which was one of the first graphical representations that tried to illustrate how emotions were related from a categorical point of view. Some of the key aspects of Plutchik's model have influenced later proposal for discrete classifications of human emotions. For instance, he introduced the concept of opposite emotions, which can not be experienced at the same time, and even proposed that emotions can be felt with different intensity, which leads up to transforming the wheel into a multidimensional discrete model of emotions. Actually, the most discussed aspect of Plutchik's approach is

Table 2.1: Main categorical models of emotions developed since 19th century.

Author(s)	Claims	Based on
Ekman & Friesen [35]	Six basic emotions: anger, disgust, fear, joy, sadness, and surprise	Universal facial expression
Carroll Izard [36]	Ten basic emotions: anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, and surprise	Motivation of behavior
Robert Plutchik [37]	Eight basic emotions: anger, acceptance, joy, anticipation, fear, disgust, sadness, and surprise	Biological evolution and survival
Nico Frijda [38]	Six basic emotions: desire, happiness, interest, surprise, wonder, and sorrow	Change in action
Oatley & Laird [39]	Five basic emotions: anger, disgust, anxiety, happiness, and sadness	Cognitively based states

the former concept, as many scientists disagree by citing different examples in which opposite emotions can be triggered at the same time. It should be noted that the contributors cited within Table 2.1 do not agree on the number and nature of basic emotions, even some of these authors along the years have modified their stated number of basic emotions. In the case of Ekman, who is considered a pioneer within the emotional research field in the current century, his research was based on the existent evidence of seven basic emotions to which later were included another ten additional enjoyable emotions [40]. Another well-known and influencing psychologist is Nico Frijda, who identified eighteen basic emotions at his early proposed emotional model [38], which afterwards evolved into a total of six [41]. In addition to the number of basic emotions, these proposals differ also with respect to the emotional theory they are based on. As it can be observed, the motivation started by a Darwinian point of view (universal facial expression), which evolved by introducing behavioural and biological factors. Finally, Oatley and Laird introduced the cognitive concept by claiming that emotions are cognitively based states in charge of coordinating quasi-autonomous processes within the nervous system [39]. Note that this latter fact is in line with some of the claims of the appraisal theory of emotions explained in Section 2.2, which admits the possibility of uncorrelated physiological response with respect to the felt emotion.

Regardless of the categorical models' diversity, most of these models include anger, happiness, sadness and fear. However, due to the cultural background of each of the authors, some postulates anger and some postulates rage to refer to the same

emotion, as well when using fear and anxiety. Despite these disagreements, the literature used to refer to these emotions as primary or basic and universal responses to stimuli. Note that there are more categorical models than the ones cited in Table 2.1. Regarding these facts, a survey realised by Ekman in 2015 [42] supposed a clarifying breakthrough advance towards the standardisation of basic emotions. Specifically, 248 scientists into the field were asked, using the same survey, with the goal to obtain any evidence of universality in any facet of emotional theories and categorical models. The highest agreement was retrieved with only five out of eighteen emotions proposed: anger (91%), fear (90%), disgust (86%), sadness (80%), and happiness (76%). The survey is concluded by claiming that, although there is a need for working toward reducing disagreements, there exist an agreement about basic emotions.

2.3.2 Dimensional Classifications of Human Emotions

To alleviate the problem derived from the different categorical terms being applied to the same emotional concept and the analysis of complex emotions by using the combination of different basic emotions lead toward the need for other scales and quantification methods rather than just categorical models. Thus, different authors have introduced what is known as affective state dimensions. For instance, Wundt was the first to introduce the use of two dimensions to classify and identify emotions already in 1896 [43]. He introduced pleasant-unpleasant and low-high intensity, which were used by many other researchers in the next years. Other early relevant author was Osgood [44], who used three different factors to evaluate affective states. These factors were defined as evaluation, activity and potency. Moreover, as commented in the previous section, Plutchik claimed that emotions are felt with different intensity. This fact implies that even authors, who have contributed to the development of categorical models of emotions, needed to assume the existence of some type of dimension to distinguish complex from basic emotions. In this sense, the inclusion of quantitative dimensions allows the creation of a multidimensional space, in which the categorical bias is diminished and basic and complex emotions can be equally identified. Moreover, the self-rating or self-reports of these dimensions after each stimulus presented to the person, as for the self-report when using discrete emotions, takes into account both the cultural differences and previous experiences

of the same stimulus. However, in this multidimensional case, the perfect understanding of the different dimensions and the emotional auto-assessment presents an arduous task.

One of the most used dimensional models is the circumplex model, postulated by Russel [45]. This model is based on two different dimensions, valence or pleasure (P) and arousal (A), which can be interpreted as the modern dimensions of those proposed by Wundt. Specifically, both are conceived to measure different key aspects of the current affective state. Thereby, the valence dimension represents the positive or negative nature of the affective state, while the arousal indicates the excitement or activation given by that affective state. Despite the fact that the circumplex model has been one of the most used dimensional models, the addition of further orthogonal axes leads to a more complete multidimensional space. For instance, Mehrabian [46] introduced dominance as a new emotional dimension and so proposed the pleasure, arousal, and dominance model PAD. Afterwards, this model has proven to be useful for disentangling emotions which are located into the same quadrant for a two-dimensional emotion space (PA). In this regard, Demaree et al. in [47] affirmed that a three-dimensional emotion classification (PAD) is required to identify an affective state properly. They compared the fear-anger distinction using the PA model and the PAD model. As a result, Demaree et al. assured that only dominance can disentangle emotions like fear and anger, associated with submission and dominance respectively.

Although in the last decades, dimensional classifications of human emotions have gained attention, there is still a profound debate going on about the interpretation of these dimensions and how that interpretation is explained and applied. This latter fact is key when trying to compare studies from different researchers who have used the same dimensions but explained them to the volunteers in different ways. Within this context and trying to provide clarity regarding the definition, understanding and explanation of the different dimensions, Bakker et al. [2] linked the PAD model to affective, cognitive and conative responses or the affect, cognition and behaviour model (ABC model of attitudes), Fig. 2-4. They concluded that pleasure, arousal, and dominance can be used together with the ABC model and the distinction between feeling, thinking, and acting respectively, to better understand the

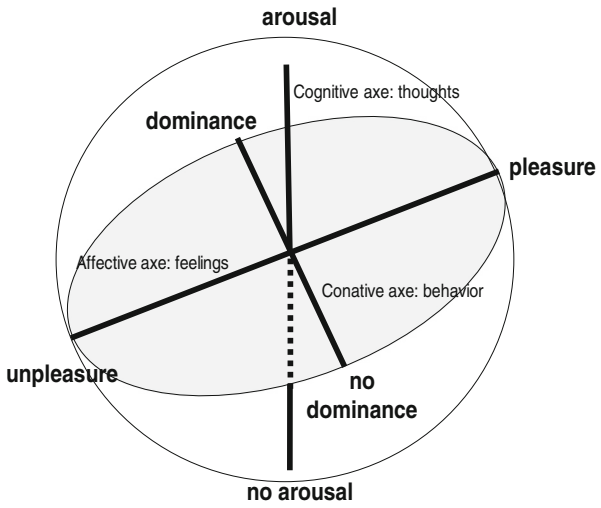


Figure 2-4: Linking between PAD model and ABC model of attitudes by Bakker et al. trying to provide a more clear vision of the original PAD dimensions [2].

original dimensions and describe the subsequent environmental experiences. This last conclusion is key for this research work, as it highlights the need for including the appraisal or evaluation process of the emotion by considering the effects of the external environment. They also indicated the need for additional research focusing on the PAD model to conceive it as a solid and proven dimensional emotion model.

Unlike qualitative emotional information provided by discrete classifications, dimensional classifications give specific quantitative metrics regarding affective states. This can be considered as an advantage at the time to design any automatic emotion recognition system, as self-reported dimensional labels are more specific and more likely to be used for affective computing. However, locating into dimensional coordinates both basic and complex emotions is not an easy task. Different authors have performed studies with a relatively large population with the goal of defining the exact dimensional positions for emotions within dimensional models. For instance, Fontaine et al. [3] used four dimensions to locate the exact space of 24 discrete emotions by using more than 600 participants, 9-point Likert scales for each dimension, and considering three different cultural backgrounds. These emotions were taken form the well known GRID instrument[1], which comprises 144 emotion characteristics representative of the different components of emotions. They achieved to map the 24 emotional terms into their proposed four dimensional schema, Figure 2-5, and pointed out that the optimal number of dimensions depend on what the researchers

[1] unige.ch/cisa/files/7214/9371/2318/Grid_questionnaire_Aug_2013.pdf

are asking or interested in. Nevertheless, they concluded that their study cannot be taken as an emotional experience dimensional representation, assured that two-dimensional models are missing key emotional variation sources such as emotion domain, and encouraged the research community to apply three or more dimensions to properly disentangle the emotion complexity.

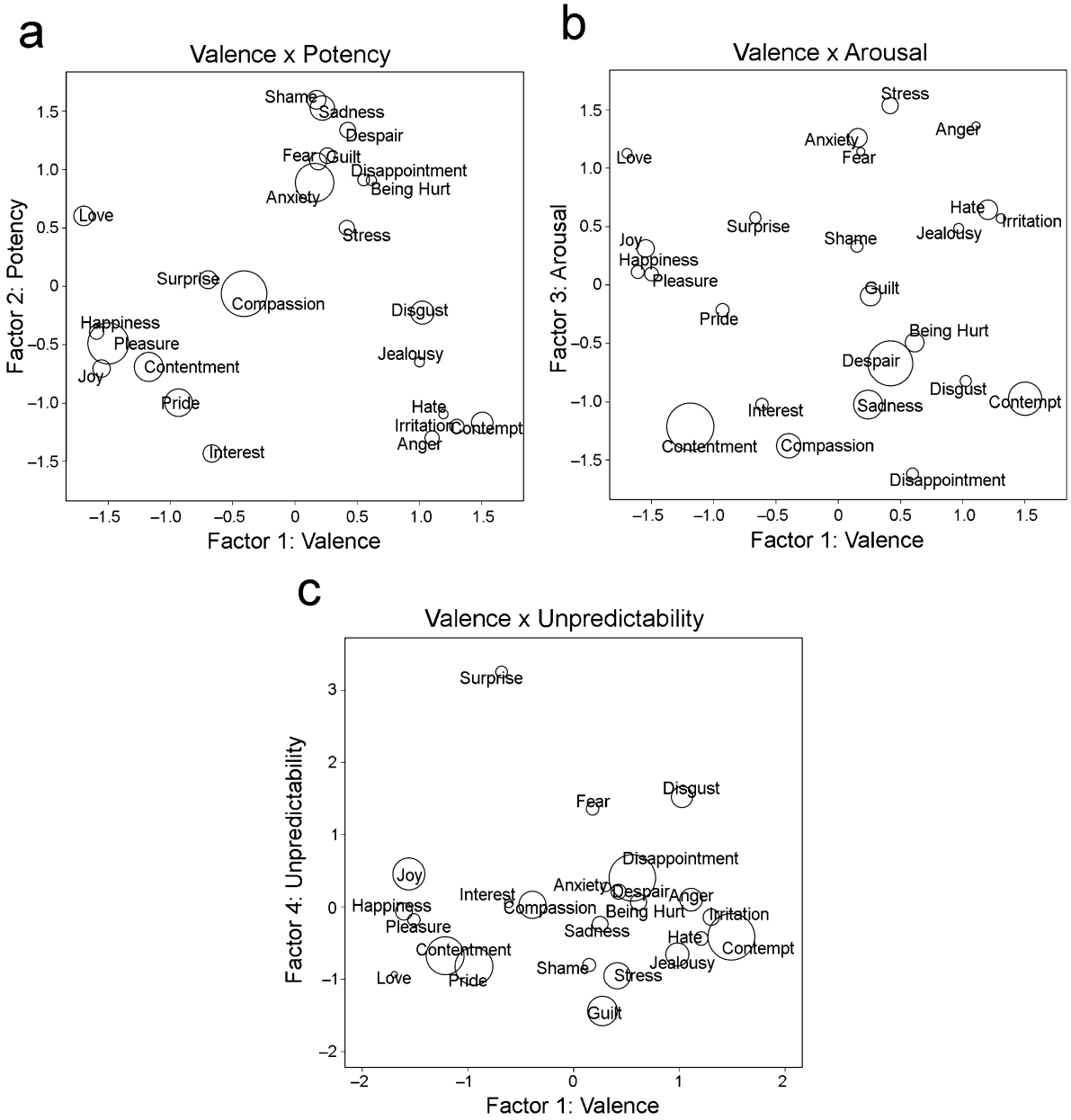


Figure 2-5: The 24 emotional terms mapped into the proposed four dimensional schema by Fontaine et al. in [3].

Disregarding the profound debate taking into consideration both discrete and dimensional meta-theoretical perspectives, there are already researchers proposing to link them together. In fact, the authors in [48] stated that both models exist but each is intended to explain different features of emotions. For instance, they claimed

that the supremacy of one classification method over another will never be assured by the psychology of emotions, which even turns into an advantage when disentangling the behaviour of emotions as it leads to increasing the understanding of emotions by not excluding any emotional perspective or information. Throughout a detailed review, they also affirmed that discrete emotions can be decomposed by including dimensions within them. Although this specific fact was already initially proposed by Plutchik [37] with his multidimensional wheel of emotions and the intensity variation within basic emotions, the combination of these two main classification approaches might be exploited by the literature toward a better psycho-physiology understanding of emotions. Regarding the effect of this issue when designing emotion recognition tools to prevent and avoid gender-based violence situations, the combination of both approaches could leverage a more consistent detection and a better understanding of any kind of fear-related emotion. The specific combination followed in this research is detailed in Section 2.3.4.

2.3.3 Personal traits, cognitive processes, attention and gender bias

In addition to the human emotional theories and emotions classifications mentioned and reviewed in the previous section, the individual thoughts and beliefs are components playing an essential role within the emotional and physiological responses that might be also considered when dealing with the design of any emotion recognition system. Actually, these elements became relevant with the acceptance and rise of cognitive psychology after 1950. However, despite of this fact, the unanswered question that still surrounds many of the human emotional theories is the specific effect that cognition has over the different emotions [49]. This is even fuzzier when cognition converges with appraisal. The latter is based on an automatic association of an affective state (emotional association) either with low or high valence, and it is the core of different human emotional theories as previously commented. From a psychological perspective, the path between a cognitive or evaluation process and an appraisal reaction can derive one from the other and viceversa, as an evaluation can be a rationale from a previous emotional association, and the latter can be also the product of an emotional posterior evaluation [50]. Within this

subjective context, personal traits and stimuli attention and interpretation might strongly affect the affective responses. In the last years some studies tried to link the different main personality five-factors grade (extraversion, neuroticism, openness to experience, agreeableness, and conscientiousness) with respect to daily life emotional processes and changes. For instance, Emma Komulainen et. al. in [51] performed an into-the-wild experiment with 104 university students (18 males, 86 females) following an experience sampling method in which the students reportted different affective metrics by about 10 times per day at semi-random intervals. They observed and concluded that personality features can influence different emotional processes. Specifically, they emphasised that those features strongly affect depressive, anxiety and stress disorders which are linked to the negative affective response to daily life contexts. This fact keeps a deep relationship with different neurophysiological concepts which are directly related to the amount of negative affective load each individual can handle [52]. Such load is known as allostatic load and it is a crucial factor to start understanding the physiological and emotional Gender-based Violence victims profile particularities, as they are subjected to chronic negative situations (fear, panic, stress) which lead up to affective restriction in traumatic contexts with the aim of recovering physiological homeostasis and behavioural balance and protecting her psychological integrity. Moreover, gender differences should be considered and accounted when adding the gender bias to the emotion recognition problem. For instance, it is proven that women are more sensitive to interpersonal expressions during social interactions than men, which is accompanied by a diathesis of mood and even Post-Traumatic Stress Disorder (PTSD) [53–55].

All these components raise different uncertainties that make the design of an intelligent emotion recognition system a task in which the different individual emotional subjective factors might be considered for achieving an optimal performance. Therefore, if an emotion recognition system is developed by using physiological and physical information, the use of the reviewed human emotions classifications methods should be accompanied by different tests or questionnaires to ease the elucidation of the effects produced by cognition, appraisal, attention, personality traits, gender, and age. Although this research work does not directly deal with, quantify or take into consideration such individual subjective factors, a set of questionnaires have

been gathered during the realisation of the UC3M4Safety database. More details are given in Chapter 6.

2.3.4 Fear Mapping within the Human Emotion Classification Methods

The fear emotion is one of the basic emotions that are common throughout most of the different categorical classifications of emotions and even represents a key distinguished emotion for the dimensional classifications when explaining the advantages of such models to deal with dominance-based emotion disentanglement. Due to the targeted application of this research work, the proper understanding of the discrete and dimensional fear bounding is essential. In this sense, Figure 2-6 illustrates an insight regarding this fact. The easiest way to look at fear is by adopting a discrete-like

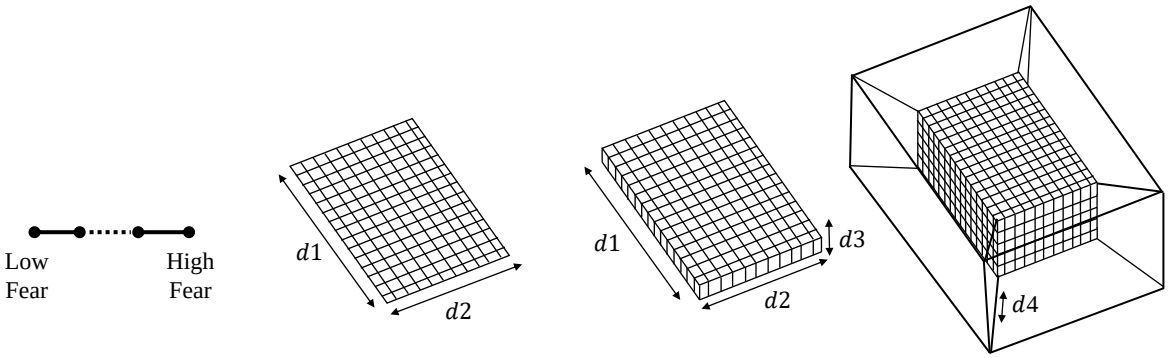


Figure 2-6: From left to right: one-dimensional fear concept (discrete intensity levels), fear contained into two-dimensional space (PA model), three (PAD model) and four (PAD model plus any individually intrinsic dimension) dimensional concepts.

one-dimensional factor form. This method is determined by the number of divisions or levels of fear intensity wanted. Moving onto more dimensions, we have more information to unequivocally determine and characterise the exact sector of fear. For instance, a two-dimensional perception, which can be related to arousal and valence, can define a specific quadrant in which the negative emotions are located. In this case, the number of levels or divisions in the different dimensions directly impact into the exact fear location uncertainty within such quadrant. Thus, the more divisions, the more limited or bounded is the area in which fear can be found. The latter fact presents also a challenging task when the levels of those dimensions are gathered directly from volunteers by using self-reported metrics, as it is not feasible or pragmatic to offer the possibility of choosing a high number of levels. Finally, as commented in Section 2.3.2, the use of three dimensions can benefit the separation

of different emotions that share two of them. Note that adding more dimensions leads to a multidimensional search space, which is translated into a more complex optimisation problem to find the sweet fear spot.

Specifically for this research work, a new merge between discrete and dimensional models for fear detection is used. Thus, taking as a reference the PAD model, we define fear to be located into low valence, high arousal and low dominance. From that specific multidimensional location, a discrete emotional binary mapping is performed by labelling that location as fear and the others as no fear. This relationship is given by the following Heaviside function,

$$\mathbf{H}(x_i) = \Theta(\varepsilon_i - x_i) \tag{2.1}$$

where $\Theta : \mathbf{H} \rightarrow (0, 1)$ and ε_i is the specified fear threshold of dimension i. Note that the final binary mapping output for proposed model is obtained by performing the following logical operation,

$$\mathbf{H}(P, A, D) = \mathbf{H}(P) \wedge \mathbf{H}(A) \wedge \mathbf{H}(D). \tag{2.2}$$

During this research work, all the conducted experiments gathered self-reported emotional labels which were used for the fear binary mapping. In this case, and following the literature [49], the same 1-9 Likert scale was used to rate each of the three dimensions. Therefore, the distinction between low and high levels was done based on half of the scale. Although a limitation is found regarding that this new approach assumes that the entire cube space formed by high arousal, low valence and low dominance, is directly related to the categorical fear emotion, it is the first time that a merge between discrete and dimensional models is done and applied using real data and non-acted emotions. Considering this research as a foundation regarding this aspect, further research might be developed toward a better multidimensional and categorical bounding of the targeted emotion. On this wise, more dimensions could be added to consider individual factors such as personality traits, cognition effect, attention processes, and gender bias. Such dimensions would produce a multidimensional shift of the fear-cube, which can lead to better disentanglement and fear detection.

2.4 Tools and Elements for Scientific Analysis of Human Emotion Responses

One of the main objectives within the affective computing community is the generation of new databases to ease and boost the emotion recognition task. An affective computing or emotion recognition database can be defined as an emotion elicitation experiment with a set of volunteers, who self-report the emotions felt for a specific set of stimuli. Moreover, amongst these essential components, the distinct physiological and physical signals sets gathered during such specific controlled, laboratory-based, experiments are fundamental to further generate emotion recognition models using that information as data. Note that the underlying goal is to disentangle the physiological and physical data patterns and variations observed within those experiments aided by the different labels recollected. On this basis, different tools, elements and methods have been presented and proposed in the literature to provide an effective emotion elicitation within those experiments. In this section, these factors are presented and detailed towards understanding the current state-of-the-art status in this regard. Note that the ones analysed here are applied and used in controlled or laboratory conditions, which constraints the straightforward application of the resultant intelligent systems to in-the-field validation. This latter fact also leads to the need for into-the-wild databases generation. More details are given in Chapter 3 regarding these latter considerations and providing a detailed description and analysis of every component involved into the generation of an emotion recognition database. Moreover, in this section, the different challenges to provide a reliable stimuli labelling ground truth are also addressed and discussed.

As analysed in the previous sections, individual factors are key in emotions. This fact makes it difficult to elicit the same emotion for a group of people being under the same experiment of the database. Although, this is commonly tackled by well defined experimental protocols within a controlled laboratory environment, such personality uncertainties always present a subjective bias introduced by the volunteers when, for example, self-reporting the emotion felt. Generally, we can divide the type of stimuli used in such protocols into two main groups: acted and non-acted. The former is mostly performed by trained actors and actresses, who follow an "emotional

Table 2.2: Review of stimuli type used in controlled laboratory environments.

Stimuli Type	Main Features	Examples
Images	Static stimulus, cognitive-driven, display duration is key, easier emotional identification than others	[58], [59], [60] [61], [62], [63]
Videos	Static stimulus, cognitive and behavioural, display duration is key, can provide more emotional content than images	[64], [65], [66] [67], [68], [10]
Gaming	Dynamic stimulus, cognitive and behavioural, latency between the input of the user and the game reaction is key	[69], [70], [71]
Stress Tests	Static and dynamic stimulus, cognitive, behavioural, and physical, strong agreement in the literature for some of the tests	[72], [73], [74] [75], [76]
VR	Dynamic stimulus, cognitive and behavioural, close to real world scenarios, offers the best ecological validity	[77], [78], [79] [80]

elicitation script" [56]. Regardless of the actors and actresses ability to go deeply into the emotional state requested, this results into a synthetic way of generating affective states that leads to a not fully emotional autonomous response. Thus, non-acted type of stimuli are preferred in the literature. These and their main characteristics are summarised in Table 2.2. There are mainly six different types of non-acted stimuli, which range from imagery up to Virtual Reality (VR). Some of these provide a static feeling by not completely involving the person into the desired emotional environment, while others provide such possibility. Differences with respect to the cognitive, behavioural and physical processes triggered by such stimuli are also found. Amongst all the different stimuli, VR is highlighted as it can offer the closer feeling to real world scenarios which is translated into a high degree of correlation between the research conditions and the emotional phenomenon under study (ecological validity). These facts led the UC3M4Safety team develop a VR environment with 2D and 3D stimuli to be used during the realisation of the different datasets [57]. More details are given in Chapter 6.

Despite the range of the different types of stimuli, one of the biggest challenges involving emotion elicitation toward the design of emotion recognition systems is obtaining a reliable ground truth, i.e. to properly determine what emotions (labels) evoke what stimulus. The assessment of the ground truth is actually one of the most critical parts within the design of those systems [81]. In fact, that process is

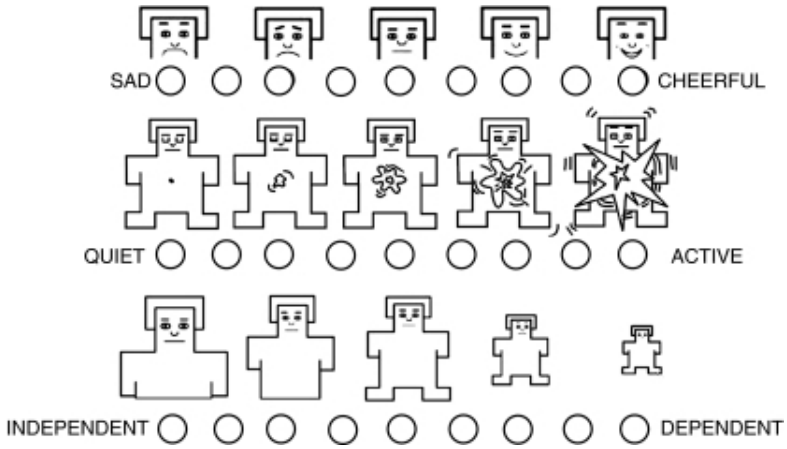


Figure 2-7: Original SAM [4].

of utmost importance which will strongly affect the training and posterior inference performance of the affect-aware system. For instance, when training a machine learning algorithm using the gathered ground truth labels, the information that is fed to such algorithm inherits such label's distribution, which can strongly bias the underlying original data patterns. In order to deal with these type of problems, there are different strategies or methodologies to collect the ground truth that are even used together within the experiment. Thus,the literature has come up with different methods to report reliable emotional self-evaluation data. One of the most used and reliable methodologies to gather ground truth is based on a well-known non-verbal pictorial technique, this are the SAM [4]. The original representation can be seen in Figure 2-7. It is based on the PAD (valence, arousal and dominance, respectively from the first to the third row) space and an 1-9 Likert-scale, in which the middle of the scale is related to a neutral affective state. However, it can be observed that this original depiction shows mostly straight lines and a very masculine-based attitude that can affect the labelling for women. Therefore, the SAM was modified by the UC3M4Safety team in order to provide less gender-bias. Note that this modification was performed based on a panel of gender-based violence experts [82]. The resultant new SAM is shown in Figure 2-8.

To summarise this section, we can state that the emotion recognition databases are needed to generate affective computing systems, but also they are essential to study emotional responses differences based on physical, physiological, gender, personal, and other types of factors of interest. Moreover, there is a wide range of tools and methods that ease such affective computing system generation process and,

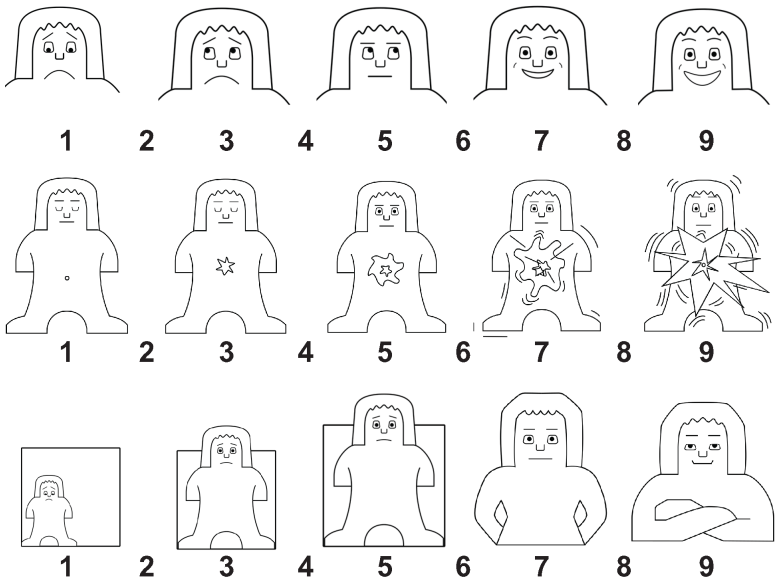


Figure 2-8: Modified SAM by the UC3M4Safety team.

although they can be improved upon avoiding possible bias (e.g.: gender bias), the state-of-the-art in this respect is solid and reliable.

2.5 Physiological indicators for Human Emotions Responses

As detailed in Section 1.2, one of the main objectives of this research work is to disentangle the relationships between physiological signals and negative emotions, e.g. fear-like, to provide the first steps towards an automatic detection of risky situations in a Gender-based Violence and/or sexual harassment context. To achieve this, a deep knowledge on physiological activity on human body under emotional responses needs to be gathered, comprehended, and applied to any technology to be developed.

First of all, the physiological signals are handled by the Autonomous Nervous System (ANS) and so they cannot be manipulated by human will [83,84]. This fact has led the literature to propose and provide different reliable affect-aware system architectures using solely physiological information [85]. However, as for emotional models and theories, there are contrary positions regarding the specific activation and behaviour of the ANS to the different emotions [86,87]. For this research work, we follow one of the latest emotion-related ANS activation claims [88], which is based on the differentiated ANS activity for behaviour preparation and body protection with

respect to the different emotions, which is essential for human adaptation. This is deeply intertwined with the functioning of the brain when receiving external stimuli. In fact, the activation of the ANS is a consequence of the internal circuitry between different parts of the brain which are in charge of decoding those stimuli and triggering the necessary mechanisms to properly adapt to them. Two of the main parts are the amygdala and the hypothalamus, Figure 2-9. The former is the one responsible for emotional processing, while the latter works like a command centre. Thus, in case of a threatening external stimuli, the amygdala sends a distress signal to the hypothalamus, which activates the Sympathetic Nervous System (SNS) through the adrenal glands. Note that the SNS is the branch of the ANS responsible for the known fight-or-flight response. Finally, those glands release different catecholamines (e.g. epinephrine) that brings on a number of physiological changes and reactions. Once the threat is gone, the Parasympathetic Nervous System (PNS) takes the lead by acting as a break for the previous physiological reactions (homeostasis). Note that the PNS is the branch of the ANS responsible for the known rest-and-digest response. Although these biological behaviours and characteristics are mostly agreed in the literature, there is still a high research interest upon providing empirical experiments regarding any of the commented facts [89]. As a consequence of that, the study of the emotional-physiological disentanglement in the literature has been done intensively since 1950, as different researchers were trying to cope with the emotional theories and understand the changes in the physiological variables due to emotional responses [90].

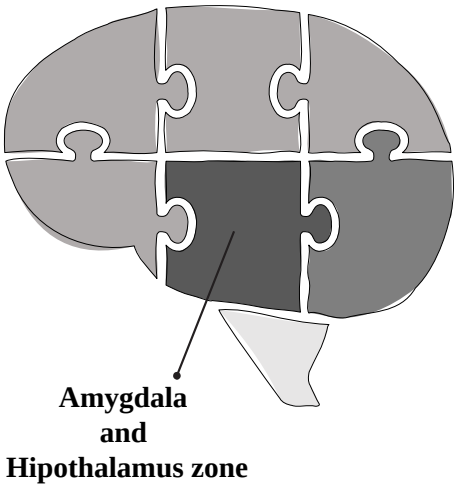


Figure 2-9: Location of the two main parts, amygdala and hypothalamus, involved in the emotional processing and autonomous nervous system regulation.

This section provides a detailed review concerning the nexus between physiological variations and negative emotions, being specifically highlighted those systems in the literature targeting for fear detection. Moreover, due to the wearable nature of the system presented in this work, just the three physiological variables that can be nowadays more wearable-ready (inconspicuous aspect) are covered. Despite of this study, there are still many factors, such as those inherent to the individual, that can directly affect the physiological signal morphology and so the underlying patterns associated with negative emotions. These components, such as age, gender, cardiovascular conditions, allostatic load, and others, used to be not considered in the literature when designing emotion recognition systems. Thus, in this section and following chapters, we analyse and tackle the influence that those elements can have on the morphology of the signals and give some insights to deal with them from a digital systems perspective.

Before going into details for each of the physiological signals to be analysed and studied, the nature of the physiological information must be highlighted, as it strongly affects the techniques applied to extract the physiological affective indicators or physiological emotional metrics. As for any other complex biological system, the human physiological signals posses a non-linear and non-stationary behaviour [91]. However, as they are intended to be digitally processed within a specific embedded platform, fixed-length processing windows are used to extract the different indicators or metrics, which leads up to the physiological quasi-stationary consideration when dealing with short processing windows. Lately, although the latter fact can restrict to the use of linear processing techniques, the application of non-linear techniques is becoming a very successful part of current emotion recognition systems based on physiological information [24] and it is boosting the understanding of complex biological systems in both health and disease [92]. Thus, in the following sections, as well as in the following chapters, the non-linear physiological behaviour is considered essential.

2.5.1 Heart Activity

The cardiac activity is one of the most used physiological information to generate emotion recognition systems [24, 49]. The different phases of the heartbeats, which is translated into different blood pressures within the muscular walls of the blood

vessels, allows for monitoring both sympathetic and parasympathetic variations or changes [6]. From a purely physiological perspective, on the one hand, the highest blood pressure is achieve during the systolic phase, in which the heart contracts to force blood through the arteries. On the other hand, the lowest pressure is achieve during the diastolic phase, in which the heart refills with blood again. This information is strongly affected by the diet of the individual, the age, and possible heart diseases [93,94]. Regardless of the type of these intra-subject factors, all of them lead to morphological modifications from an ideal expected waveform. These modifications are mainly due to peripheral blood vessel resistance changes, which range from different levels of vasoconstriction to different levels of vasodilation [95]. For this research work, being fear the targeted emotion to be detected, the understanding of these physiological principles is necessary, as it is proven that fear stimuli increases the total peripheral resistance leading up to a vasoconstriction increase. The latter is essential to properly distinguish fear-based physiological patterns against those based on any other emotion [85, 96].

The acquisition of this physiological information can be done by different sensors in a non-invasive manner as ECG, photoplethysmography (PPG), and in a invasive one as arterial catheter. Due to the wearable, low power, and inconspicuous requirements of the proposed system, we focus into PPG sensors. They are based on a optical measurement method that employs a light source (a single Light Emitting Diode (LED) or an array of LEDs) and a photodetector which are located at the surface of the skin to measure BVP. There are two types of PPG sensors, reflection and transmission. Figure 2-10 shows an example of these methods by illustrating what is the difference with respect to the path they have through the different layers of the skin. For the reflection mode, the photodector receives the emitting light that has been back-scattered or reflected by the banana effect from the inner layers [97, 98], while in the transmission mode, the photodetector is completely oposed to the LED and it receives the transmitted light passing through all the skin layers. The main difference when getting the signal out of both methods is the inverted behaviour that reflection PPG presents due to the backward direction of the received reflected light. For this research work, we focus mainly in reflection mode due to the wearable aspect and to the fact that most of the PPG sensors available are of this type.

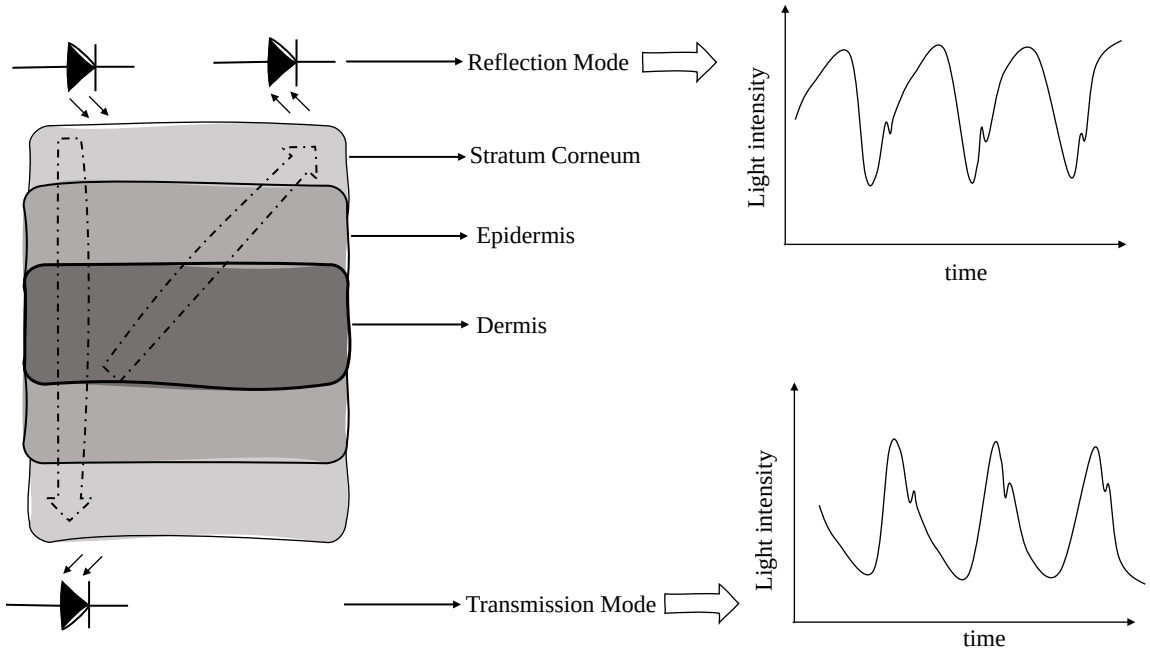


Figure 2-10: Illustration of both PPG measurement techniques, reflection and transmission. Note that the obtained signal is inverted in one method with respect the other.

Therefore, focusing on the reflection mode, it should be noted that the light intensity going throughout all the different layers decays exponentially. Specifically, this assertion is explained by the Lamberts-Beer's Law [99], which is applied to properly model the light intensity received by the photodetector as follows:

$$I = I_{in}e^{-\lambda t}, \tag{2.3}$$

where λ is the wavelength of a specific light, I is the total light detected by the photodetector and I_{in} is the transmitted or incident light. By knowing that λ can be expressed as a direct relation between the absorption coefficient of the medium and the path length, and that the former can be divided into non-pulsatile (DC component) and pulsatile (AC component) tissue contribution, equation 2.3 can also be expressed as

$$I = I_{in}e^{-(\mu_{AC}d(t)+\mu_{DC}m)}, \tag{2.4}$$

where μ_{AC} and μ_{DC} are the absorption coefficients for the pulsatile and non-pulsatile tissues respectively, and $d(t)$ and m are the lengths of the light path through such components. Moreover, the incident light intensity can also be separated into the

static reflected intensity, I_{rf}, and the banana effect intensity, I_b, as following:

$$I = I_{rf} + I_b e^{-(\mu_{AC} d(t) + \mu_{DC} m)}. \tag{2.5}$$

Thus, from equation 2.5, the relation between the DC and AC component is giving by:

$$\frac{AC}{DC} = \frac{I_b e^{-(\mu_{DC} m)\mu_{AC} d(t)}}{I_{rf} + I_b e^{-(\mu_{DC} m)}}. \tag{2.6}$$

In case of assuming that the reflected light is negligible, the amplitude of the normalised AC component would be directly proportional to the dynamic arterial light path length. This assumption is the ideal case scenario, in which the AC/DC ratio is maximised, however, in real applications, the spatial gap between the LED and the photodetector and between the sensor and the skin (air gap) will affect the DC component and minimise the AC contribution. This problem used to be tackled by applying light-blocking structures into the PPG sensors and minimising the air gap [100, 101]. Note that the location of the sensor is equally important on this matter [102]. These reviewed concepts and basics for PPG measurement are essential to properly design efficient wearable systems subjected to integrate such sensor technology. More details in order to deal with the noise of the PPG signals are given in Chapter 5.

From a signal processing perspective, a PPG signal contains different features or metrics that can be extracted and analysed towards decoding their entanglement with emotions. In this work, we distinguish between temporal, frequential and non-linear features. Regardless of the specific type of features to be extracted, the morphological analysis of the signal is required to obtain the necessary PPG characteristic points. Figure 2-11 shows a morphological example of two heart rate periods in which the two previously commented cardiac activity phases appear: systolic and diastolic. Apart from the systolic and the diastolic peaks, there are other characteristic points that will affect the delineation process of this signal. For instance, the predicrotic or incisura, which is the product of the reflections of arterial wall, can be seen in the PPG signal as well right before the dicrotic notch. This sensitive and varying morphology makes the PPG monitoring a challenging task. In fact, recently in [103], the authors presented a comparative study with a group of

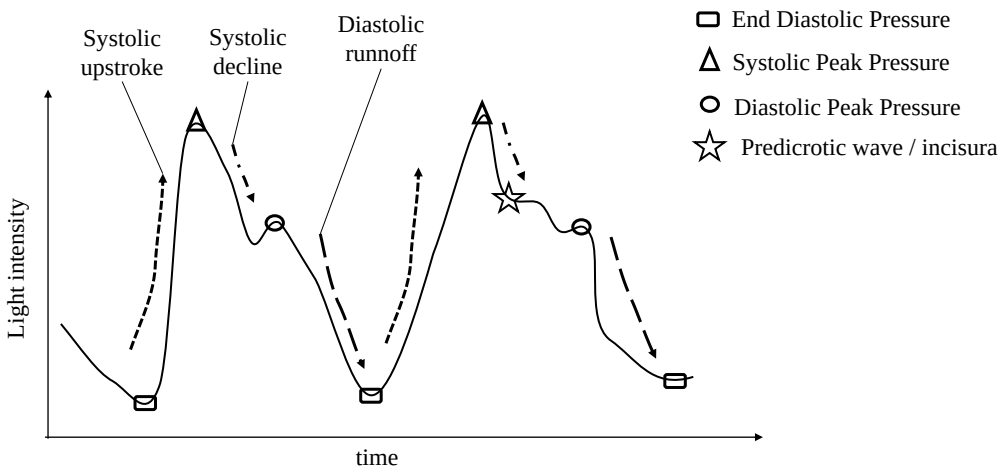


Figure 2-11: Exemplification of the different characteristic points to be extracted within the morphology of the PPG signal.

53 individuals gathering PPG data from six different consumer and research grade wearables. They compared the heart rate given by those wearables with a ground truth heart rate obtained by ECG. The experiment resulted into a maximum mean average error of 15.9 ± 8.1 Beats Per Minute (BPM), being the device type and the exercise activity the factors affecting the most to the estimated heart rate. Note that this error is relevant if wanting to approach an analogous medical equipment norm such as the UNE-EN 60601-2-27, which states that the maximal error for clinical equipment is 5 BPM. This observed problem is due to different aspects:

- Most of the wearables, either consumer or research grade, are neither thought nor designed to acquire clinical or diagnostic PPG data (where morphology is totally preserved), but they get basic PPG quality data. This fact is translated into a highly varying morphology that even depends on the device due to specific electro-mechanical considerations.
- Each wearable is using a proprietary algorithm to extract the characteristic points and calculate the heart rate. This fact is translated into variability between the measurements from the different devices.
- Motion artifacts strongly modify the morphology of the PPG signal. Some of these devices implement techniques to deal with that, while others do not.

Therefore, despite of the proliferation of PPG sensors and their acceptance from the private sector due to the better integrability and cost-effective than ECG, there is still a methodological need for delineation and Motion Artifact Removal (MAR)

standards. Generally, when wanting to extract the heart rate or the period of the PPG signal, the systolic peaks or the end diastolic valleys are a valid option to obtain that periodicity. Note that the maximum bandwidth of the heart rate is approximately $0.6 - 3.5$ Hz, which is equivalent to $36 - 210$ BPMs. Thus, in the worst case scenario for the slowest cardiac activity frequency when having digital constraint resources, i.e. constrained wearable devices, using a processing window of two seconds assure to find at least either two systolic or two end diastolic peaks. As already commented, most of the features are calculated from these points. For this research work, more details regarding the specific delineation algorithms, feature extraction and MAR techniques used are given in Chapters 4 and 5.

Regarding the relationship between cardiac activity and the emotion of fear, there is a wide range of publications in the literature [104–110]. Some of the publications tried to differentiate between positive and negative emotions solely based on heart rate extracted information, while others considered more physiological affective indicators from different physiological variables, e.g. electrodermal or cardiorespiratoty indicators. On the one hand, most of them agreed on the emotion of fear provokes an increase in cardiac acceleration, vasoconstriction, a decrease in blood flow, and an increase in both systolic and diastolic blood pressure. On the other hand, those including more physiological variables claimed the need of considering more information than just the cardiac activity due to the observed direct relation between specific heart rate metrics, such as the variability of the heart rate, and the increase of the respiratory rate or the different electrodermal activity levels. Note that, although there is a well established knowledge in the literature with respect to the cardiac activity effects produced by fear, the experiments are performed in laboratory, where conditions are under control.

2.5.2 Electrodermal Activity

Electrodermal Activity (EDA) or Galvanic Skin Response (GSR) is, along with the cardiac activity, one of the most studied physiological signals that, also, has received an important advance on its comprehension and connection with emotional responses [111, 112]. Although there are more than one type of glands involved in this process, the main responsible ones for the EDA are the eccrine sweat or merocrine glands, which are controlled by the SNS. These are located into the skin

and they are innervated only by sympathetic branch axons (long nerve sudomotor fibers). Note that each axon innervates about 1.28 cm^2 of skin [113]. Figure 2-12 shows an illustrative example of such entities distributed inside the different layers of the skin. The fact that the these glands are only innervated by the SNS makes the EDA the perfect candidate for quantifying SNS activity (fight-and-flight) and, although sweating also plays a major role into thermoregulation to achieve a proper homeostasis, it is proven that the different changes in the skin conductivity are strongly and directly correlated to the intensity of the emotion evoked by external stimuli. Many authors assure that such changes are related to the level of arousal [85,112,114]. Specifically to the evolution of the shape of those changes with respect time, the EDA is formed by a tonic and a phasic component. The former is a slowly varying component, the Skin Conductance Level (SCL), while the latter is the fast Skin Conductance Response (SCR) over time. Note that the physiological theory behind these EDA changes or variations is based on the diffusion and pore opening stated in the poral valve model of Edelberg [115].

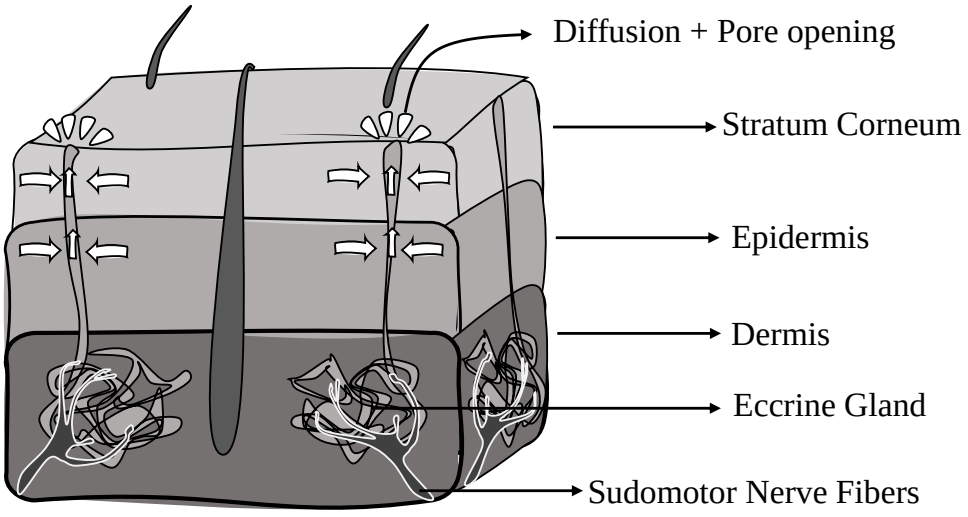


Figure 2-12: Illustration of the merocrine glands behaviour and the diffusion process through the different skin layers.

The acquisition of this physiological information can be done mainly by two different techniques: endosomatic and exosomatic. On the one hand, the former is based on measuring the electrodermal potential using two electrodes without applying neither current nor voltage between them. On the other hand, the latter is based on measuring the electrodermal resistance or conductance using two electrodes applying a small voltage or current between them. Over the years, both methods have been studied, although the wave complexity and challenging interpretation of the

endosomatic methodology have led to a wide acceptance and use of exosomatic measurements [116]. Thus, the exosomatic techniques are characterised by using either a direct or alternating source of electricity through active or passive circuits [117]. First of all, due to the electromotive force at the surface of the electrodes, using DC can lead to electrode polarisation. This problem can be mitigated with AC. However, the application of a source voltage higher than 100 mV and using Ag/AgCl electrodes also minimise the polarisation issues when using DC. Secondly, AC techniques lead to a more complex circuitry implementation, which is mainly due to the fact that both the frequency-independent and dependant information must be preserved as well as the application of posterior digital techniques to recover the real and imaginary parts of such measurements. Note that the frequency-dependent term is actually referred as to the susceptance behaviour of the skin [118]. Finally, it should be noted that the amount of research results considering exosomatic DC techniques is outstanding in comparison with AC and, although AC could undertake DC, more research needs to be done to prove this dominance. In fact, nowadays DC techniques are established as a de facto standard for EDA acquisition [119]. Table 2.3 summarises the main differences analysed between both exosomatic techniques.

Table 2.3: Main differences between DC and AC exosomatic measurements.

Property	DC	AC
Electrode Polarization	≈	✓
Simpler Circuitry	✓	✗
Conductance Information	✓	✓
Susceptance Information	✗	✓
Frequency independence	✓	✗
Amount of research	✓	≈

Therefore, focusing on exosomatic DC measurements, different active and passive electronic circuitry can be used. One of the simplest implementations is done by using single voltage dividers composed by a fixed and a variable resistor. Note that the latter is the human skin. However, this technique is prone to high perceptible measurement errors due to the difference between the voltage source and the voltage to be measured, which causes the latter not to be constant. In fact, active circuitry is widely used instead for EDA monitoring as it mitigates these problems and provides greater control on the measurement. Conventionally, quasi-constant

current and voltage methods are employed by taking advantage of the inverting and non-inverting operational amplifier configurations. For instance, Figure 2-13 depicts an inverting configuration as a possible example of such layouts. In case the electrodes are placed leaving the input resistor R_i as the skin, a quasi-constant voltage is applied which produces the conductance value of the skin be proportional to the output voltage of the circuit. Conversely, if the feedback resistor R_f is the skin, a quasi-constant current is applied over it and the resultant output voltage is proportional to the resistance value. Note that for both configurations, the current limit must be adjusted whether tuning the input and reference voltages or the input resistor, respectively. These adjustments must assure a current throughout the body not higher than $10\mu A/cm^2$, which is the current density recommended level for EDA measurements [120]. Additionally, it should be highlighted also the reference common to output and input, which is intended to avoid any endosomatic contamination of the exosomatic measurement to be performed. In the literature, different proposed EDA circuitry can be found based on inverting active circuitry op-amp configurations [121–123]. Note that these circuits used to be followed by other op-amp conditioning circuitry to adjust the signal and filter it before acquisition.

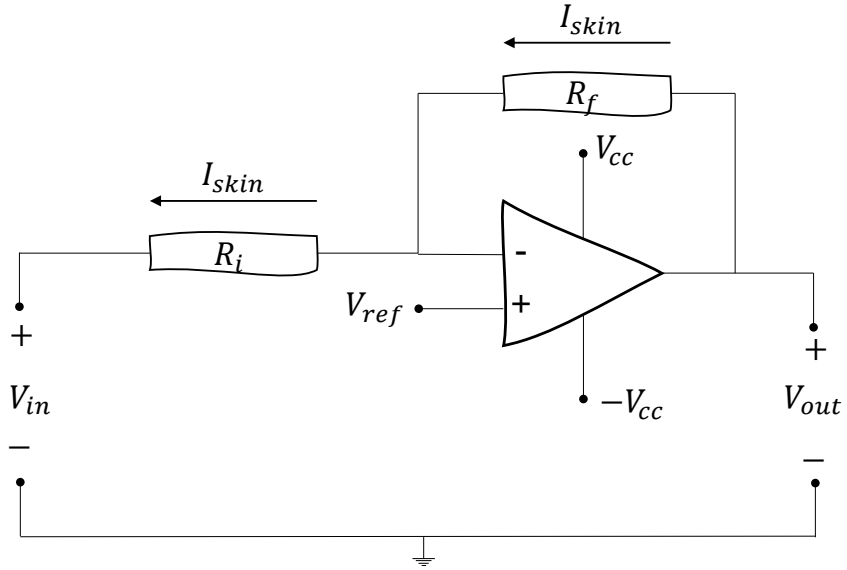


Figure 2-13: Inverting operational amplifier configuration example for exosomatic DC EDA acquisition.

Regardless of the active circuitry used, and besides the tuning trade-off to assure a safety current limit, the relation between range and sensitivity is especially relevant when measuring this physiological signal due to the relatively wide range (0 µS to 25 µS) and the 0.01 µS sensitivity that needs to be satisfied to properly record all

the SCR within the tonic and phasic components of the EDA signal [112]. To deal with this problem, a Wheatstone bridge circuitry can be applied. In that case, by reaching the bridge calibration between the two branches, i.e. by means of adjusting a potentiometer in the opposite bridge branch of the human skin resistance, so that the potential difference is zero or a specific desired voltage [124]. Once the circuit is in that state, the potential difference disturbances are the SCR, and the SCL can be obtained based on the bridge calibration. Although, this method can in fact provide a reliable measurement and assure proper range and sensitivity by adjusting in run-time some of the resistors (potentiometer), it has not been fully adopted neither extensively used in the literature. For this research work, DC exosomatic active circuitry is adopted. More details are given in Chapter 5. Note that there are other options for the DC exosomatic acquisition, such as AC-coupled amplifiers and backing-off circuits, however, they offer a higher circuitry complexity.

From a signal processing perspective, one of the first tasks to do after acquisition is to apply basic low pass filtering and to properly separate tonic and phasic components (SCL and SCR). Both are equally important regarding emotion disentanglement, thus their preservation throughout the analog acquisition and digital manipulation is desired. However, the phasic component is the one containing the ERSCR, which translates into different EDA bursts that are emotionally related with external stimuli and characterised by different metrics based on the actual level of excitement evoked. Thus, the tonic and phasic decomposition is mainly intended to the proper identification and analysis of the ERSCRs. Note that the phasic component can also present a Nonspecific Skin Conductance Response (NSSCR), which occur in the absence of an identifiable eliciting stimuli. Different thresholds for the metrics of each detected SCR peak can be assumed to determine the distinction between ERSCRs and NSSCRs [125]. Figure 2-14 shows an example of one ERSCR and some of the metrics that can be extracted from it. For this research work, more details regarding the specific features extracted are given along the following Chapters.

One of the simplest methods to overcome the tonic and phasic decomposition from the EDA signal is by assuming a linear combination of these two, as given by the

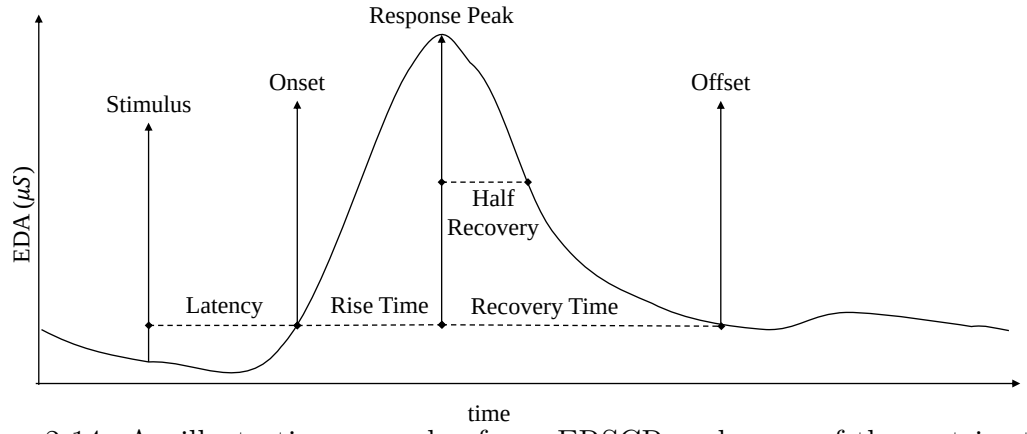


Figure 2-14: An illustrative example of one ERSCR and some of the metrics that can be extracted from it.

following approximation:

$$EDA_{total} \approx EDA_{tonic} + EDA_{phasic}, \tag{2.7}$$

where EDA_{total} is the filtered signal, EDA_{tonic} is the low frequency component or the trend associated with the SCL, and EDA_{phasic} is the resultant signal containing the different SCRs. Thus, by subtracting the trend of the filtered signal and applying a trough-to-peak technique, all the relevant peaks of the signal can be extracted. Note that by such subtraction, a pseudo-phasic signal component is obtained. Specifically, EDA_{total} can be obtained by applying a low-pass FIR filter with 1.5Hz cut-off frequency, which is selected based on the fact that EDA information remains bellow it [112]. Later, getting the EDA_{tonic} can be done by implementing a moving-median filter using a wide enough window to capture the trend bellow 0.05Hz [125]. Although this technique can be implemented in a straightforward manner and does not have a negative effect on any storage nor resource constraint, it is just an approximation and faces different problems. On the one hand, the resultant phasic component can be negative, which is never supposed to happen from a physiological perspective. On the other hand, this method does neither considers nor deals with overlapping SCRs, which can lead to an underestimation of the different response peak amplitudes. Thus, this method is recommended as a starting point. Further developments have appeared in the literature along recent years accompanied by automated tools, such as Ledalab [126], that gather the different most used algorithms to boost their applicability on EDA research. On top of those

algorithms we find cvxEDA [127] and SparsEDA [128]. The former is motivated by the deconvolution method introduced by Alexander et. al. [129], in which they stated that Sudomotor Nerve Activity (SMNA) posses a shorter time-constant than the EDA signal itself and produces bursts (pore diffusion) that arrive as separated and discrete events. They applied a deconvolution technique by means of a biexponential function that tackled the SCR overlapping problem. Thus, considering that basis and handling the negative rationale problem of the phasic component, cvxEDA uses a convex optimization that is constrained by the sparsity and non-negativity of the SMNA, which modifies equation 2.7 as following:

$$EDA_{total} = IRF * (Driver_{tonic} + Driver_{phasic}), \tag{2.8}$$

where IRF is identified as the biexponential Bateman impulse response function, and the $Drivers$ are the information coming from the SMNA. This algorithm has actually been applied succesfully to different EDA research use cases. However, although the convolution operation by itself needs low computational resources, the convex optimisation procedure needs to tune different hyperparameters which leads to a high computational time. Regarding the SparsEDA algorithm, which is one of the latest EDA decomposition methods recently published in 2017, it is based on the previous deconvolution works but introduced different features, such as the application of the least absolute shrinkage and selection operator non-negative version by using the least-angle regression algorithm, which make the deconvolution faster, more efficient and more interpretable than its predecessors. Despite of these advantages, its applicability and performance for small EDA segments (shorter than 70 seconds) is still on debate. Thus, although these two algorithms provide different advantages mainly related to the physiological EDA interpretation, their applicability into multimodal constrained wearable devices, such as the bracelet of Bindi, is a challenging task due to the high computational resources derived from specific operations, such as the convex optimisation. Therefore, alternatives that are placed between the trough-to-peak and the convex methods are needed. For instance, some authors [130, 131] have used a Regularized Least-Squares Detrending (RLSD) method [132] in which the tonic component is approximated to a low-frequency

aperiodic trend component by

$$EDA_{tonic} = \frac{EDA_{total}}{(I + \lambda D_2^T D_2)}, \tag{2.9}$$

where $\lambda D_2^T D_2$ is the regularisation term that biases the SCL to a smooth trend, I is the identity matrix, and D_2 is a discrete approximation of the 2nd derivative operator. Note that the greater the λ, the smoother the SCL component. After getting this approximate tonic component, the same subtraction to the original EDA signal applies to obtain the phasic component. For this research work, all the reviewed methods have been used, although just the trough-to-peak and the RLSD methods have been embedded. More details regarding the obtained results are given in Chapter 5 and Chapter 6.

As for any physiological signal, noise artefacts due to motion, rapid transients, and even loose electrodes can be observed during its acquisition. Figure 2-15 depicts a real example of the different noise sources that can be found within measurements. This image shows the difference between dry and wet electrodes, as the latter are more affected by the noise due to the nonexistence of Ag/AgCl which makes the skin-electrode interface less robust and being effective just through sweat. Note that this problem is specially relevant for Bindi, as it is based on dry-electrodes. To combat these type of noise sources and mitigate their possible negative effects during the EDA processing, different preprocessing steps used to be applied such as moving mean and median filters. More details regarding the implementation of such techniques are given in Chapters 4.

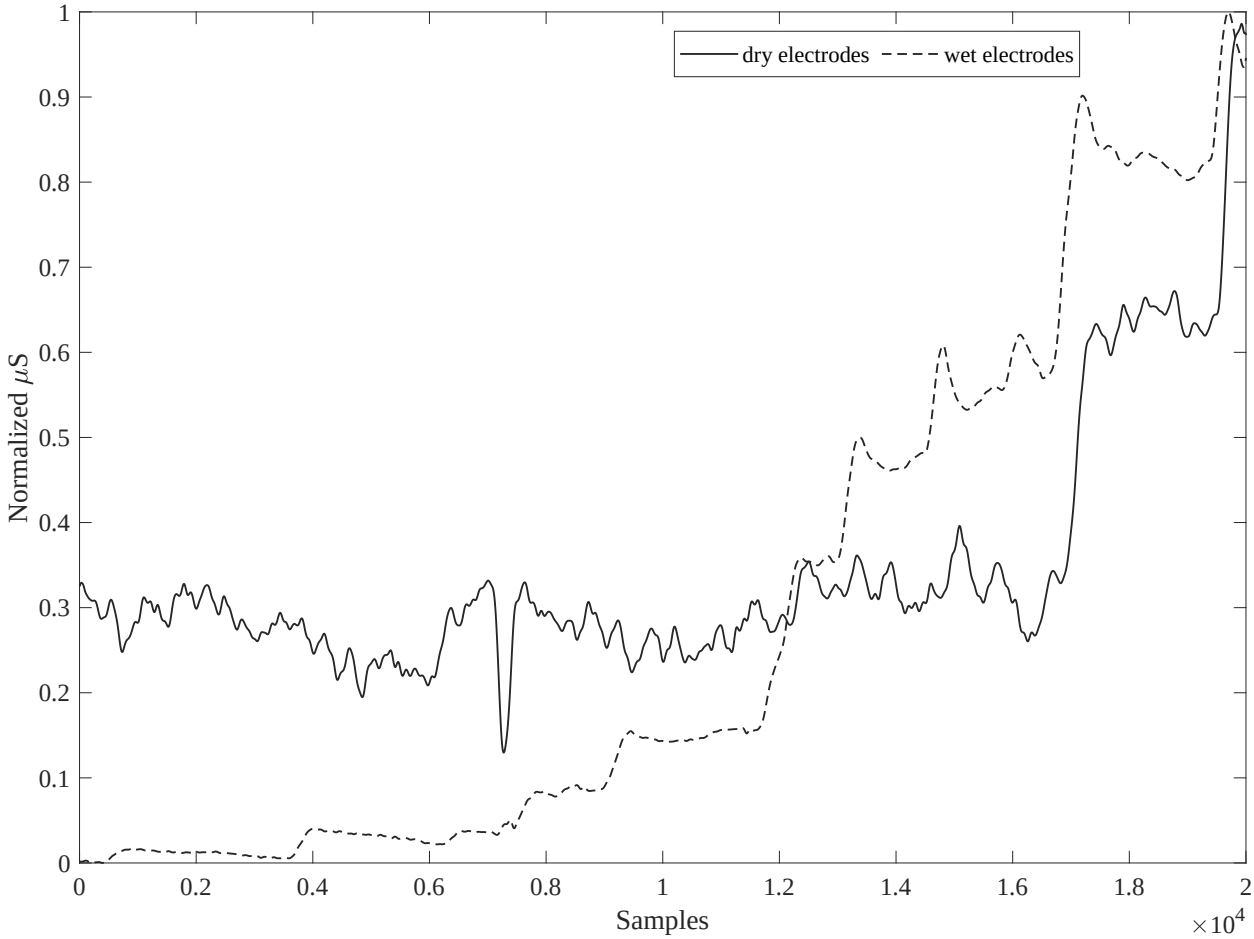


Figure 2-15: Difference between dry and wet electrodes measuring in the ventral side over the right (wet) and left (dry) part of the wrist. Note that units are normalised μ S and wet electrodes contain 0.5% chloride salt.

Regarding the relation that EDA has with the fear emotion, different research groups studied this aspect [85,133]. As previously explained, the changes observed in the EDA signal can be directly linked to the intensity of the emotion, but not to the type of it. In fact, the studies considering solely this signal targeted stress detection, arousal quantification or even sympathetic function assessment, but not emotion identification. Briefly, fear emotion cannot be detected by using only the EDA information. In any case, the information extracted from this signal can provide an excellent insight regarding changes of the sympathetic activation. For instance, when facing very stressful situations, the SNS secretes different catecholamine hormones (adrenaline and noradrenaline) that produces the EDA signal to be characterised by an increase of SCL and an increase of the different metrics to be extracted from the SCR except for the latency which tends to be decreased. Note that is such situations the perspiration increases which is directly related to homeostasis rather

than to the emotional process. The interpretation of this information regarding its emotional entanglement in case of gender-based violence victims is more complex due to the possible allostatic overload, which refers to the cumulative effects of stressful situations in daily life experienced by the individuals and can even inhibit the sympathetic activation switch-off [134]. Although addressing this latter specific fact is out of the scope of this research work, details and basis are given in Chapters 5 and 6.

2.5.3 Skin Temperature

The body skin temperature is not so popular in comparison with the two already detailed physiological signals. However, there exist several researchers dealing with emotional identification by using this information as well [24, 135, 136]. The physiological foundations for this signal are strongly interlinked with the blood flow and the electrodermal responses of the body. In fact, skin temperature is strongly related to the vasomotion changes by means of sympathetic noradrenergic fibers which regulate such process. As stated previously, the ANS does not only provides mechanisms to deal with threatening external stimuli, but it is also the main responsible for the different homeostasis processes. Specifically, in the case of the body temperature, the hypothalamus is referred to as the main thermoregulatory controller [137]. Throughout the body, we have different temperature receptors that allow the hypothalamus to continuously sense and analyse the body temperature. Once this part of our brain gathers the needed information, the subsequent actions vary depending on that negative feedback, as for any physiological control system. Thus, as for a home thermostat, based on a preset normal value, the different thermoregulatory defences will be triggered to preserve, in the case of our body, 37°C. For instance, due to such defences, the body temperature does not deviate more than a few tenths of a degree from the preset value. Moreover, there is even a called interthreshold range, over which no thermoregulatory action is activated, that is known to be around 0.2°C. Actually, those thermoregulatory defences are as follows:

- Sweating and vasodilation are the defences triggered when a heating situation is undergoing.
- Conversely, vasoconstriction is triggered to diminish the heat loss by mainly lowering skin surface radiation.

Therefore, the autonomic thermoregulatory process operates in a synchronised fashion together with other physiological negative feedback control systems, such as the blood flow and electrodermal autonomous handling. Moreover, it must be highlighted that there exists an internal source of temperature variability marked by physical, mental and behavioural changes that follow a daily cycle known as circadian rhythm. The comprehension and understanding of these physiological factors are essential to evaluate and assess properly the information to be recollected.

From a signal processing perspective, this signal does not present the same complexity as the previous ones. Conversely, its information is contained within very low frequencies, below 0.5Hz. Thus, using an ordinary FIR filter is sufficient to obtain a clean signal. After that, the literature tends to extract standard features from it, such as mean, median, standard deviation, and other high order statistics. Moreover, its extracted frequency information is divided commonly into different bands as any other physiological signal [10, 138]. However, as well as the processing is one of the easiest among physiological signals, its integration is not. In order to implement a temperature sensor within a wearable constrained device some considerations need to be addressed. For instance, the authors in [139] elaborated a survey considering 172 studies from 1960 to 2016, in which they reviewed all the factors that affect the temperature measurement when dealing with contact thermometry. They concluded with a set of recommendations and trade-offs between all these factors (skin-sensor interface, attachment, environmental protection and bias, sensor pressure into the skin, etc.) that can strongly affect the body skin temperature to be measured. These technical requirements make the integration of skin temperature sensors a challenging task. In fact, no commercially available wearable device (smartwatch like) integrates a body skin temperature sensor. There are research grade wearable devices that integrates infrared thermopile sensors, such as the E4 by Empatica®[2] [140]. However, nowadays the latter have a high cost, making its integration not as straightforward as contact thermometry.

Regarding the relation between the fear emotion and body temperature variations, different studies in the literature dealt with it. Early research on this topic can be found in [141–143], in which, although the experiments were performed with dif-

[2]https://www.empatica.com/en-eu/research/e4/

ferent experimental procedures, they agreed that the body temperature decreases under fear elicitation. Note that for the three studies, the temperature sensors were placed in the palm. Recently, the research targeting body temperature variations with respect to emotions is more focused on facial thermal mapping using functional infrared thermal imaging. For instance, the authors in [135] used 60 pictures from [58] and asked twenty four student (19 females) to rate the pictures based on the SAM scales while measuring the facial skin temperature and the EDA. They observed that the highest decrease in temperature was produced for the pictures with the highest arousal. Thus, they stated that the autonomous regulation of arousal is actually carried out by two sympathetic cutaneous responses, thermal and electrodermal. However, one of the main disadvantages of body skin temperature in comparison with other physiological information is the large latency of the signal. This causes a limitation when using solely this information to infer the emotional state. Thus, its integration used to be accompanied together, i.e. compensated, with other physiological signals, such as EDA and BVP [133].

Despite these thermal-emotional patterns and characteristics observed, up to my knowledge, there is no research dealing with the body skin temperature variations in the wrist neither with its behaviour under fear-related gender-based violence situations. These facts are essential for this research work considering the proposed bracelet within the Bindi system, as the temperature sensor is directly attached to the wrist due to the factor-form itself. More details regarding the obtained results are given in Chapters 5 and 6.

2.6 Conclusion

In this Chapter, we have provided the foundations needed to raise an emotion recognition system. Note that technical aspects related to the specific design for the training of such system are provided in the following Chapter.

Thus, the different emotional theories and human emotion classification techniques have been reviewed and detailed. Specifically and targeting the particular use case of this research, a new pragmatic approach to merge discrete and dimensional classifications of human emotions towards the identification of the fear emotion is proposed. Moreover, a comprehensive analysis into the intrapersonal factors affecting

emotion modulation such as personal traits, attention and gender bias, is provided to establish future research possibilities to be further developed as an extension of this research work. Additionally, the different emotion elicitation tools used within the affective computing community are presented and compared, highlighting the recent inclusion of VR, which is overtaking emotion elicitation experiments. On the other hand, the physiological signals of interest for this research work have been reviewed and analysed by detailing their behaviour and characteristics and studying their relationship with the fear emotion. Note that the understanding of such physiological information is essential to properly quantify and distinguish the different physiological patterns that are product of an emotional reaction.

On this basis, we can conclude that, although felt emotions are biased by different intrapersonal factors, the physiological information can be used as an indirect quantification or measurement of those affective states, as these signals are controlled by the ANS, together with their subjective self-reported evaluations. In this context, the conjunction of different physiological signals, rather than the use of just one of them, can be used to give rise to an intelligent affective computing system able to distinguish different affective states. In the pursuit of such an emotion recognition system, which can be further extended to be used on a daily basis, two main factors are highlighted and taken as essential for the development of Bindi in this case. First, the need for accounting for both human emotion classifications, discrete and dimensional, can be an advantage to explain different features of emotions. Secondly, the analysis of multiple physiological sources of information in real-time is a complex task as, from a wearable perspective, they are subjected to different noise sources which directly affects the quality of the signals and so the emotion recognition inference. More details on the application of all the detailed aspects regarding human emotion classifications, emotion elicitation tools and physiological and emotion disentanglement are provided in the following Chapters.

Chapter 3

Databases and Machine Learning for emotion recognition

In this chapter, on the one hand, we provide a complete analysis regarding the structure and experimental procedures used for the generation of databases designed for emotion recognition. Note that, as specified in Chapter 2, these databases are essential to gather emotional responses and train emotion recognition systems. Besides that, each part of the whole data processing chain for the affective computing system using such databases is also explained. Note that the understanding of the current database generation state-of-the-art has been essential to properly design the database presented in this work. Moreover, a critical review is made along the different sections, providing recommendations on what should be considered for the generation of an emotion recognition database and insights into what has been finally applied for the generation of ours, which is fully detailed in Chapter 6.

Before going into details for each part involved within both the database generation and the affective computing system design, a general representation of such elements and actions is shown in Figure 3-1. As stated in Chapter 2, a database for training an affective computing system is composed of the following main elements: 1) stimuli, 2) physical and physiological signals, 3) labels, and 4) volunteers. The second and third elements will be used for training and validating the affective computing system, while the first is required to provoke emotional reactions on volunteers. Within this context, the process of building a database implies the following tasks regarding these elements:

a) Prior to the database generation:
   - A pool of stimuli are recollected.
   - In case of facing a time-limited experiment, different methods are applied to reduced the number of stimuli from the previous pool.
   - The final set of stimuli are arranged to be used during the database generation.
   - The different sensors are selected validated to collect all information during emotion elicitation.

b) During the database generation:
   - The different variables to be measured during stimuli reception are recollected and stored.
   - Self-reported data (emotion labels) is gathered and stored to identify the physiological and physical information with respect to the specific stimuli.

c) After the database generation:
   - Digital filtering and conditioning is used to clean the different signals.
   - Exploratory data analysis is performed to identify abnormalities and even physical or physiological problems.
   - Extraction of different synthetic metrics and/or features from the data. From those, reduction, selection and optimisation is applied.
   - In case of being under a multimodal use case, different alternatives can be approached toward the data fusion.
   - Application of an iterative process among the data fusion architecture itself, the classification algorithm and hyper-parameter fine tuning processes.
   - Releasing of the model with the best performance.

This chapter is structured as follows. Within the first Section, the common elements, processes, and actions required for the generation of an emotion recognition database are explained and reviewed one by one. The next Section provides a detailed summary for the different emotion recognition multimodal databases that are openly available in literature. This Section also details how such databases have addressed the different points explained in the previous Section, as well as their limitations and applicability to our use case. For the third Section, the dif-

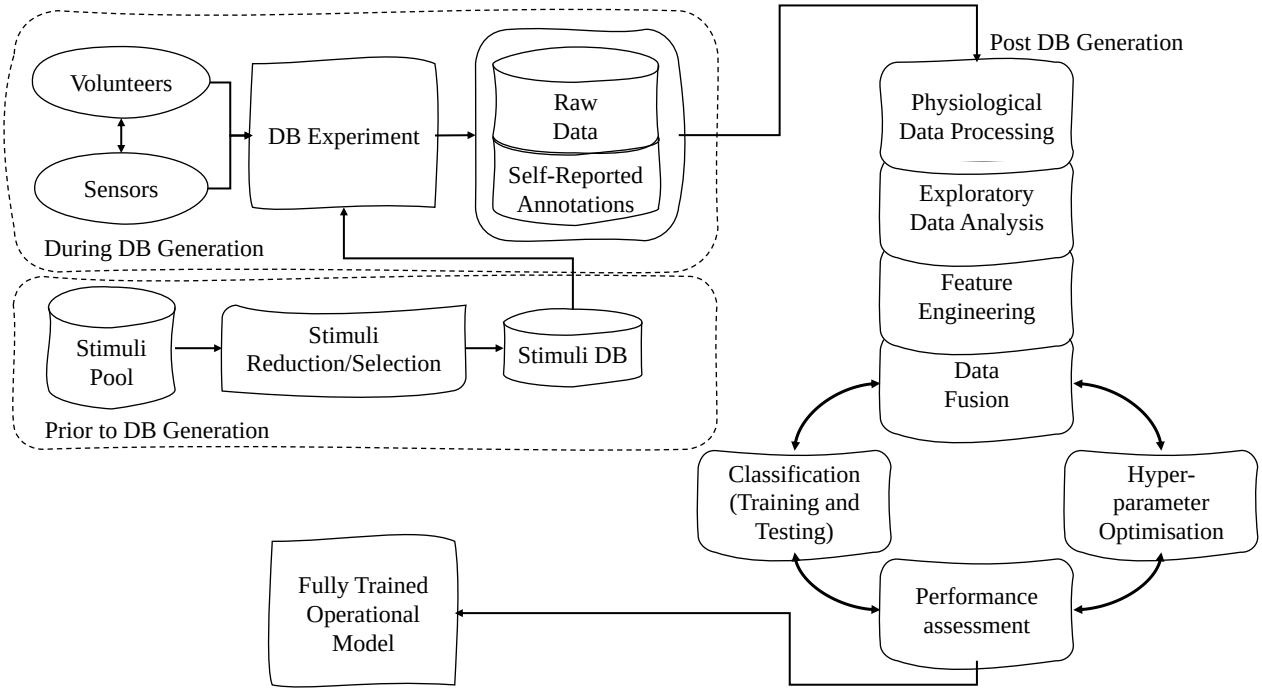


Figure 3-1: Common elements, processes, and actions required for the generation of an emotion recognition database.

ferences between laboratory and in-the-field setups for the database generation are highlighted. Note that for this research work, the only database presented in this document is based on laboratory setup settings. However, the conclusions from this last Section can be used for the near future in-the-field databases generation.

3.1 General database methodology

The elements and processes involved in the generation of an emotion recognition database are outlined in Figure 3-1. The following subsections explain their particularities, advantages and disadvantages. Note that the final output is a fully trained operational model based on offline processes that are performed after the database generation. Therefore, this chapter does not deal with any embedded digital optimisation for each process. Embedded integration and optimisation for real-time wearable devices is detailed in Chapter 5.

3.1.1 Stimuli analysis and selection

The generation of a set of adequate stimuli is the first stage for any database focused on emotion recognition. This step is essential, as the better is your stimuli pre-labelled the better the research-ground truth of the experiment will be. The

latter can be even an useful tool to be compared and analysed with self-reported annotations for every volunteer within the database, as detailed in Section 2.4. Thus, the ideal situation is that the labelling methodology were the same during the stimuli selection and during the database experiment. Note that within these type of experiments, we actually have two types of labels (ground truth), those coming from the stimuli selection and those coming from the self-reported ratings of the volunteers.

There are already a wide range of stimuli databases options in the literature. For instance, in 1997 the National Institute of Mental Health launched the international affective picture system as a database containing hundreds of labelled pictures [58]. This system has grown since then, leading up to more than 1000 labelled images, and even being adapted to other cultures, such as Spanish [144]. The labels contained within this database are characterised by the mean and standard deviation of the three dimensions of the PAD space. Apart from public available image stimuli pools, video-based databases are found as well [10, 138, 145]. Regardless of the stimuli type, the stimuli analysis and selection are essential, as some of those databases are not practical to be entirely (all the stimuli) applied within one experiment. Thus, different approaches can be observed in the literature to filter and select smaller sets of stimuli. Some of them [138] used the called emotional highlight score based on the PA space and given by

$$\mid e \mid = \sqrt{a^2 + v^2}, \tag{3.1}$$

where a and v are the arousal and valence values obtained from the gathered labels respectively. Thus, the stimuli with higher $\mid e \mid$ will result into the ones providing the stronger emotional intensity in terms of dimensions a and v. Note that this equation can be expanded to be valid and applicable for further dimensions. The obtained scores for all the stimuli can be ranked to further select a smaller group with the highest values. Others, such as [10], performed a pre-labelled experiment in which they started from a big stimuli pool and asked people to label those stimuli by using an online affective annotation system. They assured a minimum number of labels per video, and finally selected the ones that obtained the highest degree of agreement. The evolution of the latter method includes higher order statistics applied to the pre-labelled experiments to assess the agreement between different

annotators. For instance, the authors in [145] computed Jaccard distances for each pair of annotators and calculated the mean absolute deviation of the cumulative distance distribution to finally consider as outliers those that deviated more than a specific threshold. Nevertheless, one of the steps followed together with the stimuli selection is the stimuli balance assessment. This process assures that the selected stimuli are equally distributed along all the different emotions to be detected or classified. For instance, one common approach is to deal with the PA model, just for the sake of simplicity in comparison with the PAD model, and normalise the ratings using the mean and standard deviation (μ/σ), to later plot the normalised arousal versus the normalised valence and proceed to assess the balance stimuli status.

In reality, this step is strongly conditioned by the goal of the experimentation. Most of the open available stimuli databases for emotion recognition are thought and designed from a general emotional perspective, i.e. with the goal of identifying emotions in general without targeting binary specialised emotional models. Even the pre-labelled procedures are usually done by the general public, without considering any expert evaluation. This approach is totally understandable from a general and massive use perspective, however, for research works like the one being addressed in this document may not be suitable. Considering that the main goal of this research work is based on generating affective computing systems for fear detection in gender-based violence situations, the stimuli selection must be done with special care and such databases may not be adequate to fulfil the requirements in terms of specific emotion elicitation. More specific details on how we addressed the stimuli selection and analysis for the generation of our database are given in Chapter 6.

3.1.2 Sensors acquisition and processing

During the experiment of any database, different sensors are acquiring physiological and/or physical signals from the volunteer while stimuli are applied. These sensory systems need to be properly designed by considering the following aspects:

- As the generation of a database is just a huge recollection of data to further create intelligent systems from that, it is recommended to have relatively high sampling frequencies. This allows experimenting with any lower sampling frequency at the training stage to observe how that constraints and affect the

different classification models.

- Regardless of the sampling frequency, the synchronisation between all the different sensors during the experiment must be guaranteed. However, to alleviate this process, another option is to store global timestamps from each sensor data received to further ensure that they correspond to the same time-slot of the experiment.
- The use of an approved or standard sensing toolkit is recommended, to be used as golden measurement system. This will allow further comparison with other databases as well as detecting malfunctioning in the proposed sensor system.
- The sensors should be preferably located in the best measuring position as close as possible to the final body location and, in case a further wearable integration of the resultant affective computing system is expected, where been intended.

These factors are recommendations based on the state-of-the-art [24] and the knowledge gathered along the development of this research work.

Regarding the data processing, the first task is to apply basic filtering by using digital low and high band pass filters. For signals that possess a high sensitivity to noise, such as PPG, specific Signal Quality Assessment (SQA) procedures can be applied as well as MAR algorithms [146, 147]. The same applies for signals that need special component separation algorithms, such as EDA, as explained in Section 2-14. Chapter 5 provides more details on the different techniques and algorithms designed and applied in this work. Note that nowadays there already exist in the literature open available tools designed for physiological processing. For instance, Soleymani et. al. in [148] designed an open toolbox for processing a complete set of physiological signals processing and emotional related feature extraction. Besides that, different physiological signal processing toolboxes are found, being specialised in just one physiological signal [149, 150]. However, these toolboxes are intended and oriented for a PC-based or offline system design process, leaving aside the embedded wearable constraints.

Together with the data processing, data segmentation used also to be applied over the different signals. In fact, most of the emotion recognition systems in the literature use segmented processing windows to treat and analyse the acquired physiolog-

ical data. When dealing with data segmentation, window-related aspects, such as their temporal and frequency resolution and emotional latency, should be considered. On the one hand, the temporal resolution has a direct relationship with frequency resolution. This is due to a specific frequency resolution needing to be guaranteed to extract useful emotional information for some physiological features [6]. On the other hand, emotional latency is related to the fact that a person does not experience the same physiological response (emotion) during the entire reception of a stimulus [151,152]. The latter aspect can definitively affect the system performance, as it is related to the possible incorrect labelling of the samples.

3.1.3 Exploratory data analysis

Once the signals have gone through all the needed data processing, it is recommended to perform an exploratory data analysis. This process can be done by using the filtered data and/or the extracted features. this process can provide excellent insights into what is actually, at first glance, happening during the experiment. Moreover, it can also give insights into those cases in which the sensor is malfunctioning and the filtering or processing stages could not fix it. This type of exploratory data analysis allows us to determine some of the physiological behaviour during the different stages of the experiments and to carry out specific actions to deal with some problems, such as physiological recovery not working as expected or lengthen stimuli since emotional latency was affecting some of the physiological responses. More details regarding these facts are given in Chapter 6.

The different and public available emotion recognition databases are not reporting this exploratory data analysis, the papers published are focused on the generation of the database (the gathering data process). Exploratory data analysis is a very time consuming task, but the physiological effect of the experiment is very useful from the point of view of emotion detection. Therefore, other researches have performed this analysis after the release of the different databases. For instance, the authors in [153] used one of the open public databases in the literature [138], seven years after its release, and concluded that induced emotions were stronger in the final part of the stimuli based on an exploratory data analysis over the filtered physiological data. That conclusion led them to train their proposed system using only the last 20 seconds of each stimulus. Thus, these and other technical considerations

resulted into a significantly increase of the accuracy of emotion recognition rate as compared to the existing state-of-the-art emotion classification techniques. However, despite the advantages that this process can bring related to the effectiveness of the generated affective computing systems, it involves a considerable amount of time as well as the need for a good knowledge in physiological signals. The latter consideration is actually the most challenging factor, as the different physiological patterns may largely vary across subjects and experimental sessions. In fact, the recent emotion recognition reviews in the literature do not even address anything related to this topic [24, 49, 133, 154, 155].

3.1.4 Feature engineering

Feature engineering involves the utilisation of different mechanisms to improve the emotion recognition model performance. Note that it is only applied for conventional machine learning strategies and deep learning strategies in which the inputs are the extracted features. Thus, an essential distinction must be done before going into details for feature engineering. On the one hand, conventional machine learning and deep learning using feature extraction require ad hoc extraction techniques as well as optimisation, Figure 3-2. On the other hand, there are deep learning methods that do not need a feature extraction stage, as they can learn patterns and inherent principles directly from the processed data to extract already optimised features automatically. These latter methods are known as end-to-end solutions, and they seem to be very promising for emotion recognition in physiological and multimodal problems [156–158]. However, deep learning methodologies, whether they depend on hand-crafted or learned features, still require a considerable amount of resources. For instance, TensorFlow Lite, which is nowadays one of the most used open-source machine learning frameworks for low power and very constraint devices, can deploy deep learning models with a footprint from 300KB to 1MB[1]. Unfortunately, when considering the design of ultra-low-power wearable devices using current System on Chip (SoC) technology, these memory sizes can impair other critical tasks to be performed within such devices. Even though, it is worth mentioning that a tremendous effort is being applied to boosting deep learning into edge computing systems, such

[1]https://www.tensorflow.org/lite/guide (Accessed: 01/03/2022)

as the TinyML foundation[2] or the Subthreshold Power Optimised Technology® by Ambiq Micro Inc.[3]. Thus, for this research work, we are focused on the conventional machine learning architecture, leaving embedded deep learning and/or end-to-end deep systems to further research. The following subsections discuss the different processes that can be performed for the feature engineering strategy. Note that these are carried out just once during the training of the system, posterior to the database generation, but prior to the deployment of the system.

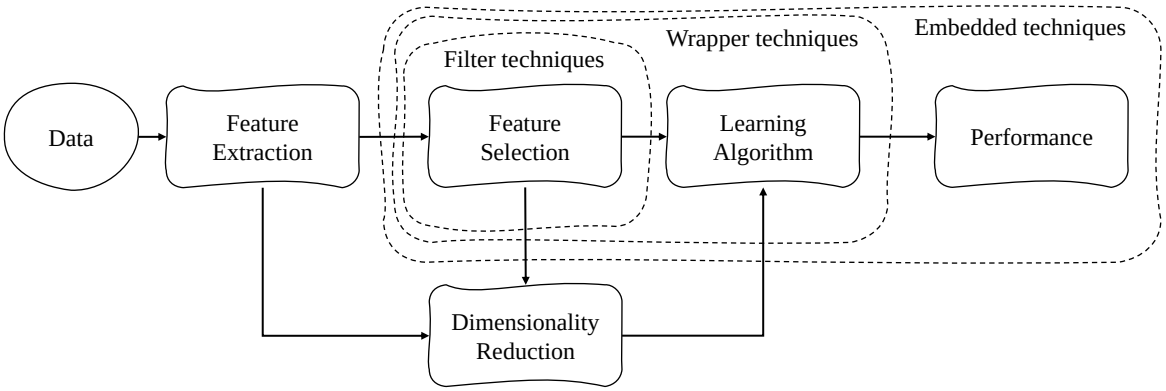


Figure 3-2: Conventional feature engineering processes for supervised feature selection.

3.1.4.1 Feature Extraction

The first task within the feature engineering process is the extraction of synthetic metrics from the previous filtered data. Regardless of the vast amount of feature extraction techniques that can be found in the literature for emotion recognition [24, 49, 154], they can be divided into three categories:

- Temporal domain. These features possess the lowest computational complexity amongst the different types of feature extraction techniques. Note that most of the temporal domain features can be implemented in a linear fashion ($\mathcal{O}(n)$). Specifically, they provide information regarding the stationary and linear aspects of the time series being analysed. Most of them are extracted through higher-order statistics computations. One of the biggest disadvantages of these features is the inability to capture the non-stationary physiological behaviour.
- Frequency domain. The goal of these features is to obtain the PSD in specific frequency bands for the different physiological signals. The common method to get the PSD is based on the Discrete Fourier Transform (DFT) by using

[2]https://www.tinyml.org/ (Accessed: 01/03/2022)
[3]https://ambiq.com/ (Accessed: 01/03/2022)

the Fast Fourier Transform (FFT) algorithm. For that specific case, the time complexity generally levels up to $\mathcal{O}(n \log n)$ in comparison with the temporal features. Moreover, dealing with the frequency domain is synonymous with the time-frequency resolution problem. The latter fact is of special relevance in the case of physiological information, as some of them are slow-changing signals, such as the EDA, which is known to have time-varying responses from 1 to 30 s based on the type of stimulus [119]. Apart from this problem, there is physiological information that is an unevenly or non-uniformed sampled signal, which makes the application of the FFT algorithm impossible. This problem is tackled in the literature by employing different techniques such as prior-interpolation or the Lomb-Scargle periodogram [159]. Thus, although the frequency content is proven to be a reliable measure to track emotions, different trade-offs must be considered regarding the optimisation of these techniques as well as the frequency resolution needs (temporal window storage size and processing capabilities).

- Non-linear methods. To disentangle the dynamic and non-stationary properties of the physiological signals, different methods are used. Note that these types of features are also referred to as chaotic features. In fact, the works that have contemplated, used, and even compared non-linear against temporal or frequency features obtained a considerable performance improvement in their specific emotion recognition objective [160, 161]. Besides that, in the last years, the applicability of deep learning to emotion recognition problems has increased due to the promising obtained results [153, 162, 163], which is an indicative of the physiological non-linear emotional component as well. Within this context and in line with the non-linear importance, different studies and reviews recollected and analysed the non-linear physiological behavior [91, 92], as noted in Section 2.5. The main and biggest disadvantage of these techniques is the time complexity they have, as it can be up to $\mathcal{O}(n^2)$.

Most emotion recognition systems based on physiological signals and using well known emotional labels are based on conventional temporal and frequency feature extraction. Thus, the combination of the three domains (temporal, frequency and non-linear) should be extended in the literature. This approach might be exploited

to gain a better understanding of the physiological variations and changes concerning the self-reported metrics used as labels within these types of systems. It should be noted that, the categorisation provided by this research work is based on the different commented reviews in the literature. However, there can be more specific feature extraction techniques or even different names for the proposed categories. For instance, morphological features [164], being those referring to specific physiological signal properties (amplitudes, times, number of peaks, etc.), used to be also employed interchangeably to the temporal domain techniques. In the case of this research work, we have elaborated a compendium of the most relevant and successful features considering the three categories reviewed. More details are given in Chapter 4 for their specific implementation.

Once features have been successfully extracted, it is time to optimise them. Such optimisation can be done by feature selection and/or feature reduction [165]. The former is based on identifying the most relevant features and creating new subsets of features with those, whereas the latter deals with the dimensionality reduction of the problem by means of different types of basis transformation. Note that the basis transformation process refers to the conversion of the high-dimensional extracted features, i.e. high number of features, into a low dimensional space with a minimal loss of information. Both of these feature optimisation methods are essential to simplify the model (less storage, improved visualisation, data reduction, Occam's razor), to avoid the curse of dimensionality, and to reduce training time.

3.1.4.2 Feature Selection

For the feature selection procedure, we differentiate three common techniques, which are outlined in Figure 3-2 and applied for emotion recognition according to the relationship with the learning methods [166]. In the first place, we can find the simplest techniques known as filter methods. These are based on general statistical metrics, such as the correlation with the dependent variable, by which the different features are ranked to further select the new subset. Although they possess the lowest computational complexity, they are more prone to fail to select the best features, as neither the interaction between them nor the effect of the new subset on the classifier performance is considered. Secondly, towards the avoidance of the filter methods problems, we find the wrapper methods. These use the classifier to

verify the performance effect of the new subsets iteratively generated. Two of the most known and used wrapper methods are Sequential Forward Selection (SFS) and Sequential Backward Elimination (SBE). Specifically, the former is initialised with an empty subset of features and starts combining them till no improvement is observed, while the latter performs the same operation backwards by starting with all the features and eliminating them one by one. The wrapper methods are known to provide better performance at the expense of: (1) high computational needs when the number of features is relatively high, and (2) over-fitting risk when the number of input samples is relatively low. Moreover, they are strongly conditioned to the type of classifier used during the different wrapping iterations. Lastly, the third type of these methods is known as embedded methods. These were created to deal with the different disadvantages of the previous two techniques and to keep their advantages. In this case, the feature selection mechanism is integrated into the core of the classification algorithm and it takes advantage of its feature selection and classification at the same time. This provides computational complexity and speeds even compared with the filter techniques and being much less prone to over-fitting. Note that the commented techniques use the labels or target variable, which is known as supervised feature selection. However, there are also methods which do not need the target variable, such as correlation-based techniques. These latter techniques can provide insights regarding the relationship among the different features to further discard redundant information.

3.1.4.3 Dimensionality Reduction

As already commented, another possibility to optimise the feature space is to apply feature reduction. This method is based on an unsupervised transformation of the extracted features into an entirely new feature space. For instance, one of the most common techniques is Principal Component Analysis (PCA), in which each new feature is obtained by a linear combination of the original features. Specifically, PCA calculates the covariance matrices of the original features to later extract their eigenvectors and each corresponding eigenvalue. Then, the eigenvectors are sorted by the eigen values in descending order (from more to less carried information) and only kept those of interest. Such stored eigenvectors are putted together giving place to the projection matrix, which will be used for the original data projection. One of

the main disadvantages is that PCA can produce independent variables to be less interpretable, as the original features become principal components. This method has been extensively used for feature reduction in emotion recognition and other machine learning problems [155, 167, 168]. Note that, besides PCA, a wide variety of methods exist in the literature regarding feature reduction, such as t-Distributed stochastic neighbour embedding, generalised discriminant analysis, or independent component analysis [169].

Taking into account that the search for the ideal subset of features, whether it is done by selection or reduction, is an NP-hard problem, the only way to obtain an optimal solution is by performing an exhaustive search within the space of the solution or within the application of different feature reduction techniques. However, even considering that this process can be done during the training of the system without involving any digital embedded constraint, this is a challenging task. Moreover, the wide variety of techniques and active research on this field introduce even more complexity to the problem. Thus, the proposal, development and/or implementation of new dimensionality reduction techniques are out of the scope of this document. Instead, along the development of this research work, different commonly used feature selection methods have been applied for our specific use case. More details regarding their implementation are given in Chapters 4 and 6.

3.1.5 Hyper-parameter optimisation

The term hyperparameter is referred to the values involved within the learning process of the different machine learning algorithms that can not be estimated from data. When dealing with conventional machine learning, the hyperparameter adjustment process can strongly enhance the classification model during training. However, as for the feature selection, this process is also an NP-hard problem, as the perfect hyperparameters are obtained after all the different and possible combinations have been verified. For the sake of simplicity, just imagine a least-squares approximation problem (a linear regression problem), in which the fitting of the model is evaluated by the residual of every point given by

$$r = y - f(x), \tag{3.2}$$

where r is the obtained residual for the observed sample y when considering the model defined by $f(x)$. Assuming that the approximation of the model is a straight line, the previous equation results into

$$r = y - (b + mx), \tag{3.3}$$

where b is the interception with the dependent variable and m is the slope of the straight line model. In fact, these are the parameters of the model directly affecting the adjustment to the observed data points. However, in order to find the optimal adjustment, a loss function must be evaluated for all the possible combinations. For instance, least-squares techniques use quadratic loss functions to minimise the residuals. Since running all the possible combinations is a very time consuming task, different techniques are used in the literature to optimise this search and provide a well optimised machine, i.e. algorithm [170]. One of the simplest methods to do this is by setting a maximum number of iterations to verify such loss function based on a specific step size or learning rate while moving toward the minimum of such loss function. These latter values are set before running the model and are external to it, being identified as hyperparameters. Although, there are a plenty of hyperparameter optimisation techniques, we have reviewed three of them:

- Grid search. This technique is based on a predefined grid of hyperparameters combinations, i.e. a preset space of possible combinations, which are sequentially run and tested. This method used to be very exhaustive but very time-consuming at the same time. For instance, if we take three hyperparameters and check 50 values for each, that results in a total of 125,000 combinations to be tested. Thus, grid search can be used for a first approximation of the problem, knowing that it is going to be neither the best nor the cheapest in terms of resource and time consumption [171].
- Random search. This technique follows the same concept as the grid search, i.e. a search is performed over a preset space of possible combinations. However, instead of evaluating those combinations sequentially, the technique uses random combinations within such space. The amount of iterations is limited explicitly by the designer. Overall, this method was proven to provide better models in most cases and required less computational time [172].

- Bayesian optimisation. One of the weaknesses of the previous two techniques is that the evaluation of new points or hyperparameter combinations within the grid does not consider any information regarding the score evolution along the optimisation process. Thus, Bayesian hyperparameter tuning is known as a Sequential Model-Based Optimisation (SMBO) technique that uses the previous iterations knowledge to concentrate on better function loss scores, i.e. it is based on a continuously updated probabilistic Gaussian model that allows choosing the next hyperparameter combination in an informed manner to boost the evaluation of more promising values [173].

More details regarding the specific use of these techniques for this research work are provided in Chapters 4 and 6.

3.1.6 Data fusion

The interdisciplinary nature of the affective computing problems aiming to recognise emotions together with the technology advancements open endless possibilities in terms of modality observation. Note that the term modality is referred to as multi-sensor data acquisition, in which each sensor is intended to capture data from totally different sources of information (e.g. audio, physiological, text, visual). For instance, our brain is already working on a multi-sensor information basis and making decisions based on data fusion. In fact, the authors in [174] performed a detailed and exhaustive review of multi-modal experiments in the literature in comparison with uni-modal. Through that survey, they confirmed that multi-modal systems outperform uni-modal systems. Moreover, they also point out that deep learning techniques are gaining ground to conventional machine learning by using end-to-end deep learning models, which do not need the feature extraction steps as they can be fed directly using the raw data [158].

Within this context, we can categorise Bindi as a multi-modal system, in which we have two different modalities: physiological and audio. These can be fused by employing different data fusion methodologies, which are described as following:

- Early fusion. This method is based on performing or applying the fusion task on the early stage of the problem, i.e. by using the data or even the features. The former can be done by removing correlated information between modalities, whereas the latter fuses the different features (from different modalities)

into just one single feature vector. For instance, Figure 3-3 depicts a possible example of early fusion, in which feature extraction is applied independently over both modalities and the resultant feature vector is just the concatenation of such. The latter process is one of the fastest methods to fuse feature, however, it can be done applying other techniques such as point-wise addition. One of the main advantages of this fusion method is that only one classification model is required to be trained.

- Late fusion. In this case, the sources of information follow totally independent paths, which may not even have the same components or processes, until a classification output is given by different and independent classification models based on the modality. Figure 3-3 depicts a possible example of late fusion, in which both modalities have independent classification models and the output of those is fused. Whether the model is providing a soft label (any output metric providing information on the predicted probability of class membership, e.g. 50% probability belonging to the positive class) or a hard label (predicted class without any probability information, e.g. label '1' and '0' for positive and negative class), different techniques can be used to perform such data fusion. For instance, one of the common techniques is performing a weighting scheme [155] given by

$$c = arg\,max\left\{\prod_{m=1}^{M} P_i(X|C_m)^{\alpha^m}\right\}, \tag{3.4}$$

where M is the total amount of modalities, X is the data input, $P_i(X|C_m)$ is the probability of X belonging to the i class and provided by the classifier of a specific modality C_m. The different weights for each modality α_m are determined during the training stage and they only need to satisfy $\Sigma_{m=1}^{M}\alpha_m = 1$. The main advantage of the late fusion is the ad-hoc design that can be performed for the different modalities independently, however, that fact also leads to the need for more than one classifier.

- Intermediate fusion. This data fusion implies the transformation of the extracted features into a new representation of the original data, i.e. basis change. It is used to be mostly applied when dealing with deep learning

models, in which the data fusion can occur anywhere along the inner layers of the neural network. This fusion is more flexible in comparison with the other two in which the information is fused whether at the beginning or at the end. However, there are very few examples of this technique in the literature, on the contrary as the previous one.

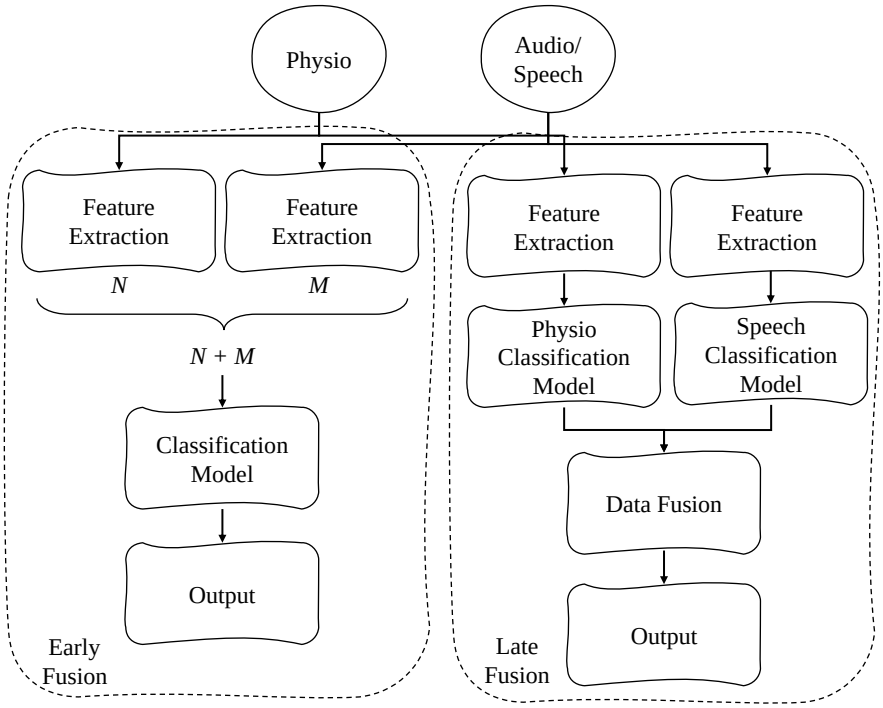


Figure 3-3: Early and late data fusion techniques for physiological and audio/speech extracted features, with dimensions N and M respectively.

More details regarding the specific use of these techniques for this research work are provided in Chapter 6.

3.1.7 Emotion Classification

This stage, together with the data fusion, is one of the last to be performed towards the achievement of a fully trained and tested affective computing model, see Figure 3-1.

3.1.7.1 Bias-Variance Trade-off

Before describing different models that are of interest for this work, the Bias-Variance trade-off might be explained and properly addressed to understand all the concepts associated within this stage.

The performance of the machine learning algorithms are mainly defined by their bias and variance. The relationship between these metrics is directly related to

under-fitting and over-fitting issues. For instance, let us consider Equation 3.2, but assuming that $f(x)$ defines the true relationship between y and x. In that case and assuming that we create or design a function $f'(x)$ that corresponds to our machine learning model, the quality of such algorithm under unseen test points can be measured by the Mean Square Error (MSE) as

$$MSE_{f'} = E[(y - f'(x))^2], \tag{3.5}$$

which can be further decomposed into

$$\begin{aligned} MSE_{f'} &= E[(f(x) + r - f'(x))^2] \\ &= E[(f(x) + r - f'(x) + E[f'(x)] - E[f'(x)])^2] \\ &= E[(f(x) - E[f'(x)])^2] + E[(E[f'(x) - f'(x))^2] + E[r^2] + \\ &+ 2E[(E[f'(x)] - f'(x))(f(x) - E[f'(x)])] \\ Note &: E[E[f'(x)]] = f'(x) \\ &: \text{The last term cancels to zero.} \\ &= E[(f(x) - E[f'(x)])^2] + E[(E[f'(x) - f'(x))^2] + E[r^2] \\ &= bias[f'(x)]^2 + variance[f'(x)] + \sigma_r^2, \end{aligned} \tag{3.6}$$

where r is the residual or random noise with zero mean and σ_r^2 variance ($E[r^2]$), *bias* is the difference between the average expected value of prediction and the actual value, and *variance* quantifies the consistency of the output prediction value based on the variation of the training data points. An illustrative example of these concepts is shown in Figure 3-4, from which different conclusions can be obtained:

- A model with high bias and low variance is within the under-fitting zone, being unable to adjust itself to the training data. This fact leads to high training and testing errors.
- A model with low bias and high variance is withing the over-fitting zone, being adjusted to much to the training data that is unable to generalise or fit new unseen test data. This fact leads to the lowest training error at the expense of a high test error.
- The best model is the one that minimises errors from wrong predicted values

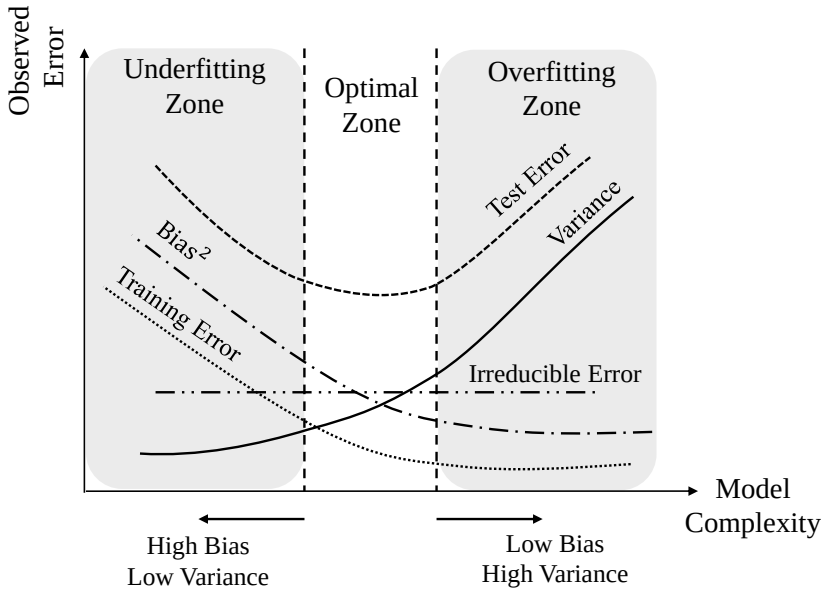


Figure 3-4: Bias-Variance trade-off with underfitting, overfitting and optimal zones.

(low bias) and presents robust consistency to the variations of the training data (low variance). This is identified as the optimal zone, in which the model achieves the perfect trade-off between the training and test error.

- Even when achieving the lowest bias and variance within the optimal zone, the quality of the model will be determined by the irreducible error, which is irrelevant to the model and related to the inherent noise within the data.

Note that considering the usual bias and variance behaviour for the different machine learning algorithms to be evaluated is essential. In fact, traditional machine learning algorithms suffer from this trade-off problem as their complexity increases.

3.1.7.2 Machine Learning Algorithms

For this research work, different well-known machine learning models are used based on current reviews focusing on emotion recognition [24]. They are described as follows:

- Support Vector Machines (SVM) [175]. This supervised classification algorithm is one of the most popular machine learning algorithms. Although originally was proposed solely for binary classification problems, it has been extended and applied for multi-class problems as well along the years. The main idea behind this classifier is based on finding a hyper-plane that best separates the data into the different classes. Note that the data are the different extracted features being fed to the classifier. In this context, two main

elements need to be defined to understand the hyper-plane concept: support vectors and margins. As depicted in Figure 3-5, assuming a binary classification with two features, the support vectors or support vector points are the ones closest to the hyper-plane (mid part of the margin). From a 2D perspective, the hyper-plane can be conceptualised as the line separating both classes given by equation 3.3.

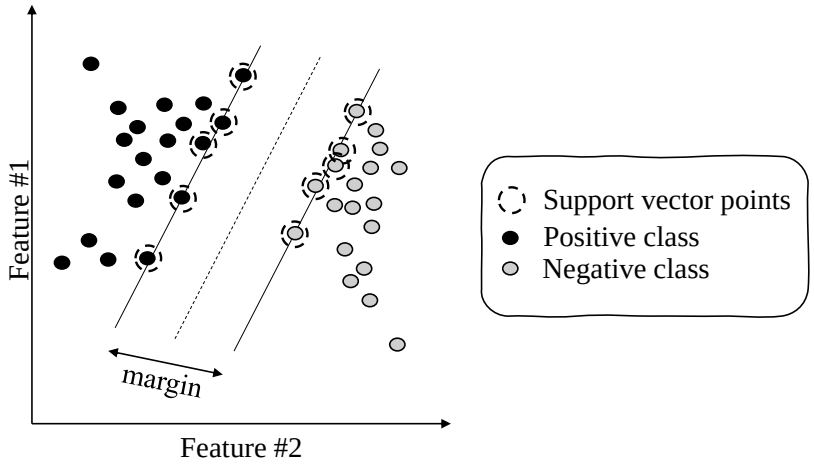


Figure 3-5: Hyper-plane illustration for the SVM classifier for binary classification (black dots are positive class, and grey dots are negative class).

However, to define the entire hyper-plane, such equation is expanded into or generalised to the M dimensions of the problem as following,

$$\begin{aligned} y &= w_0 + w_1 x_1 + w_1 x_1 + ... + w_M x_M \\ &= w_0 + \sum_{n=1}^{M} w_n x_n \\ &= b + w^T X, \end{aligned} \tag{3.7}$$

where w^T are the support vector points, X are the provided training points, b is the biased term or the offset of such hyper-plane, and y is the class label (positive or negative for a binary problem). Thus, we can define any hyper-plane as the set of points satisfying

$$w^T X + b = 0. \tag{3.8}$$

Note that, considering such equations, the optimisation problem to get the optimal hyper-plane is based on maximising the margin to best separate the data into the different classes, as already remarked before. Therefore, such

optimisation problem is actually selecting two initial hyper-planes that meet the following constraints:

$$w^T X + b \geq 1, \ for\ the\ positive\ class \ \longrightarrow y = 1, \tag{3.9}$$

$$w^T X + b \leq -1, \ for\ the\ negative\ class \ \longrightarrow y = -1. \tag{3.10}$$

These constraints can be rearranged and expressed by the following,

$$y * (w^T X + b) \geq 1. \tag{3.11}$$

Nevertheless, the above equations and assumptions are only valid if the data are linearly separable, which is not the case when dealing with physiological information due to the non-linear nature of it. In these cases, the previous equation is modified by adding an extra parameter, ζ, allowing or accounting for classification error during training. This leads up to soft-margins, rather than hard-margins, with the following formulation:

$$y_i * (w^T X + b) \geq 1 - \zeta. \tag{3.12}$$

Additionally, a hyper-parameter C is used to handle such miss-classification cost and keep control of the soft-margins. However, in most of the cases when the data are not linearly separable, the application of soft-margins is not enough, and different kernels need to be applied. The application of a kernel can be thought as mapping the data into higher dimensions, so they can be linearly separable in a new higher dimensional feature space, Figure 3-6. Moreover, these kernels use the called kernel-trick through which there is not need to know or worry about these higher dimensional transformations, as the kernel functions allow inputs in the original lower dimensional space and give back the dot product of the transformed vectors in the higher dimensional space. For instance, one of the commonly employed kernels in emotion recognition systems and also used in this research work is the Gaussian or

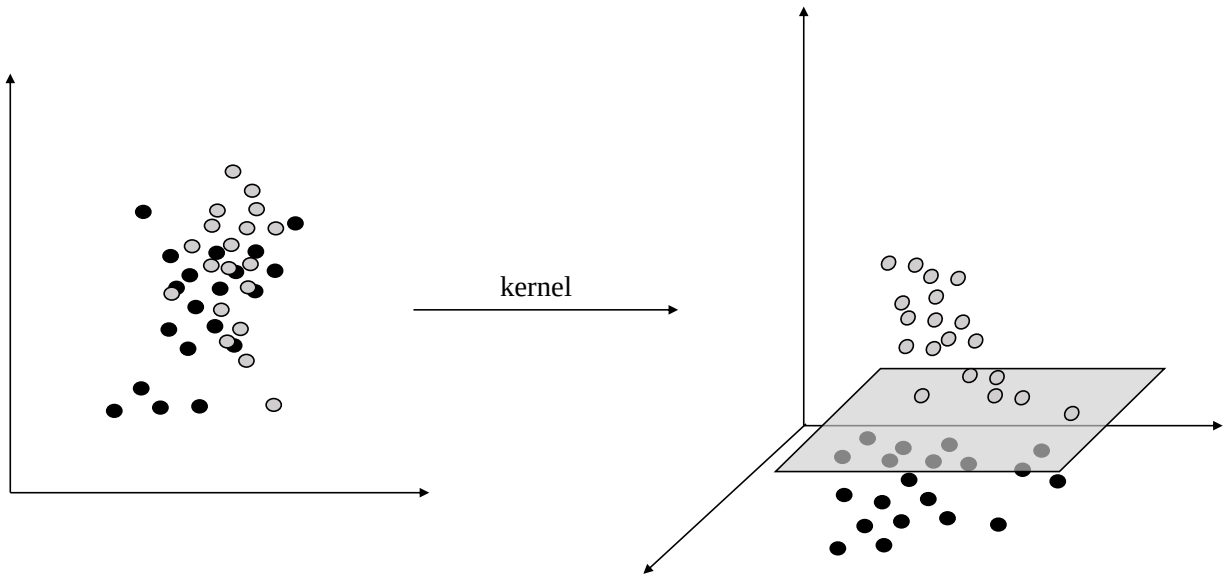


Figure 3-6: Kernel trick illustration for a binary problem.

Radial-Basis Function (RBF) kernel, which is given by

$$K(X_1, X_2) = exp(-\frac{||X_1 - X_2||}{2\sigma^2}), \tag{3.13}$$

where $||X_1 - X_2||$ is the Euclidean distance (L2-norm) between data points (feature data points) X_1 and X_2, and σ is the hyper-parameter to be tuned to consider that two points are similar (they belong to the same class). Note that this kernel is bounded superiorly by 1, as the distance between two points that are extremely similar is zero. Based on the value of σ, the region of similarity (zone where $K(X_1, X_2)$ is higher than zero) between points will change, Figure 3-7. This algorithm is a discriminative classifier whose bias and variance are determined by the C and σ hyper-parameters for the soft-margins and the RBF kernel respectively. The main advantage is that it presents a higher memory efficiency in comparison with other classifiers (just need to store the support vectors, not all the training data points), but it does not perform well when leading with too much overlapping between the different classes.

- K-Nearest Neighbours (KNN) [176]. This is also a supervised classification algorithm, however, it is called a *lazy* classifier. From a mathematical point of view, there is not an actual learning process involved within the algorithm. Instead, it seeks the best distance d and the number of neighbours k that

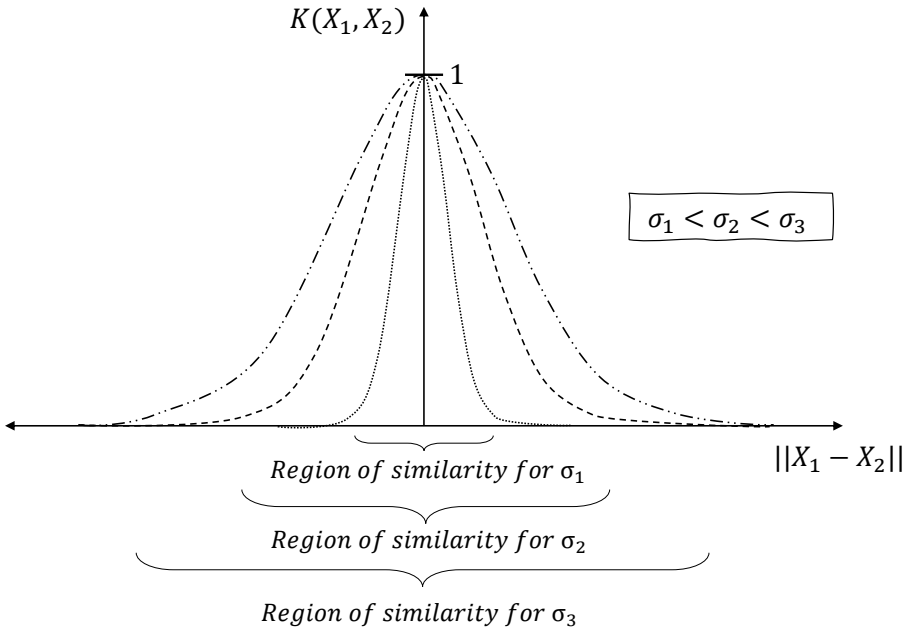


Figure 3-7: RBF kernel values based on the distance between the two points being evaluated for different σ.

maximises the separation of the classes. Thus, performing predictions upon new data arrival requires the calculation of such distance with each of the training data points and further comparison with the k surrounding neighbours to determine the belonging class. This algorithm assumes that similar things exist in close proximity. Note that different type of distances can be used (Euclidean, Minkowski, City block, Mahalanobis etc.), as well as different sorting algorithms to find the k closest neighbours after distances calculations. From a practical point of view, KNN is one of the simplest algorithms to implement. Thus, it is a right choice for a first proof-of-concept approach. However, it gets significantly slower as the number of training samples increases, as well as it affects memory efficiency.

- Ensemble Methods (ENS) [177]. These methods are actually a set of machine learning techniques, rather than a classifier. They are based on the combination of different base models or weak classifiers to produce one optimal or strong classifier. Such combination is commonly performed in a bagging or boosting manner. In bagging, each model is trained independently on the same training set, while in boosting, each weak classifier is trained considering the previous classifier performance by applying a weighting data mechanism (higher weights assign to incorrectly classified instances).

For this research work, boosting ensemble methods are used and, specifically, the Adaptive Boosting classifier or AdaBoost is applied. This classifier is very popular for binary classification and the weak classifiers employed to implement it used to be decision stumps (trees with just one node or one-level decision trees) or shallow trees (trees with very limited depth). Note that such specific type of trees enhances comprehensibility. Thus, for every weak learner (m) and all for the instances within the training set (N), this classifier computes the weighted classification error as

$$\epsilon_m = \frac{\sum_{i=1}^{N} w_i^{(m)} I(f_m(x_i) \neq y_i)}{\sum_{i=1} w_i^{(m)}}, \tag{3.14}$$

where $w_i^{(m)}$ is the weight of instance i for the learner m, and I is the loss function defined by

$$I(f_m(x), y) = \begin{cases} 0, & \text{if } f_m(x_i) = y_i \\ 1, & \text{if } f_m(x_i) \neq y_i \end{cases} \tag{3.15}$$

After training, this classifier predicts the label of new unseen information following a weighted linear combination of all the considered weak classifiers (M), which is given as following:

$$g(x) = sign\left(\sum_{m=1}^{M} \alpha_m f_m(x)\right), \tag{3.16}$$

where α_m is the total weight assigned to each weak learner given by

$$\alpha_m = \frac{1}{2} log \frac{(1-\epsilon_m)}{\epsilon_m}. \tag{3.17}$$

From a practical perspective, the inference stage of this machine learning algorithm requires less storage and possess a lower computational and temporal complexity in comparison with the two previously algorithms reviewed. However, it is more sensitive to noisy data and outliers, which requires data properly filtered and noise-free before feeding them to the machine.

3.1.7.3 Cross-Validation Techniques

Besides the specific model to be applied, the split of the data into train, validation and test sets must be performed prior to the training process. Although the separation of such datasets is used to be embedded within the tasks of the classification procedure, it can be also conceptualised as an additional operation, see "Performance Assessment" in Figure 3-1. Within this context, the training set can be defined as the set from which the model is going to learn the underlying patterns and adjust its hyperparameters. The validation set is the one affected by the cross-validation techniques, through which an estimation of the model performance can be obtained. Note that this set is actually part of the training set (the model sees this sets in training). Finally, the test part is the one that is not seen by the model during training and provides the final and unbiased evaluation of a fully trained model. However, there are different methods to separate these sets, and the selection of one technique or another depends mainly on the amount of data and the need to hyper-parameter tuning. These splitting techniques are grouped under the CV term. For this research work, different CV techniques adapted to emotion recognition databases have been used. They are described as follows:

- Hold-Out. This is the simplest CV method in which the data is divided into two sets (train and validation). During this CV, the model is fitted with the first one and evaluated using the data within the second one. Note that the final trained model is obtained by using the whole dataset (train and validation). Although, a third dataset can also be obtained to be considered as the test dataset, unseen data not used at all during the training stage. A typical split ratio is 80% for training and 20% for testing, although this ratio depends on the dataset. The main disadvantage of this method is the risk of over-fitting (high variance), as different sets (different distributions of the split) can even affect to the obtained results. Note that, as the training dataset is reduced when using this technique, it can even lead to the risk of losing inherent patterns of the signals or data.
- $k - fold$. To overcome the limitations of the previous technique and decrease training variance, this method is based on splitting the training set into k partitions, which can provide up to k different possibilities to train and validate

the system. Comparing to the previous technique, this method is usually preferred as it can give a more realistic (less overoptimistic) measurement for the model performance. The main disadvantage of this method is the computational time needed to run k times the training of the model.

- Leave-One-Trial-Out (LOTO). In this method, a sample is left out of the training process to later test the model with it. However, for the emotion recognition use case, this technique can be modified to identify a sample as a trial of the experiment. For instance, in an experiment based on the physiological recording while visualising different images, a trial would be identified as the physiological data captured during one of the images visualisation. Moreover, the fact that the number of possible training combinations is defined by the number of trials, brings this technique to the same advantages and disadvantages as for the $k - fold$ CV with k equal the number of trials.
- Leave-One-Subject-Out (LOSO). This technique follows the same concept as LOTO, but in this case the sample that is left out of the training is an entire subject or volunteer. As previously stated, considering the same image emotion recognition example, all the data recollected from one subject is used for testing purposes while training with the rest of the subjects or volunteers. The main difference of this technique in comparison to LOTO is the data variability observed within the test set. In fact, the test set in LOTO is based on just one trial, which used to be identified with one label, whereas the test set in LOSO is based on different trials from the same volunteers. Thus, while LOSO can assure, at least for a subject, a representative test distribution, LOTO is always subjected to the uncertainty of having a test set represented by just one label.

For this research work, some of these techniques are implemented to handle the generation of the different sets (training, validation and test) for the machine learning models. Specifically, the emotion recognition problem requires these strategies to be applied to generate two type of models: subject dependent and independent models. The former are trained, validated and tested using the data from one single volunteer, while the latter use the data from all volunteers to create a global model. The main difference between these models is the personalization. In fact, most of

the variability between subjects lie in the dynamic nature of their affective states and their previous experience. This fact can be proven by the superiority of the subject dependent models over the subject independent models in the literature [154]. Hence, in line with Chapter 2, stimuli interpretation and physiological changes are strongly volunteer-dependent. Thus, personalization emerges, as it was done in [24], in which the authors concluded that an emotion recognition subject-independent model could be deployed but, at some point, user customisation will be necessary to improve the system. For these reasons, there is a need in the literature, when facing emotion recognition using conventional machine learning, to come up with new CV techniques which provide some type of personalisation. In fact, looking at other fields that use as well human information, it can be observed the application of hybrid CV techniques which basically combine subject independent and subject dependent models [5]. Within this context and, up to my knowledge, there is no emotion recognition research work applying hybrid CV techniques. Thus, besides applying some of the reviewed techniques, this research work proposes the utilisation of the called Leave-hAlf-Subject-Out (LASO) CV technique as well. Note that a graphical depiction for LOSO, LOTO, and LASO CV techniques is shown in Figure 3-8.

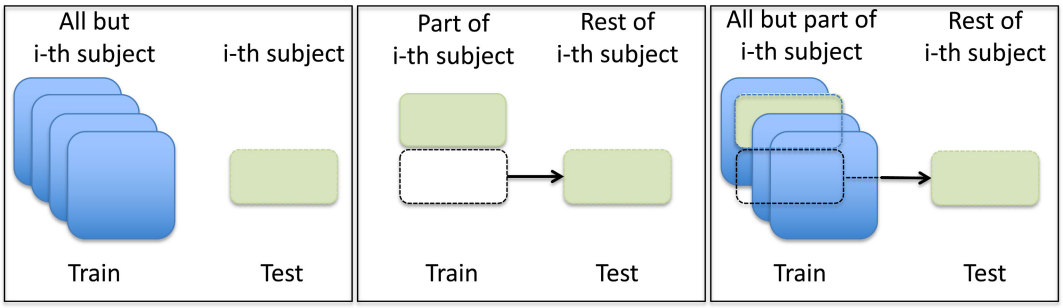


Figure 3-8: Graphical depiction for LOSO, LOTO, and LASO CV techniques [5].

Regardless of the type of model and as stated in Section 3.1.2, the physiological data collected during the experiments is segmented into processing windows. These are subjected to overlapping to increase physiological delineation performance, which can strongly affect the interpretation of the results obtained when using the detailed CV techniques. For instance, in case of applying a $k-fold$ CV over a vector of features extracted from filtered, windowed and overlapped physiological signals, there could exist folds, i.e. processing windows, that contains some of the information of the previous fold. This fact can lead to an overoptimistic interpretation of the results

and, although it depends on the overlapping length, it should be avoided. Thus, focusing on emotion recognition by using physiological signals, strategies that do not inquire into this problem, such as LOTO, LOSO, and LASO, might be preferred. Note that the latter recommendation can be affected by the amount of available data for training, which can make it impossible to apply some techniques such as LOTO for subject dependent models.

3.2 Open available databases

Within the affective computing community, different datasets deal with emotion recognition using physiological signals. The most common are MIT [178], DEAP [138], MAHNOB [10], DECAF [145], ASCERTAIN [179], and WESAD [180]. Table 3.1 summarises the main details of such databases. These open databases are considered as a solid benchmark by the scientific community. In this Section, the emotion recognition databases of interest for this research work are reviewed. Note that previously, in [181] and [182], we performed a detailed analysis of some of these open available databases and provided conclusions about their methodologies and emotional recognition approaches. Due to the similarity with respect to some of the emotional elicitation mechanisms, experimental methodologies, and, above all, the physiological information of interest for this work, only two of them are chosen to further perform fear detection proof of concepts based on their signals and stimuli, which is detailed in Chapter 4.

Table 3.1: The most common emotion recognition databases with a laboratory set-up used within the affective computing scientific community.

Database	Subjects (M/F)	Labels	Use Case	Accuracy	Year
MIT [178]	1 (0/1)	Discrete	General	81.00%	2005
DEAP [138]	32 (16/16)	Arousal/Valence	General	57.00/62.70%	2012
MAHNOB-HCI [10]	30 (13/17)	Arousal/Valence	General	46.20/45.50%	2012
DECAF [145]	30 (16/14)	PAD	General	55.00/60.00/50.00%	2015
ASCERTAIN [179]	58 (37/21)	Arousal/Valence	General	66.00/68.00%	2017
WESAD [180]	15 (12/3)	Research-based	Stress	86.46%	2018

On the one hand, the first proofs of concept of this research work were developed using DEAP [138]. This database contains physiological information of 32 volunteers (16 female). The experiment consists of a total of 40 video-clips of one minute duration each. The stimuli were selected from a larger pool or pre-tagging stage based on valence, arousal, and dominance ratings gathered by SAM. The included peripheral

(physiological) sensors are electroencephalogram, electromyogram, respiration amplitude, GSR, electrooculogram, PPG and SKT. Regarding some of the limitations of this dataset, due to the laboratory setup, the volunteers were very limited in terms of movement, and so the trained models are not valid for real-life conditions. It should be noted that a five second baseline recording was done between stimuli by using a fixation cross in the screen. Finally, the authors of the database created three binary systems, each of them inferring low or high level of arousal, valence and liking, in which they used the self-reported ratings as ground truth (labels). They presented these results as a benchmarking and obtained the following average Accuracy (ACC) and F1-score metrics: 57.00% (ACC) and 53.30% (F1-score) for arousal, 62.70% (ACC) and 60.80% (F1-score) for valence, and 59.10% (ACC) and 53.80% (F1-score) for liking. This database is of special interest mainly due to two factors. First, it contains the same physiological information as the bracelet of Bindi. Second, the self-reported labels gathered during the experiments contain PAD space. More technical details about the DEAP affective computing system and the ones proposed after its publication are provided in Chapter 4.

On the other hand, the MAHNOB-HCI dataset includes physiological data from 30 study participants (17 female) [10]. This lab-based emotion recognition dataset contains data for a total of 20 video clips per volunteer, which were selected based on a larger pool or pre-tagging stage as DEAP and were approximately 81 seconds long on average .The recorded physiological responses were acquired using the Biosemi active II system, and they included ECG, GSR, respiration amplitude, SKT, electroencephalogram, eye gaze and face and body videos. As for the DEAP dataset, the laboratory setup makes the trained models no valid for real-life conditions. However, one of the main differences with DEAP is that in MAHNOB the authors considered the emotional recoveries of volunteers between stimuli, rather than just waiting five seconds between them. In fact, before watching any emotional video, different neutral clips were shown to the participants. This process was used to recover a basal physiological level, decrease the emotional bias after experiencing an emotion and, ultimately, handle physiological intra-subject differences. Finally, the authors of the database created two non-binary emotion recognition systems, each of them inferring low, medium and high level of arousal and valence, respectively. To obtain the

ground truth, they used a mapping between the self-reported discrete emotion ratings and the emotional dimensions based on [3]. They achieved an average ACC and F1-score metrics of up to 46.20% and 38.00% for arousal and 45.50% and 39.00% for valence, respectively. This database is of special interest mainly due to three factors: it contains the same physiological information as the bracelet of Bindi, the self-reported labels contain PAD space information, and recovery between stimuli was considered during the protocol experiment.

Despite the benefits that these databases bring to this research work, they are not intended to elicit specifically fear to further detect risky situation in Gender-based Violence contexts. Thus, they can be used to generate and to study proofs of concepts for the fear machine learning engine that this research work is focused on and even provide preliminary conclusions for the wide casuistry within this complex task. However, as commented in previous Chapters, being the disentanglement between physiological reactions and fear under Gender-based Violence situations one of the main goals of this work, a new database might be created to actually target for our specific use case. Moreover, such database might use VR to provide stronger emotion elicitation immersive experiments. More details regarding the database created during this research work and its particularities are given in Chapter 6.

3.3 Conclusion

In this Chapter, we have provided a complete review and analysis for the emotion recognition databases generation and processing from an experimental point of view to the data processing procedures that can be applied after the database is finished.

First of all, we concluded that there is not an standard protocol for stimuli analysis and selection. All the public available datasets are thought from a general emotional perspective, i.e. with the goal of identifying emotions in general without targeting binary specialised emotional models. This fact makes the evaluation of the stimuli by experts not so critical. But, for research works like the one being addressed in this document, this strategy can not be applied and may not be suitable. The stimuli-conditioned situation of Gender-based Violence Victims, as well as their possible PTSD episodes, make neccesary the help of experst to adjust and select the stimuli to be presented during our experiments. Secondly, an exploratory data analysis

is strongly recommended to determine some of the physiological behaviour and to carry out specific actions to deal with some problems, such as physiological recovery from emotion elicitation. Third, we have detailed different recommendations for the CV techniques to be applied when dealing with emotion recognition problems. This fact is of special relevance due to the inter and intra variability that can exist between the different volunteers in such experiments. Thus, new CV techniques that consider intra and inter variability, such as LASO, are preferably selected to be used an applied over common techniques.

The research work presented in this document deals with the proposal, study, design and implementation of a new emotion recognition database, fear machine learning design, and wearable edge device development. This makes the knowledge of this Chapter essential to understand the topics in the following Chapters.

Part II

Fear classification using the State-Of-The-Art

Chapter 4

Fear classification Proof-Of-Concept

Once reviewed the whole state of the art regarding emotions, physiological information, databases for emotion recognition, and the different post-processing procedures to design a fully tested machine learning model, we will apply such knowledge to design different fear binary emotion recognition systems using the two databases detailed in Section 3.2. Specifically, in this chapter, the proposed architectures are solely based on the physiological uni-modal part of Bindi, considering the description of Bindi in Section 5.2. Thus, these proposals are intended to boost the first embedded implementations of the whole data processing chain, including the machine learning engine, within the smart bracelet of Bindi. Note that the multi-modal casuistry and possibilities are dealt and detailed in Chapter 6.

In the following sections we will start by approaching three initial systems developed upon the DEAP database. Thereafter, due to some limitations observed in DEAP, the MAHNOB database will be used to design another two fear binary emotion recognition systems. Finally, all the generated performance metrics will be compared with respect to the current state-of-the-art regarding emotion recognition and, more specifically, fear detection. Moreover, to contextualise the scope of the obtained results, key aspects such as class balance, feature selection, and other processes are dealt and discussed. Note that the different systems presented in this Chapter were designed and validated on a personal computer. Specifically, Matlab® was used as the software platform and all the developed code took Toolbox for Emotional feAture extraction from Physiological signals (TEAP) [148] as a reference, which is a current available open source toolbox for physiological data processing and

feature extraction. In fact, we have been in contact with the developers of TEAP and contributed with fixes to their repository. Finally, a new fully automatised toolbox from that basis has been developed, which accounts from signal pre-processing to machine learning training and testing. This tool has been applied to design the systems presented in this Chapter as well as for different experiments with other datasets and projects within the department under which this research work has been carried out.

The proposed emotion recognition systems are ordered from lower to higher complexity within this Chapter. In such a way, the research and development strategies followed along this work have fed in an incremental manner the different implementations carried out with Bindi. Thus, regardless of the architecture complexity, most of them are based on the components shown in Figure 4-1, which depicts an overall and general description of the training for the proposed fear recognition system. It includes the typical steps in the processing chain discussed in Chapter 3, from the analysis of the physiological signals dataset to raw data pre-processing, feature extraction and emotion classification. In fact, most emotion recognition systems in the literature follow this architecture but focusing on classifying emotions from a general-purpose point of view by detecting a set of emotions without considering if the user is male or female [133]. However, targeting the identification of a single emotion that could be related to a specific situation and considering gender-related particularities might be exploited towards a more accurate system. This last assertion is based on the idea that women recognise nonverbal communication or emotional prosody more accurately [54], as reviewed in Section 2.3.3. These concepts are not considered in any current emotion recognition system using physiological signals presented in the literature. Currently, up to my knowledge, there is no affective computing detection system developed to identify different critical social situations, such as Gender-based Violence episodes. Within this context, a specialised fear detection system could be designed to trigger a protection protocol that could include a connection to a trusted circle or even to law enforcement agencies, in order to provide immediately the necessary help. The latter is one of the main goals of Bindi, as stated in Chapter 1. Thus, the added value of the proposed architecture in this Chapter is twofold: (1) the generation of a first proof of concept for a specialised

fear binary recognition system by using solely physiological information (so far the state-of-the-art approaches deal with several emotions), and (2) the consideration of digital processing constraints to further properly adapt such system to be integrated into a wearable edge-device platform for allowing protection of vulnerable people.

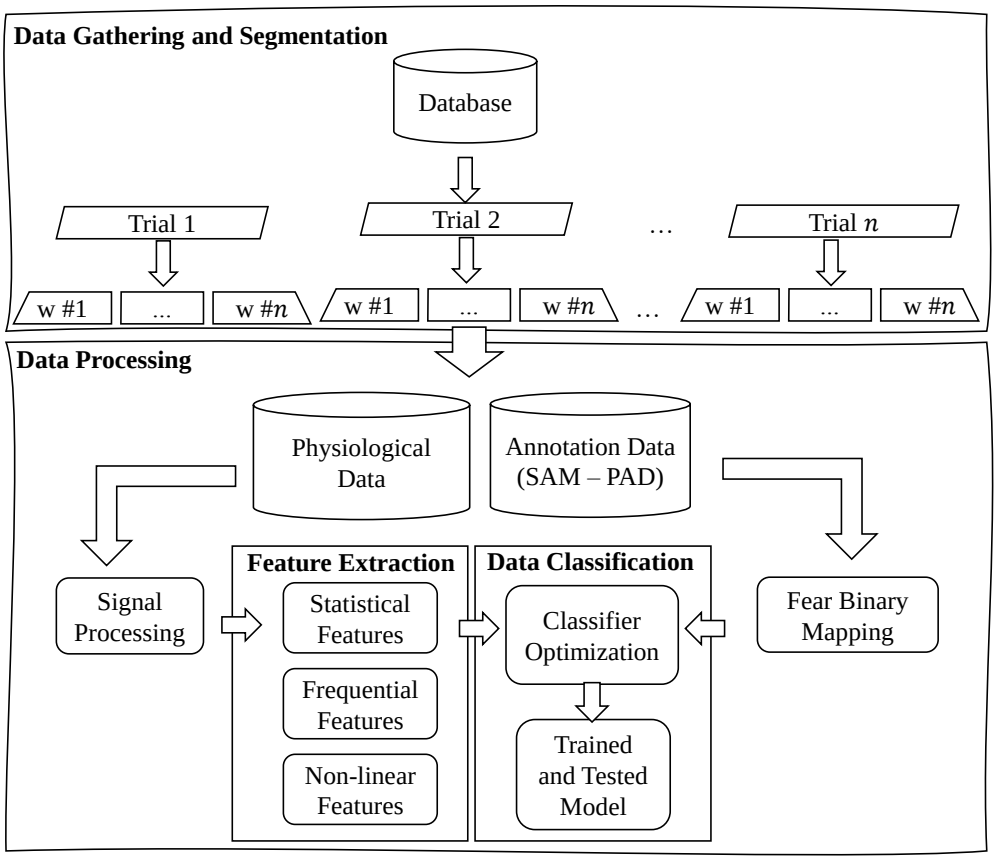


Figure 4-1: Overview of the training process for the proposed fear recognition system employing physiological sensor data and PAD dimensional approach emotion labelling. The latter is fed into the fear binary mapping procedure. Note that $w\#n$ denotes the different windows obtained after data segmentation if applicable.

It is noteworthy to highlight two specific design considerations regarding the type of generated models and the specific embedding constraints of Bindi. On the one hand, throughout the design of these initial proof of concepts, the need for a robust and reliable subject-independent model has been something tackled and chased. The design of a fully subject-independent model would allow boosting the initial deployment process of any technological tool able to detect any emotion using machine learning. This fact can be specified in Bindi for fear detection under Gender-based Violence situations. This consideration is essential to understand some of the decisions made and parameters studied within this and the following Chapters. Note that such subject-independent system, which is deployed at an initial configuration

process, is subjected to be later customised and personalised for the specific subject to improve performance, as stated in Section 3.1.7.2. On the other hand, the design process of all the different fear binary emotion recognition systems presented in this research was also biased by specific embedded resources and capabilities constraints. These were fixed as 64 kB of RAM and 512 kB of Flash. Note that such resources were imposed by the research team to narrow down the design to a lightweight implementation, however, different limitations can be set considering the respective performance improvement or worsening. More specific details for the embedded implementation are given in Chapter 5.

4.1 Fear classification using DEAP

As detailed in Section 3.2, the DEAP database is one of the most used databases in the literature regarding emotion recognition with peripheral or physiological signals. Although it is not a fear specialised dataset, i.e. the different stimuli were selected from a general emotional perspective without focusing specifically on any particular emotion, it contains the necessary elements for us to design the first proof of concept for the fear detection system based exactly on the same physiological signals of our interest, i.e. PPG, GSR, and SKT. DEAP contains data of 32 participants for a total of 40 video clips, which were selected based on a pre-tagging stage following arousal, valence and dominance ratings. However, it should be highlighted that the measurement equipment of this database was Biosemi ActiveTwo system[1], which is a professional measurement kit thought to be employed in laboratory conditions. This fact makes the acquired signals to be far from real measurements obtained with wearable devices. Thus, the proposed systems presented here serve as initial proof of concepts and allowed us to identify different key aspects to be considered when both designing a database and training a machine learning model from such data.

Regarding the specific methodology followed during the DEAP experiments, Figure 4-2 shows a simplified diagram for the experimentation applied for every volunteer and each stimulus. Note that the 2-minute baseline was applied just at the beginning of the experiment. From this figure, a very short transition is appreciated between consecutive stimuli and, therefore, between two elicited emotions. This

[1] http://www.biosemi.com

fact can strongly affect the emotional state of a volunteer, and hence the physiological recovery, before the next video clip. Moreover, a mandatory break was performed at the half of the experiment (stimulus number 20), during which cookies and non-caffeinated, non-alcoholic beverages were offered to the volunteer. This experimental methodology may introduce a very harmful bias depending on the order of the stimuli and their targeted emotion. In the following subsections, we analyse, from a physiological point of view, the possible effects detected in the collected data and labels. Regardless of these facts, up to my knowledge, DEAP was the first database that proposed well-documented selection stimuli and experimental laboratory methodology, together with a relatively high number of volunteers, and made everything fully open access.

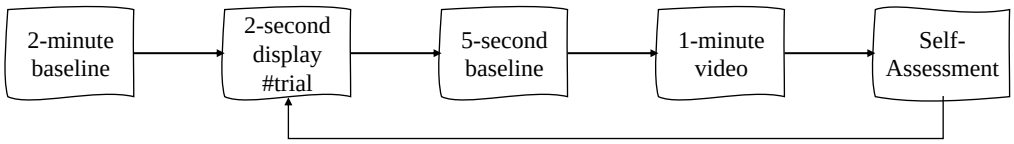


Figure 4-2: Simplified diagram for the experimentation applied for every volunteer and each stimulus for DEAP database.

Before going into details for the analysis performed and systems designed using the DEAP data in this research work, a review of the data processing and machine learning techniques applied by the original work of the database, and by subsequent research using it, might be provided. The authors of the original work of the DEAP database applied basic pre-processing procedures to remove the temporal low frequency drifts of some signals and smooth them by using moving average filters. They extracted 106 physiological features and employed a filter feature selection method to use only the highest-ranked ones. Specifically, they applied the Fisher linear discriminant score given by equation 4.1,

$$J_f = \frac{|\mu_1 - \mu_2|}{\sigma_1^2 + \sigma_2^2}, \tag{4.1}$$

where μ_k and σ_k represent the mean and variance of the class k for each feature f. Note that this equation is valid for $k = 2$, i.e. a binary classification problem. The bigger this score, the more important that specific feature will be. Thus, the goal is to maximise the score to obtain a large between-class variance (numerator) and a small within-class variance (denominator). However, this methodology neglects

the combination of features and neither handle redundant ones, which leads to a sub-optimal selected feature space with empirical threshold discrimination. For the classification, they used a Gaussian naïve Bayes classifier for a two-class problem and three different use cases, low and high levels for arousal, valence and liking. This specific classifier is characterised by being a generative model, i.e. it has a high bias and a low variance derived by the assumed Gaussian distributions learnt from the features, which can produce under-fitting issues. The output of such classifier considering N classes is provided by equation 4.2,

$$y = argmax_y P(y) \prod_{i=1}^{N} P(x_i|y), \tag{4.2}$$

where we can obtain the inferred class y for a given set of features or feature vector x_i. Note that this classifier makes two key assumptions by considering that features are independent and normally distributed. Being the latter given by

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(- \frac{(x_i - \mu_y)^2}{2\sigma_y^2} \right), \tag{4.3}$$

where μ_y and σ_y^2 are the mean and the variance of the values in x associated with class i. Lastly, the CV applied was LOTO considering the 40 audiovisual stimuli used during the experiments. By employing all the peripheral signals, they provided average Accuracy (ACC) and F1-score metrics and obtained 57.00% (ACC) and 53.30% (F1-score) for arousal, 62.70% (ACC) and 60.80% (F1-score) for valence, and 59.10% (ACC) and 53.80% (F1-score) for liking. Note that they did not provide the associated standard deviations of such average values.

Since the release of the DEAP database, different machine learning systems have been proposed in the literature using its data. Some publications studied the feature importance by applying different methods and improving such process. For instance, the authors in [183] considered the same classification problem as the authors of DEAP, but used recursive feature elimination to tackle the mutual and redundant information. They applied a SVM classifier and obtained 66.36% (ACC) and 63.99% (F1-score) for arousal, and 68.71% (ACC) and 63.25% (F1-score) for valence, which surpassed the original DEAP work. They used all the features from all modalities including those extracted from non-wearable-ready sensors. In fact,

they concluded that the electroencephalogram signals were playing a key role into the classes separation (distinction between classes). Although their use case can not be directly extrapolated to ours, as we are just based on three peripheral signals and towards a fear binary detection system. The improvement of the results in comparison with the original work due to the application of less restrictive feature selection techniques and a discriminative classifier is valuable and can help in our research.

There are also other publications that did not consider the whole set of signals and, instead, they reduced their number towards a more wearable-ready concept. For instance, [153] is one of the latest emotion recognition systems based on DEAP. The authors designed a five class (Happy, Relaxed, Disgust, Sad, Neutral) emotion recognition system using the PA model. They applied a feature fusion level technique by leveraging a deep belief network architecture together with conventional statistical feature extraction over only three physiological signals (PPG, EDA, and EMG). Finally, they trained a SVM classifier and obtained up to 89.53% average accuracy for a subject-independent model following a LOSO configuration, which outperformed the state-of-the-art. In their work, they did not consider any real implementation constraint related to the data segmentation, frequency resolution, storage and complexity applied or needed. Also, they only took the physiological data recorded during the last 20-seconds of every stimulus based on their hypothesis that the emotional immersion was greater at the end of the video clip. This hypothesis has not been demonstrated with a statistical, objective and/or quantifiable method, but just assessed by physiological visual exploration.

Among the rest of the research conducted on the DEAP database and regarding specifically the fear recognition use case, four systems are found in the literature. On the one hand, the first two [182, 184] are our publications and they are detailed in the following subsections. These will be hereinafter referred to as DEAP-b1 and DEAP-b2 for this and the following Chapters. On the other hand, the authors in [185] and in [186] employed the same binary fear paradigm that is described in Section 2.3.4. In [185], they used all the DEAP volunteers and all the available physiological signals, including the ones providing electrooculogram and electroencephalogram data. By performing a Design Space Exploration (DSE) for different

feature selection techniques, as well as for nine different classification machines, including deep neural networks, they achieved up to 90.07% average accuracy for a subject-independent model without feature selection, just by using the filtered data, and following a Hold-Out strategy with a 70/30 train-test split ratio. Note that [185] was published after our fear binary emotion recognition work, DEAP-b1 [184], which, up to my knowledge, was the first research that applied such fear labelling paradigm to emotion recognition through physiological signals. In [186], they took our research from [182,187] as main reference and elaborated a comprehensive analysis that comprised a detailed study of the effects for fear binary emotion recognition when using different machine learning elements methods, and techniques. In contrast with the research in [185], in this one they applied filtering stages, data segmentation with and without overlapping, feature selection, dimensionality reduction, and imbalanced adjustment with Synthetic Minority Over-sampling TEchnique (SMOTE). They also relied upon the DEAP volunteers data, but discarded most of the signals and just employed GSR and PPG. Finally, they used 20 second data processing windows and achieved a maximum fear recognition accuracy rate of up to 93.50% for a SVM classifier together with PCA by considering a non-overlapping strategy and 5 $k-fold$ CV. Although this latter research is a valuable work towards exploring the wide DSE regarding the fear recognition, their main limitation is the CV technique applied as there could exist folds, i.e. processing windows, that contains some of the information of the previous fold, see Section 3.1.7.3. Thus, this fact may lead to overoptimistic results.

4.1.1 Stimuli balance and labels considerations

As stated in Section 3.1.1, one of the common approaches followed during the generation of a database is related to the stimuli balance assessment. This is referred as to the statistical representation of the different classes. For instance, in any classification problem is desirable to have the same amount of instances for all the classes. Otherwise, the classification algorithm could derive into favouring the learning of the class with greater representation compared to the rest of the classes. Thus, the analysis of the labels during the whole generation and processing of the database is essential to contextualise and understand both the emotion elicitation and the obtained results.

In DEAP, the 40 video clips used during the experiment were chosen from a larger stimuli pool. A pre-tagging stage started with 120 video clips and gathered around 14 ratings per video. After that process and by using equation 3.1, the authors selected the videos that were located on the extreme corners of the normalised quadrants within the PA space, which resulted in a set of 40 video clips with extreme labelling and used for provoking emotions on the volunteers while measuring their physiological signal. After visualising these video clips, the volunteers labelled the emotion felt by them. Thus, this methodology led to the generation of two different sets of labels, the ones from the pre-tagging stage and those self-reported and recollected during the experiment. Generally, the latter are the preferred ground truth for training machine learning models that are based upon the gathered physiological and/or physical data. However, the distribution of the self-reported labels can be very different with respect to the pre-tagging stage labels. For instance, Figure 4-3 depicts the differences in pre-tagging and self-reported labels for the DEAP database and the selected video clips. For the pre-tagging stimuli, the obtained labels are categorised with a different symbol based on the normalised quadrant location (Q1 - positive arousal, positive valence -, Q2 - positive arousal, negative valence -, Q3 - negative arousal, negative valence -, Q4 - negative arousal, positive valence). On the contrary, the self-reported labels are represented using the same symbol and colour. Note that the authors of DEAP selected the stimuli that achieved the highest means and smallest variations among the different ratings. As it can be observed, the pre-tagging ratings do not follow the same distribution as the self-reported ratings, which even results into the same stimuli being located in different emotional quadrants, for instance the stimulus 83.

Table 4.1 presents the videos that are in a different quadrant. Without even considering those with different locations within the same quadrant and those that are right on the frontier lines, 20% of the stimuli are not evoking the targeted or pre-tagged emotion during the realisation of the experiments. This is translated into distinct label distribution and so can lead to system performance differences when training with pre-tagging or self-reported labels. However, the stimuli related to the second quadrant (Q2), based on pre-tagging ratings, are the only ones presenting a complete agreement in comparison with the self-reported labels. Thus, it can be

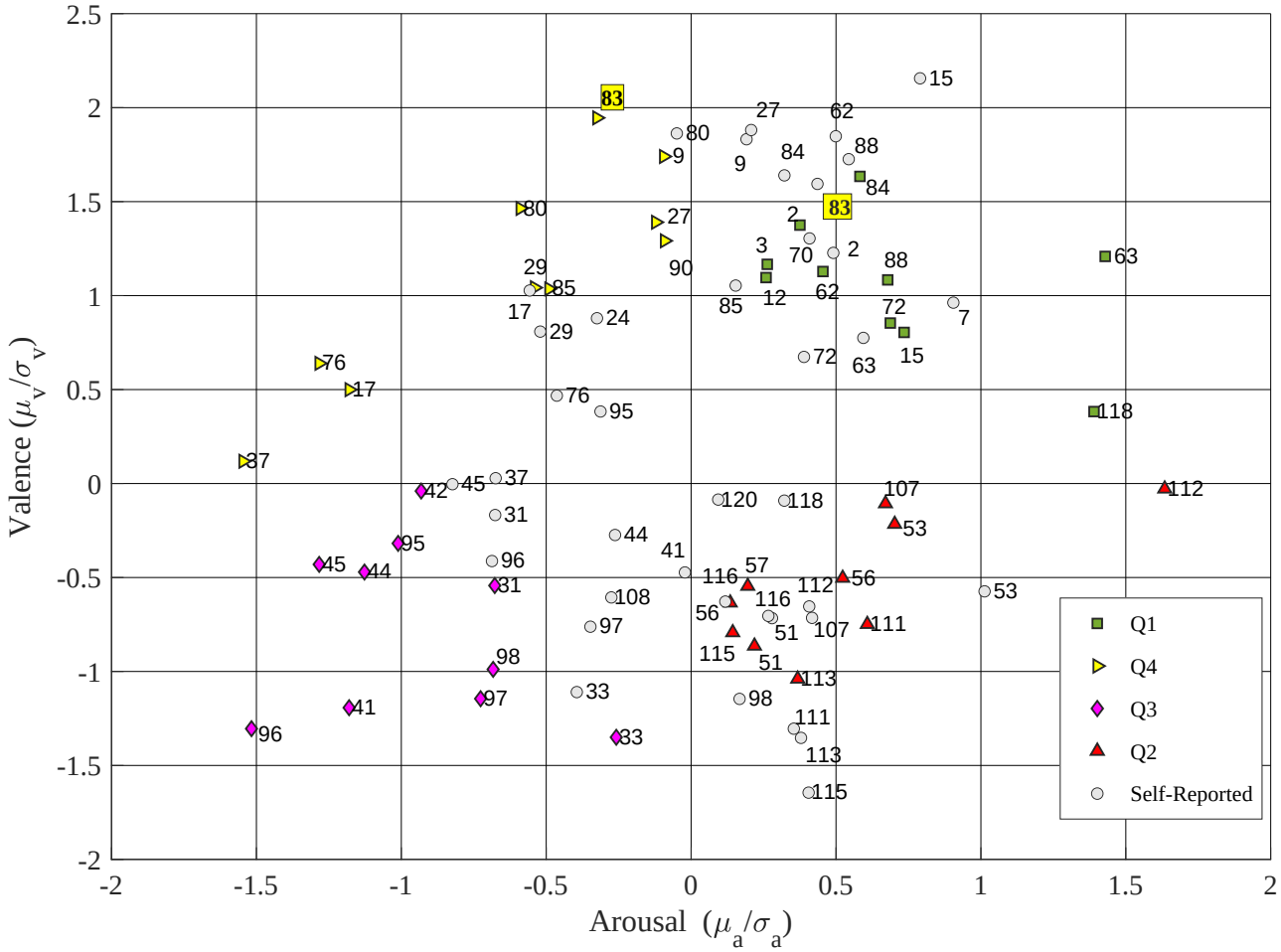


Figure 4-3: Labelling differences for the DEAP database and the original numbering for the selected video clips.

concluded that the stimuli intending to evoke negative emotions (high arousal, negative valence), such as fear-related emotions, are less prone to be confused with other quadrants. Although this conclusion strongly supports an approach for a 2D-based labelling strategy when facing the development of a fear binary recognition system, the consideration of more dimensions might be considered and explored as well.

From a 3D perspective and considering the PAD model, the self-reported ratings and the Fear mapping proposed in Section 2.3.4, only five stimuli are located within the fear-related emotional cube (51, 98, 111, 115), Figure 4-4. This fact is specially relevant due to disentanglement of the stimuli contained within Q2 in Figure 4-3. Thus, from a subjective-rating point of view, the expansion to one more dimension (dominance) can benefit the emotional location of stimuli and so their posterior label assignment. However, such benefit is overshadowed in this specific case by knowing that those five stimuli represent just 12.5% of the total amount of video clips. This latter consideration represents a very imbalanced situation in the case of a fear

Table 4.1: video clips that are in a different quadrant regarding the pre-tagging versus the self-reported labels.

Stimuli ID	Pre-tagging Quadrant	Self-Reported Quadrant
9	4	1
27	4	1
45	3	4
83	4	1
85	4	1
95	3	4
98	3	2
118	1	2

Table 4.2: Self-reported imbalanced ratios for the DEAP database.

Low:High Class	Arousal	Valence	Dominance	Liking
Imbalance Ratio	1.4:1	1.2:1	1:1.6	2:1

binary emotion recognition system, but it can be mitigated by using oversampling techniques over the minority class, as explained in the following subsections.

To provide a specific quantification for the self-reported imbalanced distribution, Table 4.2 shows the self-reported imbalanced ratios for every collected type of label. On the one hand, these imbalance ratios are calculated based on a two-class problem by dividing each dimension into two levels (High and Low), which is equivalent to what was done in the original publication of DEAP. On the other hand, as the stimuli selection in DEAP was done exclusively based on arousal and valence, the resultant imbalance for the other two gathered ratings (dominance and liking) is higher. For instance, as this database was not focused on eliciting negative emotions, the imbalance ratio observed in dominance indicates the presence of more positive stimuli in which the volunteers rated a high degree of control over the evoked emotion.

Nevertheless, in the case of dealing with the design of a fear binary emotion recognition system, in which a binary transformation from the PA or the PAD spaces is performed rather than targeting multiple-level classification using one-dimensional models, the imbalance ratios are detailed in Table 4.3. Such imbalance ratios must be contextualised based on the individual balance obtained through the specific self-reported ratings. For instance, Figure 4-5 shows the class balance for each volunteer after having applied the fear binary mapping from a PA space, which resulted in '1' or the positive class for the Q2 (high arousal and low valence) and '0' or the negative class for the rest of the quadrants. Note that a 25% threshold is highlighted as a

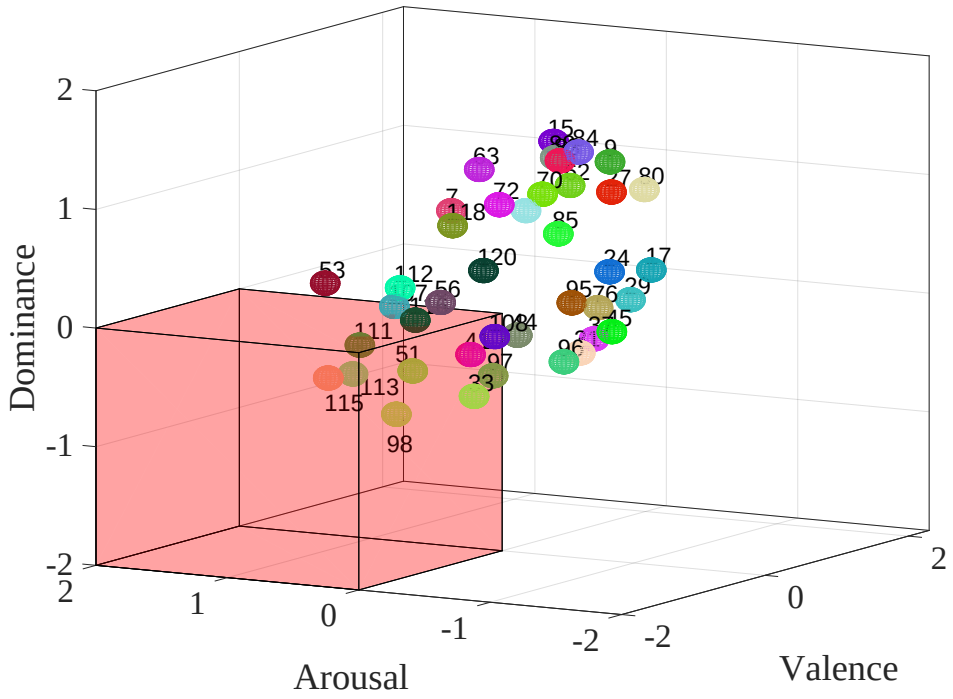


Figure 4-4: PAD model for the self-reported labels of the volunteers. Fear mapping proposed in Section 2.3.4 is marked with coloured cube.

reference mark indicating that, based on the original ground truth expected from researchers, this amount of stimuli should evoke an emotion located into Q2. As it can be observed, 17 out of the total volunteers (32) present less than the expected threshold within the positive class, which strongly affects the imbalance ratio for this binary mapping. Moreover, there is even a volunteer (23) that did not report any rating into the positive class. These facts, besides explaining the high imbalance ratio obtained, give an understanding of the complexity of the inter-individual differences. Nonetheless, the average class percentages considering the 32 volunteers are up to 76.50% and 23.50% for the negative and positive classes respectively. Note that the average positive class percentage is close to the expected 25%. This latter fact supports the conclusions obtained with Figure 4-3, by which we claimed that a 2D-based labelling strategy can be approached to design a fear binary emotion recognition system using DEAP. For the applied binary mapping when considering a PAD space, the positive class is determined by low dominance, high arousal and low valence, whereas the negative is given by the other possible combinations. Figure 4-6 shows the class balance per subject in such case. Note that in this graph there is

Table 4.3: PA and PAD imbalance ratios for the DEAP database.

NoFear:Fear Class	PA	PAD
Imbalance Ratio	3.2:1	6.3:1

not expected threshold to be achieved as the pre-tagging stimuli selection stage was solely based on arousal and valence. In this case, there are three volunteers (23, 27, 28) that did not exhibit any positive class rating and the average class percentages are 86.33% and 13.67% for the negative and positive classes respectively. These facts clearly explain the higher imbalance ratio with respect to the PA binarization and indicate that for this database the balance of the dominance dimension was not crucial. The latter is essential for our use case due to the need for distinguishing between specific emotions that only differ in the dominance dimension, such as fear and anger, as commented in Chapter 2. Nonetheless, despite this problem, a fear binary emotion recognition system using this database might be explored as a proof of concept. Moreover, as already pointed out previously, different oversampling techniques can be applied to deal with such extreme imbalance conditions.

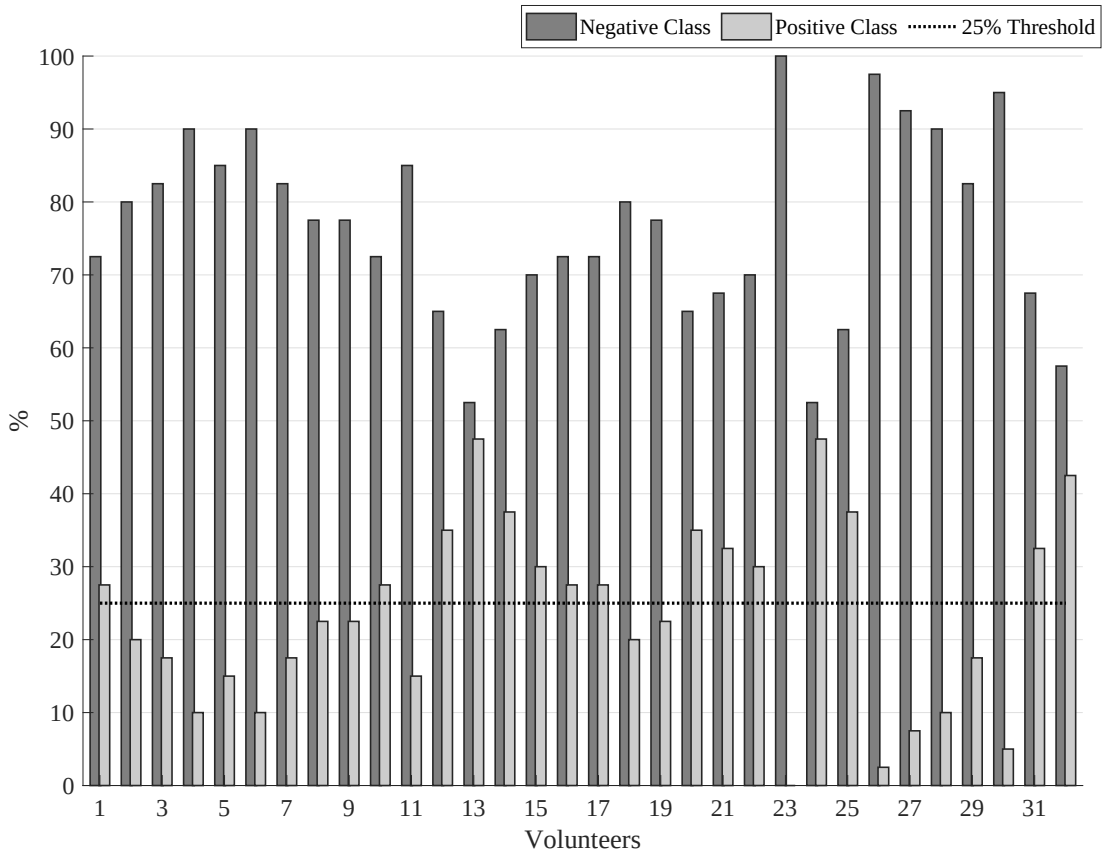


Figure 4-5: Class balance per volunteer after having applied the fear binary mapping from a PA space.

Another essential process when assessing the labelling consistency over the different

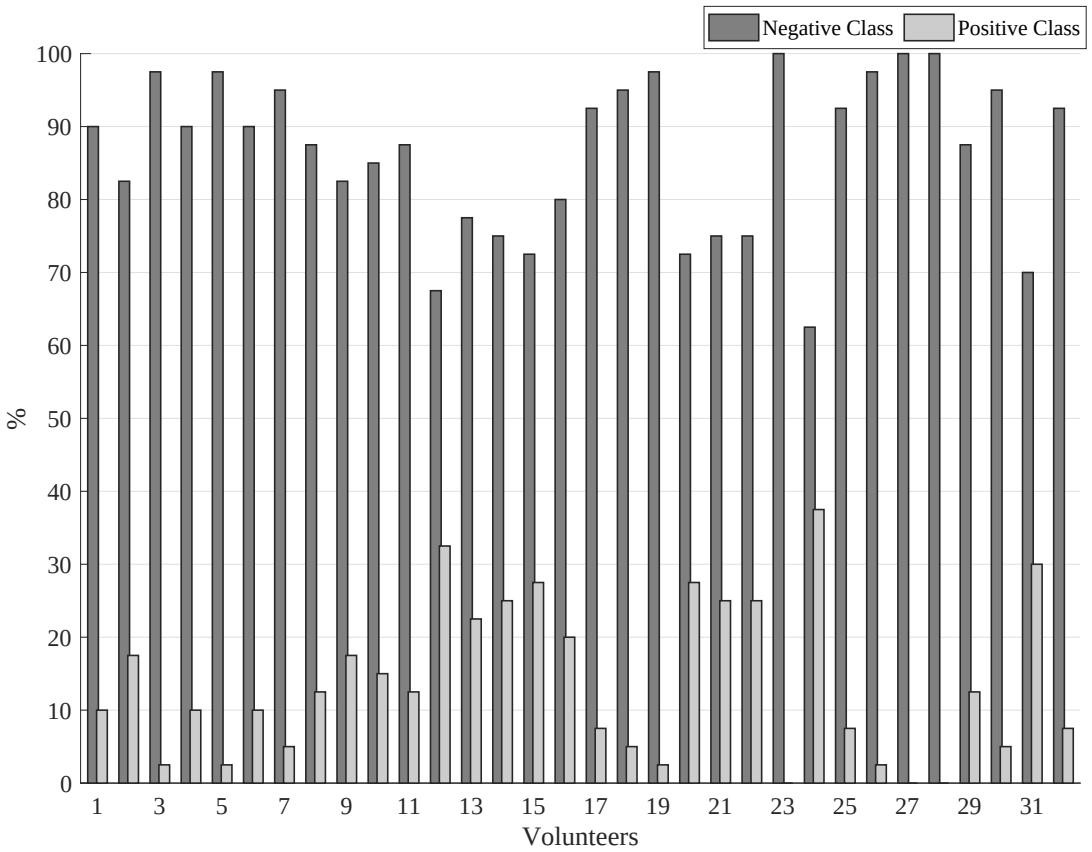


Figure 4-6: Class balance per volunteer after having applied the fear binary mapping from a PAD space.

volunteers is to observe the label inter-individual correlations. Such a task provides information that can be directly and further related to the obtained results from the different machine learning models. For instance, considering both fear binary mappings performed, i.e. from PA and from PAD, the results obtained after a Levene's test and a Kruskal-Wallis test rejected the null hypothesis that the variances are equal across all volunteers ($p < 0.001$). Note that both sets of binarized labels exhibit a non-normal distribution and that the significance level was set at $p < 0.05$. These facts lead to the evaluation and application of correlation and independence tests to study the labelling behaviour of the different volunteers. Thus, Figure 4-7a and Figure 4-7b show the averaged $p - values$ for the Spearman correlation and the Chi-square test of independence for the PA fear binary based mapping, respectively. Note that both are non-parametric methods to assess the different associations between variables. However, the former responds to monotonic associations, while the latter provides information related to the independence of the variables considering any type of association. The results given by both processes are close, in fact,

both of them fail to reject the null hypothesis. This indicates that no statistical difference exists between the different groups, i.e. the correlation is considered not significant and the different variables are independent. Therefore, we can conclude that there is not enough evidence to suggest that an association between the binary label of the volunteers exist. Moreover, some volunteers (4, 8, 16, 21, 26) show high $p-values$ in comparison with the others, which can be interpreted as a stronger decorrelation and independence of their labels. Figures 4-8a and 4-8b present the averaged $p-values$ for the PAD fear binary based mapping and the same correlation and independence tests. In this case, we can observe a stronger decorrelation and independence for the different volunteers in comparison with the previous tests. This gives insight into the label dataset distribution and can even guide the design process. Thus, two main conclusions can be extracted from the comparison of these figures and from the consistency study of the labels: (1) these processes allow identifying the volunteers that provided very distinct labelling during the experiments, and (2) although the PAD space provides more information in terms of emotional modelling, the more dimensions added the lesser agreement could be obtained from the self-reported ratings of the volunteers.

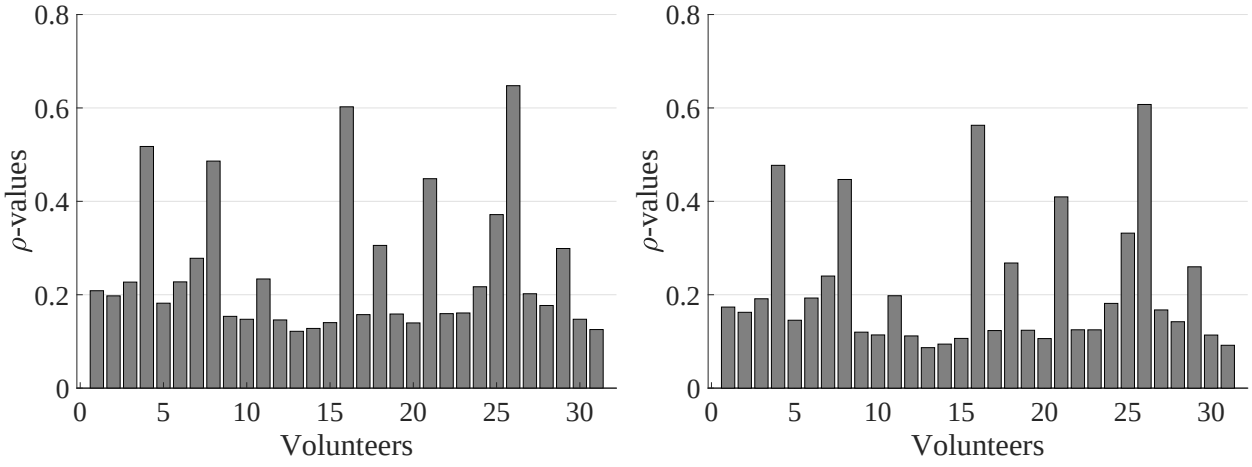


Figure 4-7: Averaged $p-values$ for all considered volunteers and their labels applying: a) the Spearman correlation, and b) for the Chi-square test of independence. In this case, the labels are binarized using the PA fear binary based mapping.

Despite the balance and agreement differences observed when applying the fear binary transformation from both emotional models, the usage of the dominance dimension to properly distinguish the fear emotion, led us to design DEAP-b1 by using the fear binary transformation from the PAD space. The results obtained

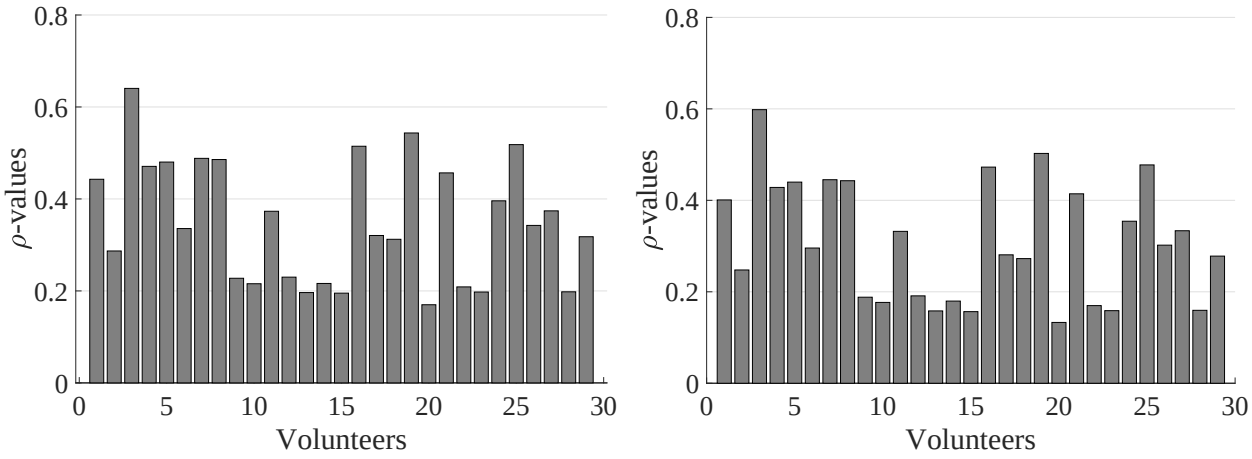


Figure 4-8: Averaged $p-values$ for all considered volunteers and their labels applying: a) the Spearman correlation, and b) for the Chi-square test of independence. In this case, the labels are binarized using the PAD fear binary based mapping.

from that system, when considering a subject-independent perspective, were not promising [184]. Thus, we decided to simplify the problem by designing a system using a fear binary mapping from the PA space [182], DEAP-b2. The latter improved the first results and proved that a fear-related binary emotion recognition system was feasible using solely physiological information. Specific details of each of these systems are given in Section 4.1.4. Moreover, it should be noted that the different results gathered from the stimuli balance and label consideration study provided in this section were always present during the design of such systems.

4.1.2 Exploratory data analysis and filtering processing

During the generation of the DEAP database, the researchers incurred some issues affecting the sensor acquisition for some volunteers and the designed physiological recovery. In this Section, we generated different plots synchronised with the experimental methodology to perform and exploratory data analysis and assess the behaviour from both a physiological and a sensor functioning perspective. For instance, Figure 4-9 shows an example of one of the graphical representations for the physiological visual assessment performed during this step. Specifically, the plots are the full experiment for volunteer number 22. The represented signals are GSR, BVP, and SKT, from high to low order respectively. The Synch Signal indicates the different states of the experiment: 20 stimuli represented by each saw tooth and 20 labelling assignment tasks in each decay of those, which are followed by a break and

the final 20 stimuli with their respective labelling. It should be noted that the displayed data was obtained directly from the ".bdf" (BioSemi's data format generated by the Actiview recording software) files provided by the database. Note that they also uploaded a preprocessed version of the data, however, only applying a downsampling with no other additional filtering stage. Thus, from the visual exploratory analysis of all the participants raw data, we got three main conclusions:

- While GSR and SKT showed an acceptable quality, BVP needed to be filtered to remove not just the high frequency noises but also the baseline wander, i.e. the very low frequency trend that is produced by the respiration effect on the PPG acquisition.
- The expectation for a non-controlled physiological break (after stimulus 20th) stated during the analysis of the database methodology in Section 4.1 is confirmed at first glance at least by the GSR. Note the tonic level increment during the break and that this behaviour is repeated along with all the volunteers. From a laboratory database perspective in which the conditions should be properly controlled, this type of recovery can lead to unknown effects for the emotion recognition systems to be trained. Thus, in case of performing a physiological recovery or break, other strategies might be applied which accounted for an actual stabilisation or detrending of the physiological signals.
- Physiological skin temperature inconsistencies were observed for different volunteers. Those were referred to very low skin temperature values. For instance, the SKT signal in Figure 4-9 presents a variation from 29 ^{o}C to 25 ^{o}C, which is not within the normal and/or valid SKT ranges under controlled laboratory conditions. This problem can be due to different factors, such as sensor acquisition malfunctioning or wrong sensor attachment to the body.

To tackle the different noise problems observed in the BVP, different filters can be designed. On the one hand, the high-frequency noise can be filtered out by a direct-form low-pass Finite Impulse Response (FIR) filter. On the other hand, the residual baseline wander or low frequency drift effect presented in the signal can be removed using a forward-backwards low-pass Butterworth Infinite Impulse Response (IIR) filtering stage [188]. Specifically, the forward-backwards technique handles the nonlinear phase of such filters. For instance, an example of the application of these

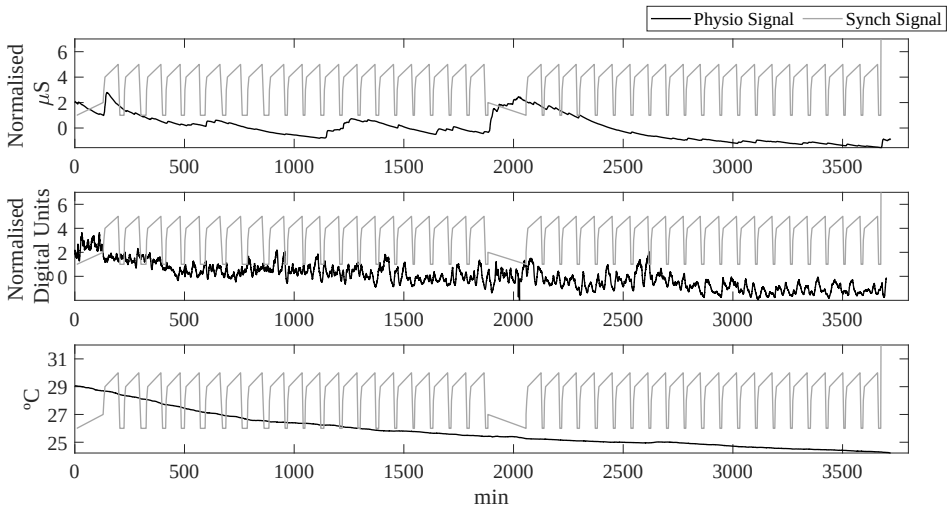


Figure 4-9: Example of one of the graphical representations for the physiological visual assessment performed.

different filtering processes for one specific fragment of the previous signal is shown in Figure 4-10. It should be noted that these two detailed filtering stages can be independent, i.e. they do not have to be strictly applied one after another. In fact, we only employed the FIR filter for DEAP-b1 and DEAP-b2. This consideration was based on the fact of observing low-frequency drifts within the signal during the feature extraction. The designed FIR filter resulted into a 3.5 Hz cut-off frequency with -6 dB attenuation. Note that at Hamming window was used during the design process to properly minimise the first side lobe.

Regarding the SKT problem with some of the volunteers, 11 out of 32 volunteers were affected. Thus, only 21 valid volunteers were considered for DEAP-b1. However, after that, the need for increasing the dataset led to considering the complete set of volunteers for DEAP-b2 at expense of omitting SKT and using just GSR and BVP.

As already highlighted in Chapter 3, most of the public available emotion recognition databases do not deal with exhaustive exploratory data analysis during or after the dataset generation. This fact can lead to unexpected behaviours when designing emotion recognition systems. Thus, this step is needed to guarantee the quality of the provided data.

4.1.3 Feature extraction

The design of DEAP-b1 was our first fear binary emotion recognition system [184], which followed an approximate computing approach by applying no feature ex-

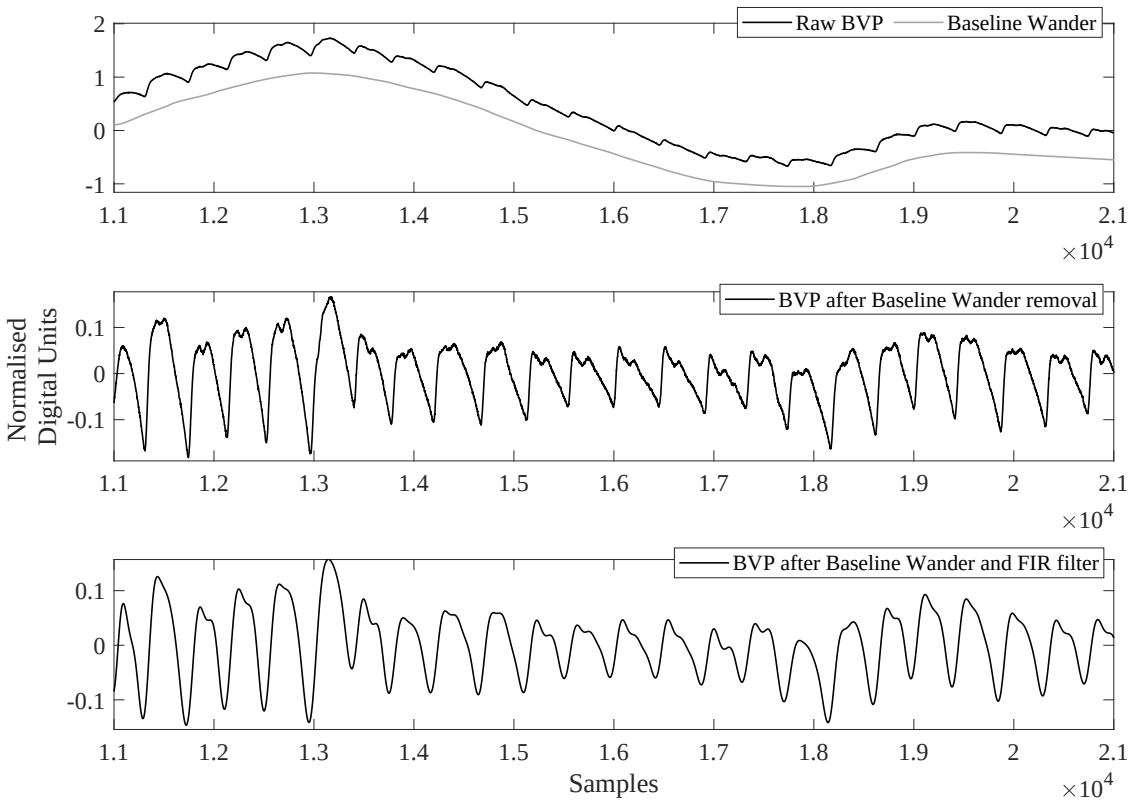


Figure 4-10: Filtering example for baseline wander extraction and removal through IIR filtering, and high noise removal using .

traction procedure, i.e., reducing the system complexity in exchange for decreasing accuracy [189]. Instead, we only considered the filtered value of each physiological variable. This brute-force-like focus is opposite to other approaches, which employed more than a hundred features extracted from physiological signals, such as DEAP original work [138]. The decision of such a bare-metal feature extraction system was led by an initial exploration of the design possibilities when dealing directly with filtered and/or raw data, and by the limited embedded resources that the first version of Bindi presented (64KB of RAM). Note that other recent publications, such as [185], also used directly the raw data to generate fear machine learning engines.

On the contrary to DEAP-b1, our second proposed fear binary recognition system, DEAP-b2 [182], implemented conventional feature extraction techniques usually employed in the literature. They were extracted using the entire video clip duration, i.e., 60 seconds processing windows. Table 4.4 presents the complete list of features for the two physiological signals considered in this system. Specific details and features rationales are provided in the following subsections.

Different delineation processes needs to be applied to obtain specific physiological

Table 4.4: Features extracted for DEAP-b2 system.

Sensor	Domain	Features
PPG/BVP (13)	Time-domain: (3)	Average of filtered signal Mean of Inter-Beat-Interval Heart Rate Variability
	Frequency-domain: (5)	Power spectral density of four bands (0–0.1 Hz, 0.1–0.2 Hz, 0.2–0.3 Hz and 0.3–0.4 Hz) Inter-Beat-Interval spectral density ratio between 0–0.08 Hz and 0.15–0.5 Hz bands
	Non-linear domain: (5)	Inter-Beat-Interval Multi Scale Entropy (five levels)
GSR (7)	Time-domain: (7)	Average of filtered signal Number of ERSCR peaks per second Average relative amplitude of ERSCR peaks per second Average rise time of ERSCR peaks per second Standard deviation of filtered signal 25th percentile value 75th percentile value

points for every signal before extracting the features of the signals, as explained in Section 2.5. For instance, the BVP signal requires peak and valleys identification. For the DEAP-b2 system, this is done by implementing the same BVP delineation algorithm that is proposed in [148]. For the GSR signal, the tonic and phasic components, SCL and SCR, must be extracted as well. In this case, we assumed a linear combination of these two components represented in equation 2.7. The trend of the GSR signal is obtained by a moving median filter with a $+/-$ four seconds sliding window, which is based on replacing each entry with the median of neighbouring entries for such window. After that, the trend is directly subtracted to the GSR signal, which gives the SCR component.

4.1.3.1 Time-domain

Time domain features can be divided into two main groups: higher-order statistics and morphological features.

Within the first group, the main block is the calculation of the average signal within a processing window, in which a total of N samples are acquired at a specific sampling frequency f_s. The average follows

$$\mu_X = \frac{1}{N} \sum_{n=1}^{N} X_n, \tag{4.4}$$

where X represents the BVP or GSR signals. For the BVP case, the mean value is related to the peripheral resistance, which is responsible for the vascular tone, as stated in Chapter 2. Moreover, when no Baseline Wander removal method is applied, this information is mixed together with respiration amplitude effects, which can affect the DC and very low frequency parts of the signal (equation 2.6). For the GSR case, the signal average contains information of the stationary part from the tonic level of the signal or SCL. Thus, it is strongly related to the arousal quantification. To account for variability of such information, the standard deviation of the GSR is also considered given by the square root of the variance, equation 4.5

$$s_X = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}|X_n - \mu_X|^2}. \tag{4.5}$$

Finally, the remaining higher-order statistics procedures are the 25th and 75th quartiles of the current window processing elements. These are also applied only to the GSR signal and obtained using a sorting-based algorithm.

Regarding the morphological features, they are characterised by the identification of the delineation physiological points within the current processing window. The BVP signal is subjected to the extraction of two morphological features: the average and the variability of the IBI. This metric is the temporal difference between the different systolic peaks identified. Its average and variability are related to the ANS response. Specifically, they indicate cardiac variability changes in response to acute stressors (videos). Note that this information allows to track the response of the cardiovascular system, and this variability is expected from a healthy person. Conversely, it could happen that for persons with a chronic stress condition, such as PTSD, it presents minimum or no variability. The IBI time series is given by equation 4.6

$$IBI_n = t_{sys_{n+1}} - t_{sys_n}, \tag{4.6}$$

where $t_{sys_{n+1}}$ and t_{sys_n} are the temporal positions for the $n+1$ and n systolic peaks, respectively. It is usually expressed in milliseconds and it ranges from 1000 ms to 600 ms (60-100 BPM) when being under resting conditions. The IBI variability or Heart Rate Variability (HRV) is calculated as the standard deviation of the recollected IBIs along the processing window, which is also known as SDNN. Note that

this is one out of different possible options to calculate the HRV. When dealing with the GSR signal, the number, amplitude and rise time of the different ERSCR peaks are extracted using the SCR component obtained after the trending subtraction. Thus, to extract such three features, a trough-to-peak method is run over the SCR. For this implementation, we assumed that the normally accepted amplitude criterion to discern ERSCRs over an external stimulus is 0.01 µS [112]. Note that these three features are expressed in $\mu S/sec$, i.e. they are calculated and normalised by the window processing time, in this case the video duration.

Although the time-domain based features can not deal with the non-stationary physiological information, they provide a strongly supported and validated starting point for any emotion recognition system.

4.1.3.2 Frequency-domain

Before dealing with any frequency information extraction, the frequency resolution must be set accordingly to be able to obtain all the established Power Spectral Density (PSD) bands. In fact, such resolution only depends on the temporal length of the processing window. For instance, in this case, a 60 seconds window size results into a frequency resolution of 0.016 Hz/bin given by equation 4.7

$$f_{res} = \frac{f_s}{f_s * T_{len}} = \frac{1}{T_{len}}, \tag{4.7}$$

where T_{len} is the window size in seconds, f_s is the sampling frequency of the discrete signal, and f_{res} is the frequency resolution in Hz/bin. Note that the latter is referred as the difference in frequency between each bin, i.e. the results or bins of a FFT algorithm indicate the frequency magnitude response for specific centred frequencies separated by f_{res}, Figure 4-11. In our case, the first frequency bin is centred at 0.016 Hz, the second is centred at 0.033 Hz, and so is done for the following consecutive bins. Thus, considering that the lowest PSD band to be extracted is bounded from 0 to 0.1 Hz for the BVP signal, using this temporal window length is enough to deal with all the needed PSD bands and get a proper separation between them.

The frequency-domain features for DEAP-b2 were only contemplated using the BVP signal. Four low-frequency PSD bands from the filtered signal were extracted

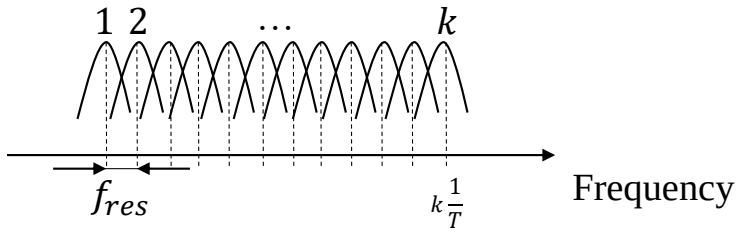


Figure 4-11: Frequency resolution illustration and frequency bins location based on a T seconds processing window.

together with the PSD ratio of the low-frequency and the high-frequency contribution from the extracted IBIs. On the one hand, the four low-frequency bands go from 0 to 0.4 Hz in 0.1 Hz step. Such information allows to recollect information regarding the low-frequency components within the BVP, i.e. mainly the respiratory effects. On the other hand, the PSD ratio of the low and high frequency bands for the extracted IBIs is based on the sympathetic and parasympathetic activation. For instance, in case the IBI variance observed were very low, the cardiac activity would be stable or constant, which from a frequency point of view implies the very low-frequency bands have more power than the high-frequency bands. Note this physiological state could be triggered by acute stressors, i.e. in our case negative emotions leading to sympathetic activation. Nevertheless, when being under resting conditions, the IBI variance will be high, which leads to high-frequency bands activation. This is depicted by Figure 4-12, which shows an ideal representation and relationship between the low-frequency (LF) and high-frequency (HF) parts of the IBI PSD, given by the Task Force of The European Society of Cardiology [6]. As it can be observed, although the HF part occupies a greater spectral range in comparison to the LF part, an evident increment is obtained in LF when not being under resting conditions and normalising both factors. Note that they can be divided into more internal bands, providing information relative to ultra-low, very-low, very-high, and ultra-high frequencies. In this case, we used a thick distinction by grouping them into two main bands: from 0 to 0.08 Hz for LF and from 0.15 to 0.5 Hz for HF.

The PSD was calculated using Welch's overlapped segment averaging estimator. It should be noted that the frequency resolution for the IBI is not the same as for the filtered BVP signal, as the IBI is an unevenly sampled or acquired signal. This implies that for a fixed temporal window, the number of gathered IBIs may

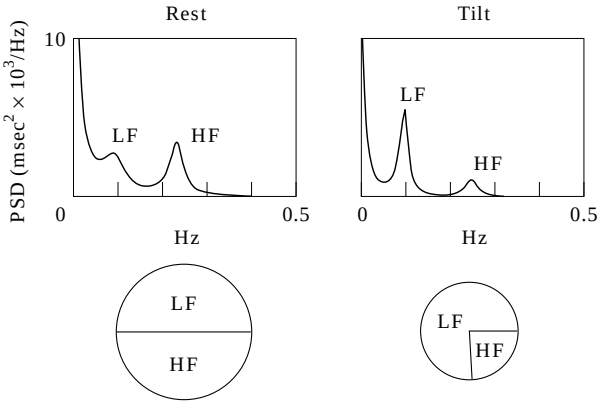


Figure 4-12: Ideal representation and relationship between the low-frequency (LF) and high-frequency (HF) parts of the IBI PSD [6].

not been the same for different processing windows. For instance, considering a fixed processing window of 60 second, a stable 40 BPM would lead to a frequency resolution of 0.025 Hz/bin, while a stable 100 BPM would reach up to 0.01 Hz/bin. Note that the worst case scenario will be the lowest cardiac frequency. To keep a fixed frequency resolution regardless of the BPMs variability, the obtained IBI points are interpolated and resampled at 8 Hz, which achieves 0.031 Hz/bin. The latter is sufficient to deal with all the needed PSD bands and get a proper separation between them.

4.1.3.3 Non-linear domain

The last set of features to be extracted are based on non-linear information. For the DEAP-b2 system, these are solely applied to the extracted IBI signal. Just as the information provided by the GSR signal is more directly related to the SNS activation due to the eccrine sweat glands, the information obtained from the BVP signal can present a wide range of behaviours that are produced by different physiological (SNS and PNS) and physical combinations or non-linearities. These can be identified as vascular or haemodynamic factors being modified by external stressors or even by homeostasis towards thermoregulation under different physical conditions. Thus, non-linear features can provide information that linear methods are losing. In fact, the superiority of the non-linear methods when applied to emotion recognition using physiological information is a hot topic nowadays [190].

It is known that the physiological non-linear behaviour can be seen in different timescales, such as circadian rhythms. Therefore, for this system we used the Multi-

Scale Entropy (MSE) introduced in [191] to consider the non-linear aspect of the IBI time series and a time scale dependency. This metric extends sample entropy to different timescales to provide an additional perspective when the time scale of relevance is unknown, which is our case. All the calculations are based on the statistical entropy given by equation 4.8

$$H(X) = -\sum_{i} p(X_i) log(p(X_i)), \tag{4.8}$$

where $p(X_i)$ is the probability mass function for the data block i. Note that this is a measure for the average uncertainty of the signal under test, i.e. a time series with non-periodic fluctuations will generate higher values than a pure sinusoidal (not chaotic) signal. Thus, the calculation of the sample entropy starts by segmenting the N point time series by an embedding dimension m, where $m < N$. This leads up to $N - m + 1$ such segments. After that, the distance d between the different m-dimensional points is calculated and compared to a predefined threshold r. In case $d < r$, the two segments are considered similar and a positive ranking of '1' is stored m, otherwise a null ranking of '0' is annotated. This is equally done for $m+1$. Finally, the results are expressed by the matrices: expressed in equations 4.9 and 4.10

$$A(m, r) = \frac{1}{N-m} \sum_{i=1}^{N-m} \Theta(r - d_i), \tag{4.9}$$

$$B(m+1, r) = \frac{1}{N-m+1} \sum_{i=1}^{N-m} \Theta(r - d_{i+1}), \tag{4.10}$$

which are used to provided the final sample entropy expressed in equation 4.11

$$SampEn(X) = -\log\left(\frac{A(m, r)}{B(m+1, r)}\right). \tag{4.11}$$

The extension that MSE introduces within the sample entropy is the coarse graining or down-sampling of the time series at different time scales. Thus, at every level (time scale), the coarse-grained time series is obtained by averaging the respective time series points. This is illustrated in Figure 4-13 and mathematically expressed by the equation:

$$y_j^{\tau} = \frac{1}{\tau} \sum_{i=(j-1)\tau}^{j\tau} x_i, 1 \leq j \leq N\tau, \tag{4.12}$$

where τ is the time scale or level. Finally, the sample entropy is calculated over the obtained y_j^τ. In our analysis, we employed a five level ($\tau = 5$) MSE, set $m = 2$, and $r = 0.2\sigma$, where σ is the standard deviation of the IBI time series. Note that these parameters were chosen based on previous works in the literature dealing with emotion recognition and this type of non-linear features [192].

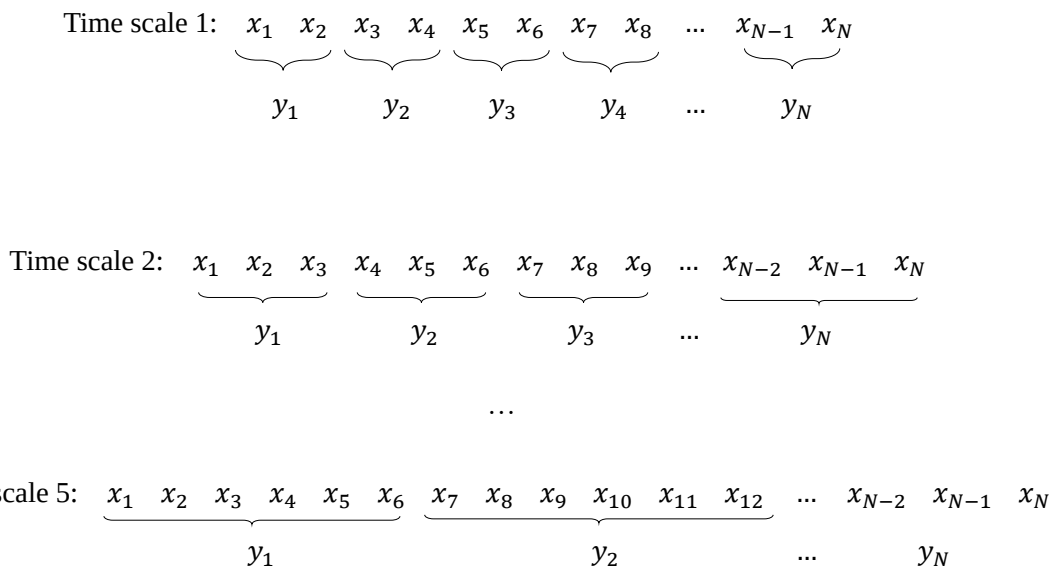


Figure 4-13: Every coarse-grained time series obtained for every level of the MSE feature extraction technique or algorithm.

4.1.4 Fear classification systems

In the following sections, the results obtained with the DEAP-b1 and DEAP-b2 systems are detailed and explained. Note that all specifications regarding labelling mapping transformation, exploratory data analysis, data processing and feature extraction have already been detailed in the previous sections of this Chapter.

4.1.4.1 DEAP-b1 system

This system [184] was, up to my knowledge, the first in the literature to propose and validate the specific fear binary mapping using the PAD space and only three physiological variables. As already stated in previous sections, only 21 out of 32 volunteers from DEAP were employed, and no feature extraction as such was applied, instead, the filtered value of each physiological variable was considered. Every volunteer was subjected to $0-1$ scaling for the complete set of physiological values gathered during the experiment. For the classification, towards a first embedded implementation proof of concept of the system, we consider a lazy algorithm, specifically, a KNN. Note that the value of k was fixed to the square root of the size

of the training set, which is a commonly applied practice. For simplicity in the computation, the Euclidean distance is considered to compare two samples. To be fair in this comparison and avoid problems related to values in different units and scales, every of the three values in a sample is normalised as stated previously ($0-1$ scaling). Finally, for the validation, we implemented a Hold-Out CV strategy and performed an experimental parameter sweep for the Hold-Out ratio. Moreover, to deal with the strong imbalance labelling situation, we decided to apply cost-sensitive learning by tuning a miss-classification cost parameter. This practice is commonly used in imbalance binary classification problems. In this case, such parameter defines a penalty that gives more importance (weight) to the false negatives produced. Thus, the usage of this penalty is useful to reduce the false negative rate in our system, which is critical for the Gender-based Violence application. For instance, when considering Bindi, the bracelet is actually at the bottom of a cascade of more powerful devices, and so this constrained wearable device could act as a trigger for running more complex algorithms in upper layers if needed. Hence, it is essential to reduce the number of false negatives in this first step, although it penalises accuracy.

As a first step in the evaluation of DEAP-b1, we focus on the data from one arbitrary volunteer, number 18 ($p18$). For this specific volunteer, there are a total of 256,358 samples or instances of each physiological variable collected during the entire experiment (40 clips). The KNN algorithm is trained using different values of Hold-Out and miss-classification costs. For each combination of Hold-Out and miss-classification, 30 independent intelligence systems are randomly generated to have statistical validity in the results obtained. Different metrics such as accuracy, specificity (or true negative rate), sensitivity (or true positive rate), and geometric mean between sensitivity and specificity are analysed and compared. It is noteworthy to mention that the memory usage is strongly affected by the Hold-Out ratio, as the complete training space needs to be stored to process further samples and provide future inferences. Thus, the design space for the Hold-Out ratio goes from 0.7 to 0.9999, which leads up to a training set size from 30% to 0.01%.

Figure 4-14 shows accuracy vs. miss-classification cost for the different values of Hold-Out in $p18$. Analysing this figure, we check that i) accuracy is better for lower values of Hold-Out (the training set is bigger and then, the system is

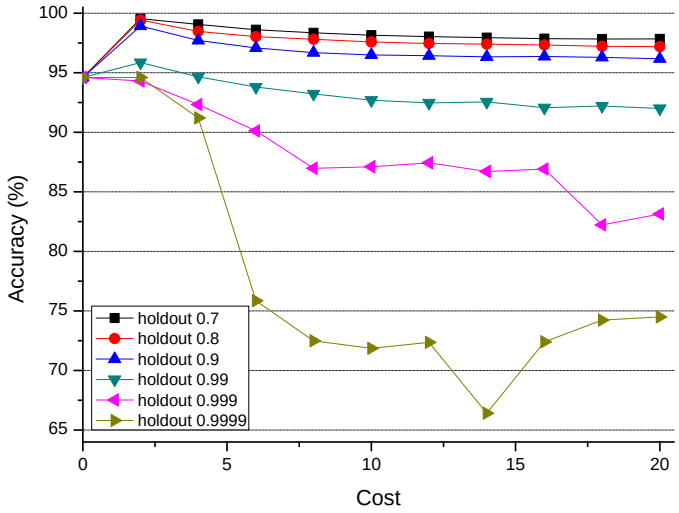


Figure 4-14: Accuracy vs. miss-classification cost for $p18$.

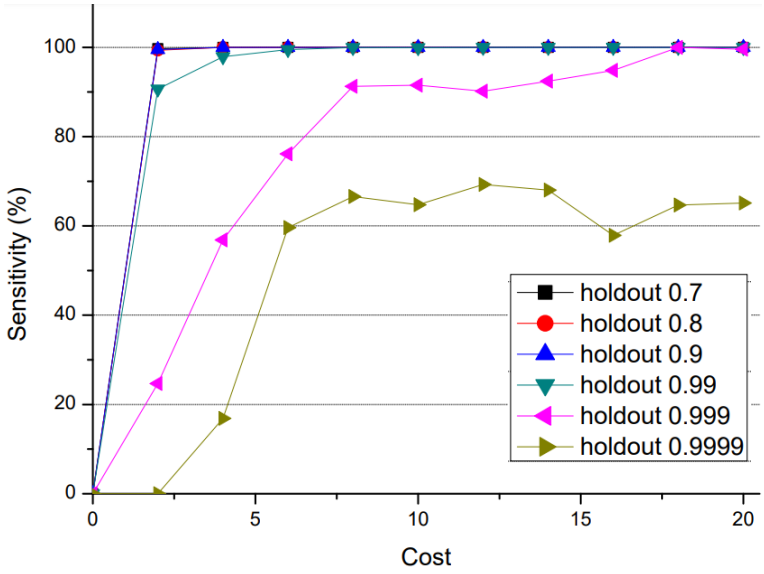


Figure 4-15: Sensitivity vs. miss-classification cost for $p18$.

better characterised) and ii) accuracy usually decreases as miss-classification penalty increases (the number of false negatives is reduced, but also increases the number of false positives). Figure 4-15 shows sensitivity vs. miss-classification cost for the different values of Hold-Out in $p18$. Analysing this figure, we check that sensitivity increases with the miss-classification cost based on the mechanism of this penalty. Figure 4-16 shows specificity vs. miss-classification cost for the different values of Hold-Out in $p18$. In this figure, specificity decreases with the miss-classification cost based on this penalty. Note that the legend of the two last figures is the same as the first one.

Applying this analysis to the rest of the volunteers and observing similar behaviours, we determined that a miss-classification cost of 8 units can be adequate for

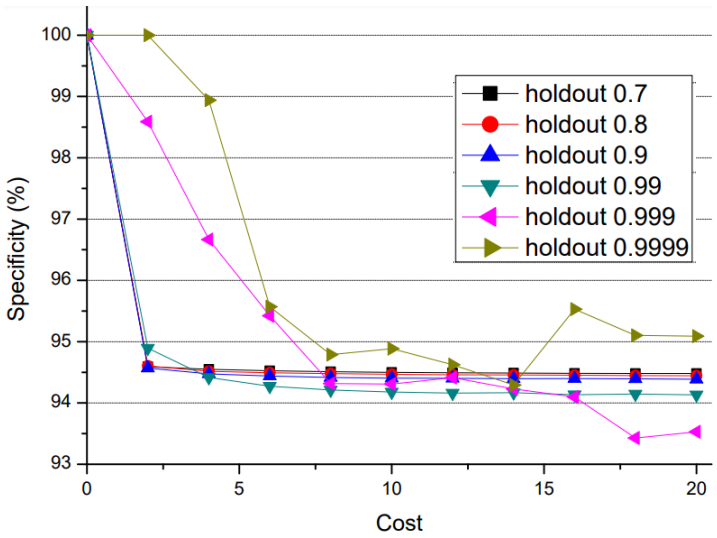


Figure 4-16: Specificity vs. miss-classification cost for $p18$.

Hold-Out	training set size	memory used (kB)	operations
0.7	76907	976.36	375767
0.8	51272	650.91	241483
0.9	25636	325.45	113024
0.99	2564	32.54	8739
0.999	256	3.25	618
0.9999	26	0.32	36

Table 4.5: Impact of the size of the training set on memory and computation for $p18$. Subject-dependent approach.

the current dataset. However, for the Hold-Out ratio, this decision is not immediate or purely based on performance, as it is needed to study the impact in memory and computation of this parameter. Thus, Table 4.5 shows, for each value of Hold-Out for volunteer p_{18}, the size of the training set, the required memory needed considering such training set allocation and the memory consumed by the KNN algorithm, as well as an estimation for the number of operations. Note that the memory used in KB is based on a 32-bit integer data type, and the number of operations are based on the average computational complexity of the quick-sort method usually found in KNN implementations. This complexity is $\mathcal{O}(n \log n)$, where n is the number of elements to sort, i.e., the size of the training dataset. Analysing this table, we check that the Hold-Out value has an important impact on the memory used and the number of operations to compute. In fact, going from 0.999 to 0.99 leads up to more than 13x times more operations. These aspects are critical for an edge-computing system as Bindi. As a consequence, after analysing the trends in Figures 4-14, 4-15 and 4-16, the Hold-Out value is fixed to 0.99.

Accuracy	Sensitivity	Specificity	Geometric Mean	Volunteer
0.71	0.97	0.67	0.81	p_1
0.93	1.00	0.85	0.92	p_2
0.80	0.97	0.77	0.86	p_3
0.79	0.97	0.85	0.90	p_4
0.96	0.97	0.97	0.97	p_5
0.91	0.99	0.94	0.96	p_6
0.99	1.00	0.97	0.98	p_7
0.86	0.97	0.89	0.92	p_8
0.86	1.00	0.80	0.89	p_9
0.83	0.98	0.88	0.92	p_{10}
0.94	1.00	0.87	0.93	p_{11}
0.69	1.00	0.52	0.72	p_{12}
0.84	1.00	0.73	0.85	p_{13}
0.76	0.99	0.70	0.83	p_{14}
0.79	1.00	0.66	0.81	p_{15}
0.71	0.99	0.74	0.85	p_{16}
0.87	1.00	0.86	0.92	p_{17}
0.93	1.00	0.94	0.96	p_{18}
0.82	0.99	0.79	0.88	p_{19}
0.84	1.00	0.81	0.90	p_{20}
0.77	0.99	0.74	0.85	p_{21}
0.84 (0.08)	0.99 (0.01)	0.81 (0.11)	0.88 (0.06)	$\mu(\sigma)$

Table 4.6: Accuracy, sensitivity, specificity and geometric mean metrics for each volunteer by assuming hold-out and miss-classification cost of 0.99 and 8, respectively. Subject-dependent approach.

Based on the previously defined values for Hold-Out and miss-classification cost, Table 4.6 shows accuracy, sensitivity, specificity, and geometric mean metrics for all the volunteers considered in the dataset. Note that all the volunteers have comparable size to $p18$, and so we fixed the same Hold-Out value. In this case, also 30 independent systems have been trained for each model to have statistical validity. Analysing this table, we check that, on average, accuracy, sensitivity, specificity, and geometric mean were 0.85, 0.99, 0.81, and 0.81 respectively. Note that the higher average and lower variance for the sensitivity in comparison to the other metrics is due to the miss-classification cost, i.e. the system is more biased towards the positive class.

In case of considering a subject-independent approach and applying the same miss-classification cost, the Hold-Out might be re-evaluated in terms of memory consumption. Thus, Table 4.7 shows, for each value of Hold-Out, the size of the training set

when mixing data from all the considered volunteers in the dataset, the memory used in kB is based on a 32-bit integer data type, and the number of operations follow the same calculation approach as for the subject-dependent case. Analysing this table, we check that, for this system, a value of Hold-Out equal to 0.999 is comparable to a value of 0.99 for the subject-dependent case in Table 4.5, i.e., we need to reduce the Hold-Out ratio to achieve similiar number of points in the training dataset. Table 4.8 shows accuracy, sensitivity, specificity and geometric mean metrics for the subject-independent case by assuming different Hold-Out values while considering the miss-classification cost of 8 from before. Specifically the obtained metrics for the 0.999 Hold-Out are significantly lower than for the subject-dependent case. Moreover, we check that the three metrics are improved when considering lower values of holdout, as for the subject-dependent case. However, their high impact in space and time complexity makes their implementation into constrained edge-devices not feasible.

Thus, based on these experiments, we can conclude that a subject-dependent implementation can significantly improve the performance of the emotional state inference in a tiny constrained wearable device. Specifically, the subject-dependent approach provides up to 0.84, 0.99, 0.81, and 0.88 of accuracy, sensitivity, specificity, and geometric mean on average while the subject-independent approach provides up to 0.54, 0.88, 0.47, and 0.62, for the chosen configurations. For this latter approach, in the case of Bindi, which intends to provide a fear machine learning engine to be deployed in real-life, the sensitivity should be close to 1.00 to maximise the true positive detection, as occurs with the subject-dependent approach.

Notwithstanding the evidence of the results, the Hold-Out strategy used for both models, subject-dependent and subject-independent, can lead towards overoptimistic results. This is due to the fact that the inputs of the system are the filtered physiological values and the applied Hold-Out did not take into account if they belong to the same video. Thus, the training and the testing processes could be using information from the same physiological data set collected during a given video clip visualisation.

Hold-Out	training set size	memory used (kB)	operations
0.98	76555	971.89	373892
0.99	38277	485.94	175423
0.999	3828	48.59	13715
0.9999	383	4.86	989

Table 4.7: Impact of the size of the training set on memory and computation. Subject-independent approach.

Accuracy	Sensitivity	Specificity	Geometric Mean	Hold-Out
0.66	0.96	0.52	0.71	0.980
0.64	0.95	0.50	0.69	0.990
0.54	0.88	0.47	0.62	0.999
0.50	0.81	0.45	0.60	0.9999

Table 4.8: Accuracy, sensitivity, specificity, and geometric mean metrics for each tested hold-out assuming a miss-classification cost of 8. Subject-independent approach.

4.1.4.2 DEAP-b2 system

During the research and development of the DEAP-b1 system, we identified five different drawbacks:

1. The low number of volunteers could be affecting to the data variability.
2. The complexity and imbalance of the fear binary mapping from PAD was particularly high for this database.
3. The fact that we did not consider feature extraction could be leading to loosing physiological information of interest.
4. The application of a Hold-Out strategy over the filtered physiological values could be resulting into overoptimistic metrics.
5. The space complexity for such lazy KNN algorithm was considerably high when considering lower Hold-Out values.

The first drawback can be fixed by considering the complete set of volunteers at the expense to skip the SKT signal from all of them. Note that SKT inaccuracies were found during the exploratory data analysis for a total of 11 volunteers. The second shortcoming can be alleviated by applying the fear binary mapping from PA, which showed a lower imbalance ratio. The third reason is one of the most sensitive from

a physiological point of view, as by considering just the raw filtered values, we are loosing all temporal, morphological, frequency-based and non-linear information. The fourth drawback needs to be solved to properly assess the system performance and assure that no testing or even testing-related information is provided to the training stage. Finally, the fifth drawback affects to the training set size and motivates the evaluation of different classification algorithms that provide less space requirements. Based on these identified problems, the DEAP-b2 system [182] tried to overcome them by considering the 32 volunteers from DEAP, only data from PPG and GSR sensors, a fear binary mapping using the PA space, and a complete set of 20 features including temporal, frequency and non-linear domain, (enumerated in Table 4.4 and detailed in Section 4.1.3). Moreover, the feature extraction process was applied considering a 60 seconds window processing, which corresponded to the stimulus duration and provided a set of 20 features per video. Thus, in this way we assure that no information within the same video is given to both training and testing when performing the Hold-Out strategy.

In addition to simplifying the labelling problem and reducing the imbalance ratio by choosing a fear binary mapping based on the PA space, an oversampling technique was applied over the minority class data (fear). Specifically, SMOTE was implemented to deal with the observed balance problems [193]. This technique is based on an over-sampling approach of the minority class generating new samples by considering the closest k neighbours, rather than by over-sampling with replacement. Thus, instead of having 1280 instances (32 volunteers x 40 videos) with a class balance ratio of about 76/24% (negative/positive), we achieve a class balance ratio of up to 50/50% with a total of 1800 instances. Note that the k value for the SMOTE was set to 5.

Regarding the specific classifiers used in this system, Gaussian naïve Bayes (equation 4.2) and SVM with RBF kernel were used. This decision was based taking into consideration two main facts:

- We decided to use the same classifier used by DEAP (Gaussian naïve Bayes) to provide a fair comparison with respect to the original dataset.
- To overcome the space complexity of KNN, a SVM classifier with RBF kernel is applied, as it preserves all the advantages of the KNN algorithm, storing

only the support vectors during training rather than the entire training space. In addition to the $0-1$ scaling performed in the previous system, z-score was applied per volunteer to normalise the data in this case. For the testing methodology of DEAP-b2 system, the Hold-Out strategy was run from 0.01 (1%) to 0.9 (90%) for 100 iterations every 0.01 step. Furthermore, for the validation of the SVM classifier during training, a $k-fold$ is implemented with $k=5$.

The system topology, in this case, is based on a subject-independent approach, as a subject-dependent was not feasible due to the small amount of data (40 sets of 20 features per volunteer). Moreover, besides the classification metrics used for DEAP-b1, the Area Under the Curve (AUC) is also given in this case, which provides a measurement of performance across all possible classification thresholds and presents the probability of the model ranking a random positive more highly than a random negative.

All these DEAP-b2 considerations are structured and combined into six different configurations to provide a bounded DSE for the subject-independent use case. These are given as follows:

- Case 1. The system is implemented without using any feature selection and applying the Gaussian naïve Bayes classifier.
- Case 2. The system uses the same filter feature selection method as DEAP (Fisher linear discriminant score, equation 4.1), but implements a SVM classifier with RBF kernel. The latter is taken from [148] with $\gamma = 0.15$ and $C = 1$.
- Case 3. This system configuration follows the same structure as Case 2, but without employing feature selection.
- Case 4. It implements SMOTE to deal with the balance problem, uses Fisher linear discriminant score to select the relevant features and runs the Gaussian naïve Bayes classifier.
- Case 5. It presents the same configuration as Case 4, but using the SVM classifier of Case 2.
- Case 6. It employs SMOTE, Fisher linear discriminant score to select the relevant features and a SVM classifier with RBF kernel. Moreover, a grid search is applied to find the optimal hyperparameters for such classifier.

Case	Accuracy ($\mu(\sigma)$)	Sensitivity ($\mu(\sigma)$)	Specificity ($\mu(\sigma)$)	Geometric Mean ($\mu(\sigma)$)	AUC ($\mu(\sigma)$)
1	52.48 (0.34)	50.84 (2.14)	53.36 (0.90)	52.08 (1.38)	52.55 (1.13)
2	76.47 (0.34)	0.12 (0.12)	99.87 (0.11)	3.46 (0.11)	50.00 (0.10)
3	76.54 (0.37)	0.03 (0.07)	99.96 (0.07)	1.73 (0.07)	50.00 (0.10)
4	51.80 (0.54)	52.73 (0.94)	50.86 (0.90)	51.78 (0.91)	52.24 (0.80)
5	53.27 (0.73)	58.80 (2.83)	47.72 (1.77)	52.97 (2.23)	53.50 (2.20)
6	62.80 (4.75)	62.27 (4.14)	66.99 (5.79)	62.62 (4.73)	62.79 (4.72)

Table 4.9: Accuracy, sensitivity, specificity, and AUC metrics for each case by assuming the specified conditions, respectively. Subject-independent approach.

Table 4.9 provides a comparative analysis of these six different implementations, in which random -like results can be appreciated by the obtained AUCs for all the cases except case number six. Note that some of the values of this table have been modified in comparison with the ones obtained in [182], as more tuning of the models was performed after such publication. First of all, the Gaussian naïve Bayes classifier in Case 1 achieves a poor performance. This can be affected by the independence of the extracted features, as it is known that this type of classifier provides a good performance when the features are independent of each other. Then, it is striking that a high accuracy score does not mean the model is appropriately performing. For instance, Case 2 and 3 have the highest classification rate, but there is no sensitivity, however specificity is close to 100%. Thus, accuracy paradox is happening, so the accuracy is only reflecting the underlying class distribution. Also for these two cases, the specific feature selection technique by itself, i.e. Fisher linear discriminant score, is not providing any advantage. This can be due to the non-suboptimal feature set that this technique generates, as previously stated in Section 4.1.4. Cases 4 and 5 in comparison with 6 demonstrates that finding the right hyperparameters is essential in order to achieve the best Bias-Variance trade-off. Thus, by applying a grid search, the performance of the classification scheme improves. Finally, Case 6 combines a synthetic oversamplig method, ranked feature selection, a non-linear based classifier and a grid search hyperparameter tuning process, achieving up to 62.79% AUC, which overpass the rest of the Cases.

Although the average geometric mean of the latter Case is lower than the one obtained for the subject-dependent model of DEAP-b1, it must be highlighted again the validation drawback observed when applying the Hold-Out strategy directly over the filtered physiological values. Moreover, when comparing the DEAP-b2 with the

DEAP-b1 results for the subject-independent, we can observe that both systems achieve similar geometric mean, but the DEAP-b2 outperforms the DEAP-b1 in specificity by more than 15% and in accuracy by more than 8%. This indicates that DEAP-b2 presents a better balance between false positives and false negatives, leading to a better performing system. Finally, from a time and space complexity balance, the SVM clearly outperforms the previous KNN implemented. The latter presents $\mathcal{O}(n\ \log\ n)$ and $\mathcal{O}(n)$ for the time and space complexity respectively, where n is the training set size. Conversely, the SVM with the RBF kernel achieves $\mathcal{O}(n_{sv}d)$ and $\mathcal{O}(n_{sv})$ for the time and space complexity respectively, where n_{sv} is the size or number of the support vectors and d is the number of attributes or features to be employed. For the worst case scenario, considering the 20 features and a thousand of support vectors chosen over the complete set of 1,800 instances, there will be a total of 20,000 operations per prediction, which is considerably higher than both subject-dependent and subject-independent from DEAP-b1. However, due to the feature selection that DEAP-b2 is doing, most of the iterations ended up with half of the features, which is translated into 10,000 operations. This measurement is in between both DEAP-b1 subject-dependent and subject-independent models using KNN. Concerning the space complexity, the SVM in this case requires up to 39.06 kB (1000 sv per each feature), which also lies between the two DEAP-b1 models. Note that the memory used in kB is based on a 32-bit integer data type for all the support vectors to be stored.

It is noteworthy that further SVM algorithmic optimisations, as well as other feature selection alternatives, can be implemented and applied to achieve a smaller amount of support vectors, which would reduce both time and space requirements for this system leading even in some cases to better recognition performance. In fact, Table 4.10 shows the results for the same training configuration, validation and testing as Case number six but changing the feature selection method to mrMR. This technique [194] is based on the assumption that within the given entire feature set there is a minimal-optimal set in which such features are mutually as dissimilar to each other as possible, but also marginally as similar to the classification variable as possible. With this technique, we want to select the features that have maximum relevance for the classification variable (target) and present a minimum redundancy in

comparison to the rest of the other features being evaluated. Thus, to measure such properties between two variables (X and Y), the mutual information is employed, which is given by equation 4.13

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p_{(X,Y)}(x,y) log \frac{p_{(X,Y)}(x,y)}{p_X(x) p_Y(y)}, \tag{4.13}$$

where $p(x,y)$ is the joint probabilistic distribution, and $p(x)$ and $p(y)$ are the marginal probability density functions for each variable respectively. From this information, the level of similarity (or dissimilarity) between two features (i and j) is encoded by the minimum condition; stated in equation 4.14

$$min\ W_I, W_I = \frac{1}{|S|^2} \sum_{i,j \in S} I(i,j), \tag{4.14}$$

where S is the subset of minimal-optimal features. In the same way, the discriminant power of features or relevance for the classification variable h (target) is provided by the maximum condition stated in equation 4.15

$$max\ V_I, V_I = \frac{1}{|S|} \sum_{i \in S} I(h,i). \tag{4.15}$$

From equations 4.14 and 4.15, the minimal-optimal feature set can be obtained by optimising them in a simultaneous way and by applying different criterion functions. Specifically, the particular mrMR implementation for this system applied the Mutual Information Difference (MDI) criterion to perform the ranking process. Equation 4.16

$$max(V_I - W_I). \tag{4.16}$$

As this feature selection technique is a filter feature selection method which ranks the existent features based on such premises, the number of final features, K, to be considered after such ranking should be indicated. In this case, the selection is done based on a performed experimental parameter sweep taking into account performance and storage requirements. Finally, K is fixed to 10. Therefore, the obtained results using this specific feature selection technique outperforms by more than 18% the AUC of the subject-independent model of DEAP-b2, while maintaining the same storage requirements. This last experiment demonstrates the possibilities

of optimisation are high in this complex problem.

Case	Accuracy ($\mu(\sigma)$)	Sensitivity ($\mu(\sigma)$)	Specificity ($\mu(\sigma)$)	Geometric Mean ($\mu(\sigma)$)	AUC ($\mu(\sigma)$)
6+	81.54 (8.69)	70.93 (14.92)	94.59 (3.89)	81.55 (10.21)	81.60 (8.70)

Table 4.10: Accuracy, sensitivity, specificity, and AUC metrics for mrMR feature selection and SVM with RBF kernel. Subject-independent approach.

4.2 Fear classification using MAHNOB

After having presented the results for three different fear detection systems (one subject-dependent and two subject-independent) using the DEAP dataset and being aware of the limitations found, the need for a new dataset, in which these problems are solved or alleviated, was strongly required. Some of these identified problems were referred to the physiological recovery between stimuli, the skin temperature data inaccuracies, and the class imbalance of fear mapping and binarization.

As already reviewed in Section 3.2, the MAHNOB database overcomes the physiological recovery limitations of DEAP, keeps the same recollected physiological information, and presents even more self-reported labels from volunteers. Moreover, no measurement problems are observed for the SKT, or any other physiological variable, with any of the valid volunteers. These claims are even reinforced by the literature; for instance, the authors in [195] conducted a DSE for the feature vectors of DEAP and MAHNOB to investigate the relevance of the physiological features within both datasets. One of their experiments concluded that the stimuli in MAHNOB were more emotionally immersive than the ones in DEAP. In fact, such work has motivated the realisation of further and recent research such as [196]. On this basis, in this section, the results obtained for a subject-dependent and subject independent fear binary recognition systems based on the MAHNOB dataset are detailed. Specifically, and trying to design a more specialised system towards the long term goal of this research work, two design key aspects are fixed:

- Only women volunteers are employed. This design constraint allows for the development of very specialised emotion recognition systems due to the emotional particularities between men and women, as reviewed in Chapter 2.
- Fear binarization only from the PAD space is contemplated. The dominance

factor, as reviewed in Chapter 2, is essential to distinguish between some of the main negative emotions (fear and anger).

Moreover, the data segmentation and other processes such as the feature extraction are modified or extended to improve the systems presented in the previous Section.

For the specific methodology followed during the MAHNOB database experiments, Figure 4-17 shows a simplified diagram of the experimentation applied for every volunteer and each stimulus. Unlike DEAP, MAHNOB did take into account the reduction of the emotional bias after every stimulus visualisation and, therefore, emotion responses. In fact, the neutral clips used were randomly selected from a larger pool provided by the Stanford psychophysiology laboratory [197]. This consideration together with the 30-second pre and after trial recording provided a physiological recovery intended to isolate the emotional activation between stimuli.

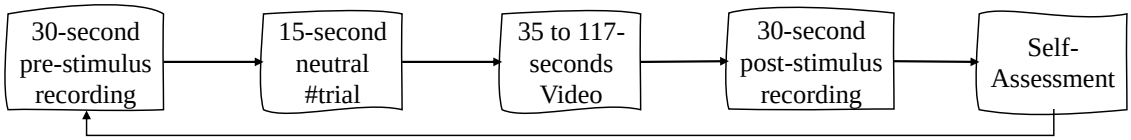


Figure 4-17: Methodology followed during the MAHNOB database experiments.

Regarding the specific technical differences between the systems presented in this Chapter and the one that could be actually integrated into Bindi, two of them must be highlighted. On the one hand, the MAHNOB dataset includes the cardiac activity information measured with an ECG sensor. Thus, the different delineation or peak detection algorithms, which are used to extract the morphological information to calculate the different features or metrics, need to be designed specifically for the ECG morphology rather than the PPG reviewed in Section 2.5.1. This fact directly affects any possible option of pre-processing integration to leveraging that algorithmic part into the embedded platform of the smart bracelet of Bindi. However, it is proven in the literature that PPG is a valid surrogate of ECG for different metrics or features such as HRV [198–200]. Therefore, the feature extraction and further processes can be applied in the same way regardless of whether the sensor is ECG or PPG. On the other hand, the equipment used during the MAHNOB experiment was the same as in DEAP. Moreover, even in case of having PPG sensor data and being able to get morphological features when using such equipment, due to the extremely challenging task of obtaining clinical quality PPG signals (morphology is totally preserved) with wearable devices, that would be worthless because of the

very high noise in real life. Thus, as in the previous section, the proposed systems here serve as a proof of concept and eases the different DSE processes that need to be performed towards an optimal fear binary emotion recognition system design and integration on the edge.

Preceding the presentation of the methods employed and results obtained, a review of the state-of-the-art, regarding the utilisation of MAHNOB for the generation of emotion recognition systems is detailed. First of all, as already described in Section 3.2, the original work of the MAHNOB dataset comprises the acquisition of different physiological signals at a sampling frequency rate of 256 Hz during the visualisation of different audiovisual stimuli (20 emotional clips interspersed with 20 neutral clips). First of all, they used basic pre-processing procedures to remove the temporal low frequency drifts of some signals and smooth them by using moving average filters. They extracted a total of 102 features from all the collected signals and applied a filter feature selection method to use only the highest-ranked ones. Specifically, they used one-way Analysis of Variance (ANOVA) and rejected any non-significant feature ($p > 0.05$). For the classification task, they provided two emotion recognition systems based upon low, medium and high levels of arousal and valence detection to be used as a benchmark for further investigations using such data. The levels resulted from the mapping between emotional keywords and classes following [3]. Regarding the classifier, they employed a SVM with RBF kernel and adjusted γ using a 20-fold CV. Lastly, the testing strategy applied was LOSO. By employing all the peripheral signals, they provided average ACC and F1-score metrics and obtained 46.20% and 38.00% for arousal and 45.50% and 39.00% for valence, respectively. It should be noted that they also performed multimodal data fusion by using EEG and eye gaze recollected data and achieved better results, 67.70% and 62.00% for arousal and 76.10% and 74.00% for valence.

Since its release, different machine learning systems have been proposed in the literature using its data. For instance, the work in [196] is highlighted due to its similarity with our research. They used multidimensional dynamic time warping as a non-linear technique to deal with physiological dynamics followed by a stacking classifier. Their results achieved up to 94.00% and 93.60% accuracy for a three emotional class subject-independent model by using all physiological signals from

MAHNOB database and a $k - fold$ CV strategy. Although they tried to diminish the possible bias effect by combining both labelling methodologies, mapping arousal and valence dimensional space into a specific discrete emotion, their model was not able to capture the difference between fear and anger. This fact is essential for our use case.

Among the rest of the state-of-the-art based on the MAHNOB database and regarding specifically the fear recognition use case, the only system proposed in the literature is our publication [187]. This is the one being detailed in the following subsections.

4.2.1 Stimuli balance and labels considerations

As for the DEAP database, MAHNOB stimuli were based on a previous larger stimuli pool. Specifically, the preliminary study contains 155 video clips from different movies [201]. Each video clip received 10 annotations on average using a 9-point Likert scale for arousal and valence dimensions by means of the SAM and discrete emotional tags. Based on the accumulative agreement of the latter, the researchers selected up to 14 stimuli from this previous study. For instance, the clip with highest number of fear tags was selected to elicit fear. The remaining six videos until reaching the 20 videos of the experiment were chosen based on popular online audiovisual content. Thus, most of the stimuli selected for this database were chosen following a discrete-like emotional criterion. Note that, as noted before, the 20 neutral videos used for physiological recovery where validated by the Stanford psycho-physiology laboratory.

Within this labelling context, the researchers of MAHNOB did not consider the emotional dimension aspects (arousal, valence and dominance) and so the set of generated ground truth labels were based on discrete emotions. However, they provided and used a discrete-dimensional mapping for arousal and valence based on [3] as shown in Table 4.11. Unfortunately, as the preliminary study in [201] is not publicly accessible, we cannot realise the same exploratory labelling analysis (ground truth reports vs self-reports of volunteers during the experiment) as with DEAP. Thus, the self-reports of the volunteers can be compared to the ground truth just by means of fear binarization of the latter based on the emotional discrete tags. After performing the fear label binarization in MAHNOB using the provided self-

Table 4.11: Discrete-dimensional mapping for arousal and valence based on [3] and adopted by MAHNOB [10].

Arousal classes	Emotional keywords
Calm	sadness, disgust, neutral
Medium arousal	joy and happiness, amusement
Excited/Activated	surprise, fear, anger, anxiety

Valence classes	Emotional keywords
Unpleasant	fear, anger, disgust, sadness, anxiety
Neutral valence	surprise, neutral
Pleasant	joy and happiness, amusement

reports for arousal, valence and dominance and following the fear mapping proposed in Section 2.3.4, the obtained distribution was analysed for all the considered female volunteers, resulting in asymmetry. That meant that the appearance of fear labels was not uniform for all of the participants. Thus, Figure 4-18 shows that 60% of the volunteers reported more than 30% of binary-fear labels, whereas the rest of the volunteers were below that amount. Note that, in this figure, the notation Vx means volunteer x, with $x \in 1 \dots 12$, and the notation G refers to the original binary-fear distribution of the experiment (the actual number of stimuli intended to elicit fear; i.e., only 20% of the total amount of videos). This unbalanced situation is especially relevant for V11, with only 5% of fear data. This analysis supports assumptions already highlighted in previous chapters, such as that the interpretation of stimuli is strongly volunteer-dependent.

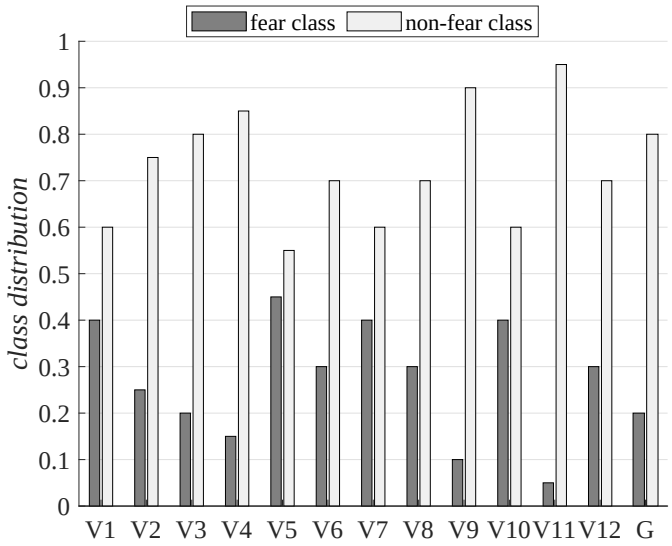


Figure 4-18: Class distribution for binary fear mapping over the subjective self-reports in MANHOB for all the different considered female volunteers, and the original intended class distribution of the experiment.

Nevertheless, in the case of assessing the average balance or average class percentages of the 12 female volunteers, the imbalance ratio is 1:2.6 (NoFear:Fear Class) and the consequent class percentages are up to 72.50% and 27.50% for the negative and positive classes respectively. On this basis, two conclusions can be obtained. On the one hand, the average positive class in this case is even higher than the expected to be achieved following the ground truth. Although the difference is less than 10%, this fact needs to be contextualised with respect to the male volunteers. For instance, the average class percentages for the nine valid male volunteers is 81.11% and 18.89% for the negative and positive classes respectively. Without considering the realisation of any statistical test to assess if the difference between men and women is significant, from a emotional point of view and considering the women emotional processing differences stated in Section 2.3.3, this could be one of the factors being influencing. On the other hand, the obtained balance for this database using the fear binary mapping from the PAD space is smaller than the imbalance ratio observed when performing the fear binary mapping from the PA space with DEAP. This conclusion can not be directly interpreted as that MAHNOB is better than DEAP, but it provides insights of the differences regarding the stimuli perception or efficacy from both databases, which is in line to previous research works [195].

Following the same schema analysis for this database as the one applied to DEAP, the label inter-individual correlations are assessed. In this case, the results obtained after a Levene's test and a Kruskal-Wallis test rejected the null hypothesis that the variances are equal across all volunteers ($p<0.001$). Note that the set of binarized labels exhibit a non-normal distribution and that the significance level was set at $p<0.05$. After these processes, Spearman correlation and the Chi-square test of independence are applied, Figures 4-19a and 4-19b. The obtained results are close to each other and fail to reject the null hypothesis on average for each of the 12 volunteers, which indicates that the average correlation is considered not significant and the different variables are independent. Thus, it can be conclude that there is not enough evidence to suggest that an association between the fear binary labels of the volunteers exist. Moreover, when comparing these graphs with the ones obtained for the correlation and independence study of the previous systems using DEAP and

the fear binary as well from PAD, we can observe a stronger agreement in this case.

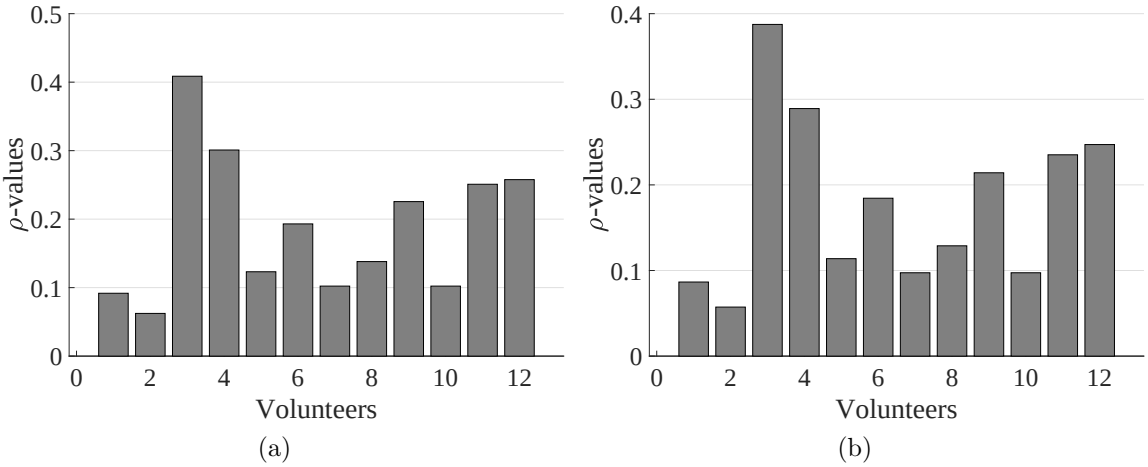


Figure 4-19: Averaged $p-values$ for all considered MAHNOB volunteers and their labels applying: a) the Spearman correlation, and b) for the Chi-square test of independence. In this case, the labels are binarized using the PAD fear binary based mapping.

Based on the analysis provided in this section, it is demonstrated that the fear binary mapping from PAD with MAHNOB is equivalent to the one done from PA using DEAP, which benefits the objectives of this research work as the same or similar techniques used for the emotion recognition paradigm of the latter can be applied for the current. It should be noted that the different results gathered from the stimuli balance and label consideration study provided in this section were always present during the design of such systems.

4.2.2 Exploratory Data Analysis, Data Segmentation and Filtering

The exploratory data analysis performed with MAHNOB followed the same procedure as with DEAP. Different plots synchronised with the experimental methodology were generated to check the physiological recoveries or neutral pre-stimulus clip as well as the normal physiological ranges for all the considered volunteers. After this analysis, it was concluded that, on average, the 30 seconds of data at the beginning and at the end of the video clip slot together with the neutral clips were actually behaving as expected leading towards the stabilisation of physiological signals and targeting the emotional isolation between stimulus. Thus, the 60 second periods corresponding to the 30 seconds before and after the stimulus in this specific experimentation were eliminated.

As stated in Section 3.1.2, data segmentation or window-based methods are used to extract emotion-related information concerning time instants. Unlike the previously presented system in this Chapter, this system operates on a data segmentation basis following the typical data segmentation procedures in the literature [85]. Regarding the DSE faced in this stage, an appropriate window length must be chosen to ensure: that (1) the frequency resolution is sufficient to deal with all the frequency-based features, and (2) the length of each window is the minimum as possible to ease the host processing tasks. For our specific use case and based on the features to be extracted, which are later described and detailed, the minimum required frequency distinction between bands is 0.05 Hz, which can be assured by using a 20 seconds window size. With this window duration, those two conditions are satisfied. Moreover, a 50% overlap is employed. To select the optimal window length and overlapping, different considerations must be assessed:

- Both the time (the bigger the window, the longer the processing) and computational complexity (the bigger the overlapping, the more operations are needed within the same time).
- Physiological facts. They are related to the non-stationary nature of these signals, which can be blurred for very large windows.
- Machine learning training size. This is referred to the final number of samples or instances provided after windowing, as the feature vector is extracted from each window and so the number of training and test points varies based on the number of windows obtained from the data.

In our case, some physiological limitations are assumed when dealing with 20 seconds windows. For instance, an ERSCR duration greater than 20 seconds cannot be captured in one single window. Note that, as stated in Section 2.5, the ERSCRs may vary between 1 to 30 seconds, although the initial configuration of a 50% overlap allows for a balanced trade-off between the amount of ERSCR information lost and memory requirements. Based on our window duration and overlapping, the average segmentation per video resulted into five windows or instances, which had the same class or label.

Concerning the storage of the acquired signals into an embedded platform, for instance assuming a maximum width of 32-bits for each data point, the parameters

set would lead to a 60 KB memory requirement (256 samples per second × 20 s × 3 sensors for 32-bit samples). This storage space could be provided by the current system-on-chips that are used for many wearable devices. Nevertheless, these requirements are application-driven and can be modified and ajusted based on the embedded platform capabilities.

Regardless of the window length, data are encapsulated in fixed time slots to be processed when filled, Figure 4-20. These segmented data (windows) obtained are pre-processed to eliminate noise and other non-useful components for the next steps. Thus, the overall signal quality is improved by denoising filters, focusing on their specific physiological characteristics. Specifically, the raw ECG signal is subjected to a band-pass FIR filter through a low and high pass filtering cascade to ease complexity. Moreover, the residual baseline wander is removed using a Butterworth IIR filtering stage, which resulted into a third-order IIR filter with -6 dB at 0.5 Hz. Note that we used a bilinear transformation with frequency prewarping to generate the digital coefficients. Afterwards, Automatic Gain Control (AGC) is applied to limit the signal and enhance the peak detection. For the GSR and SKT signals, low-pass FIR filters are employed to remove high-frequency noises.

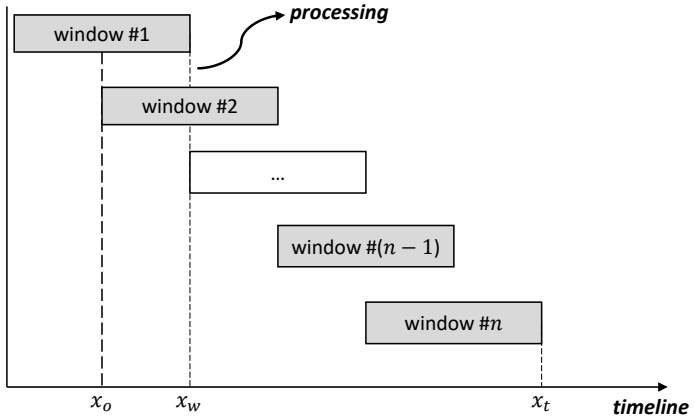


Figure 4-20: Typical data segmentation process in emotion recognition systems based on machine learning.

4.2.3 Feature extraction

In order to improve the results obtained from DEAP-b1 and DEAP-b2, the proposal presented in this Chapter considers and extends the features from the three main groups: time-domain, frequency-domain and non-linear features. This set of features comprises a total of 48 features detailed in Tables 4.12, 4.13, and 4.14, for the three

physiological sensors respectively. Specifically, 25 features for ECG (two in the time domain, nine in the frequency domain and 14 non-linear features), 17 features for GSR (six in the time-domain, three in the frequency-domain, and eight non-linear features) and six features for SKT (four in the time-domain and two in the frequency domain) are included. Note that all considered features are based on accepted, well-known physiological literature dealing with emotional-related features [6, 202, 203] as well as the previous implemented features (DEAP-b1 and DEAP-b2). Moreover, for this system, we have considerably increased the non-linear features considered in our model, which are based on recent publications that included these synthetic metrics in emotion recognition systems as well [160, 161]. The following subsections detail the specific different features extracted for the three different domains and signals.

Table 4.12: Features extracted for the ECG signal and the proposed fear binary emotion recognition using MAHNOB dataset.

Sensor	**Domain**	**Features**
ECG	Time-domain:	Mean of Inter-Beat-Interval
(25)	(2)	Heart rate variability
	Frequency-domain:	Power spectral density of four bands
	(9)	(0–0.1 Hz, 0.1–0.2 Hz, 0.2–0.3 Hz and 0.3–0.4 Hz)
		Inter-Beat-Interval Power spectral density for
		Low frequency (LF) (<0.08 Hz)
		Medium frequency (MF) (0.08–0.15 Hz)
		High frequency (HF) (0.15–0.5 Hz)
		Total energy ratio for MF
		Spectral density ratio between
		LF and HF band
	Non-linear:	Multiscale entropy at five levels
	(14)	Detrended fluctuation for filtered data
		Detrended fluctuation for Inter-Beat-Interval
		Recurrence rate
		Determinism
		Laminarity
		Longest RP diagonal line
		Diagonal lines entropy
		Trapping time
		Correlation dimension

Table 4.13: Features extracted for the GSR signal and the proposed fear binary emotion recognition using MAHNOB dataset.

Sensor	**Domain**	**Features**
GSR	Time-domain:	Filtered data mean value
(17)	(6)	ERSCR including number of peaks
		ERSCR Amplitude and rise time
		Standard deviation
		First quartile
		Third quartile
	Frequency-domain:	Power spectral density of two bands
	(3)	for SCL and SCR components
		(0–0.05 Hz, 0.05–1.5 Hz)
		Spectral density ratio for 0–0.05 Hz
	Non-linear:	Detrended fluctuation for filtered data
	(8)	Recurrence rate
		Determinism
		Laminarity
		Longest RP diagonal line
		Diagonal lines entropy
		Trapping time
		Correlation dimension

Table 4.14: Features extracted for the SKT signal and the proposed fear binary emotion recognition using MAHNOB dataset.

Sensor	**Domain**	**Features**
SKT	Time-domain:	Filtered data mean value
(6)	(4)	Standard deviation
		Skewness
		Kurtosis
	Frequency-domain:	Power spectral density of two bands
	(2)	(0–0.1 Hz, 0.1–0.2 Hz)

Previous to the feature extraction process, the physiological delineation tasks take place. For this system, the raw ECG signal is subjected to peak identification to determine the IBI and extract a valid heart rate estimation and heart rate variability-related parameters. Specifically, a ECG peak detector based on the algorithm devel-

oped by Pan and Tompkins in [204] was applied. Figure 4-21 shows the architecture of such algorithm, which is fed from the ECG filtered signal. The different stages are described as following:

- Differentiator. This is usually conceived as a derivative filter that is responsible to provide information regarding the slope of the morphological ECG wave pattern. It also attenuates low frequency components, which are referred as to the atrial depolarisation and ventricular repolarization. For our case, this process is done based on the first difference of the input filtered ECG signal.
- Squaring. This is a non-linear operation that emphasises the ECG peaks by amplifying the previous derivative result.
- Integrator. As the output of the squared derivative could present multiple peaks within the duration of a single ECG period, a moving window integration filter is used to smooth such signal. The width of this filter is usually set to 150 ms.The output signal of this process is known as integrated signal.
- Threshold Check and Search-Back. These last procedures are intended to identified and corroborate the proper location of the local peaks within the integrated signal. Different physiological constraints are applied to ensure the physiological detection of the ECG peaks, such as 200 ms lockout time between identified peaks. After all the peaks have been identified from the integrated signal, a search-back process is applied to discard and correct those peak to peak or RR intervals causing potential problems. For instance, in our case, we performed a double iteration looking five peaks ahead and assessing the median evolution of the peak to peak vector.

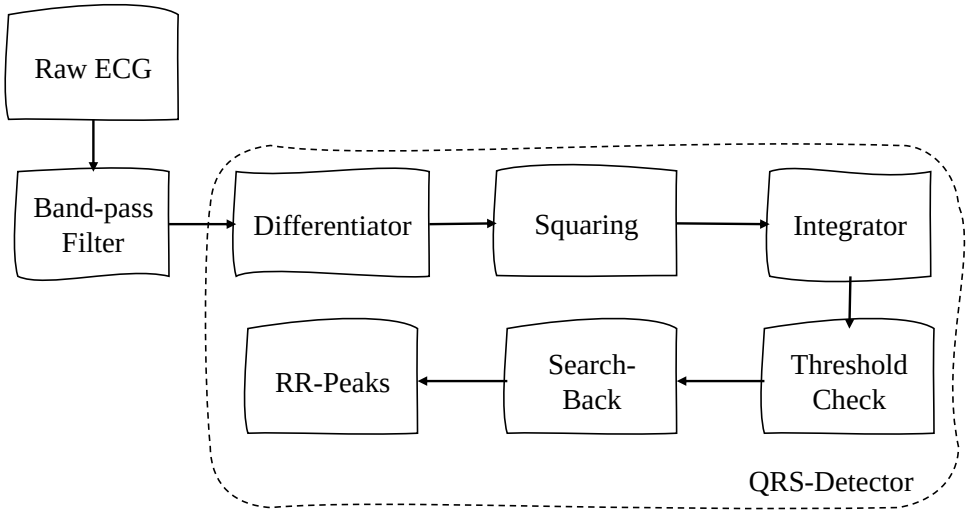


Figure 4-21: Architecture outline of the ECG peak identification algorithm applied in this work.

For the GSR signal, the applied FIR is designed to preserve information below 1.5 Hz, which is the maximum frequency for SCR activity. Such filter is also used with the SKT signal to make profit of storing only one set of filter coefficients. Regarding the GSR delineation, we applied the same processes as done with DEAP data base. Thus, a linear combination followed by equation 2.7 is used, through which the trend of the GSR signal (SCL) is obtained by a moving median filter with a four seconds sliding window. That output is subtracted to the GSR filtered signal, obtaining the SCR component. Both components, as well as the filtered GSR signal, are used to extract synthetic metrics or features detailed in the following sections.

4.2.3.1 Time and Frequency domain

For the time domain features to be extracted in this system, they follow the same distinction or grouping as the ones presented and detailed in Section 4.1.3, as they can be divided between higher-order statistics and morphological features. However, the system proposed in this section extends the higher-order statistics features into two additional metrics: *skewness* and *kurtosis*. Specifically, these are applied to the SKT signal. On the one hand, the former is referred as an indicative of the asymmetry, positive or negative, deviating from a normal distribution and it is given by

$$s = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i - \overline{x})^3}{\sigma^3}, \tag{4.17}$$

where x is the filtered SKT with N samples in this case, and $\overline{x}$ and σ are the mean and the standard deviation for the current processing window. On the other hand,

kurtosis is the statistical metric related to the shape of a probability distribution by measuring degree of concentration presented around the mean of the frequency distribution for a real-value random variable, also described as the measure of the tailedness. This higher-order statistic measurement is given by,

$$k = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_i - \overline{x})^4}{\sigma^4}. \tag{4.18}$$

These statistical moments allow characterising the temporal distribution of the SKT along the visualised stimulus.

4.2.3.2 Non-linear domain

For this system, the set of non-linear features is expanded adding up to eight new features. Most of them are based on chaos theory and time series analysis techniques. These are described and detailed as follows.

- Detrended Fluctuation Analysis (DFA). This is a powerful technique subjected to be applicable if non-stationaries signals are either suspected or known to exist. It allows the estimation of the power law (fractal) scaling or Hurst exponent of a signal coming from a system that is exposed to such non-stationaries [205]. In fact, in this case, this metric gives a measurement regarding the physiological self-similarity at different resolutions (window sizes), which can be translated into physiological complexity assessment. Thus, the time series of length N is first integrated, y, and encapsulated into boxes or windows of length n. These non-overlapped segments are fitted to a polynomial from which the local trend is obtained y_n. Finally, the integrated time series is detrended by subtracting such local trend. The root-mean-square fluctuation $F(n)$ is provided by the following equation:

$$F(n) = \sqrt{\frac{1}{N}\sum_{k=1}^{N}[y(k) - y_n(k)]^2}, \tag{4.19}$$

  which is repeated for all the window sizes to be evaluated. Note that in our case the polynomial fitting is linear (first order) and the number of window sizes at which to evaluate the fluctuations is empirically set that $n \in t_w/10 \ldots t_w$ with 10 samples steps, where t_w is the sample size of the processing window.

- Recurrence rate. This and the following features rely on the mathematical interpretation of the Recurrence Plot (RP)s. These are conceptualised as bi-dimensional plots in which the states of the phase space trajectory of a dynamical system can be represented and quantified. Such states are referred as the recurrences that the system or signal presents over a specific temporal processing window. Such bi-dimensional representation is obtained by the computation of the distances between two states and the comparison with respect a predefined threshold, following

$$R_{i,j} = \Theta(\epsilon_i - ||\vec{x_i} - \vec{x_j}||), \qquad \vec{x_i} \in R^m, \quad i,j = 1,...,N, \tag{4.20}$$

  where i and j are two arbitrary states, ϵ_i is the used threshold for the recurrence evaluation, $\vec{x_i}$ and $\vec{x_j}$ are the respective modulus for each state, and m is the embedded dimension to be considered. Note that the separation t between spaces i and j can be adjusted as desired and needed as well. For our system, we estimated t and m using mutual information [206] and false nearest neighbor [207], respectively, and define ϵ as 10% of the average phase space diameter of observations [208]. Once the RP is obtained, the recurrence rate ratio can be derived from

$$RR = \frac{1}{N^2} \sum_{i,j=1}^{N} R_{i,j}, \tag{4.21}$$

  which corresponds to the correlation sum and it quantifies the amount of detected recurrence states.
- Correlation dimension. This technique is commonly used in time series analysis to characterise the attractor of a dynamical system, i.e. in this case to measure the complexity of a physiological system. An approximated correlation dimension or $D2$ can be computed as

$$D2 \approx \frac{log(RR)}{log(\epsilon)}. \tag{4.22}$$

- Determinism. Based on the RP plot, the diagonal lines provide information about the repetitive physiological patterns of the time series being analysed.

This is quantified by the following equation:

$$DET = \frac{\sum_{l=l_{min}}^{N} lD(l)}{\sum_{i,j=1}^{N} R_{i,j}} = \frac{\sum_{l=l_{min}}^{N} lD(l)}{\sum_{l=1}^{N} lD(l)}, \tag{4.23}$$

where $D(l)$ is the histogram of the different diagonal line lengths. Note that a minimum diagonal line parameter l_{min} must be provided, which in our case is empirically set to 2.

- Laminarity. This feature counts the percentage of recurrence points that form vertical lines within the RP. Those are referred as the chaotic states of the system or non-periodic ones. It is given by

$$LAM = \frac{\sum_{l=l_{min}}^{N} lV(l)}{\sum_{i,j=1}^{N} R_{i,j}} = \frac{\sum_{l=l_{min}}^{N} lV(l)}{\sum_{l=1}^{N} lVl}, \tag{4.24}$$

where $V(l)$ is the histogram of the different vertical line lengths.

- Longest RP diagonal line. The quantification of the longest diagonal line within the RP plot allows to characterise the maximum amount of periodic time within the system. This is given by

$$L_max = max(l_i; i = 1, ..., N_l), \tag{4.25}$$

where N_l is referred as the total number of diagonal lines within the RP plot. In our case, the implementation is done by employing a quick-sorting algorithm using the diagonal lines previously identified.

- Trapping time. As the previous feature is intended to characterise the periodicity of the signal, the trapping time provides information about the amount of non-stationaries states recurred within the RP plot. It is calculated as following:

$$TT = \frac{\sum_{l=l_{min}}^{N} lV(l)}{\sum_{l=l_{min}}^{N} Vl}. \tag{4.26}$$

- Diagonal lines entropy. Finally, to consider the periodicity uncertainty of the signal, the Shannon entropy is applied to the probability distribution of the

diagonal line lengths $p(l)$. This is calculated as:

$$ENT_R = -\sum_{l=l_{min}}^{N} p(l)log(p(l)). \tag{4.27}$$

4.2.4 Fear classification systems

In the following sections, the results obtained with the proposed system using MAHNOB are detailed and explained. Specifically, two systems are presented: subject-dependent and subject-independent. Both use the fear binary mapping from the PAD space analysed in Section 4.2.1. Moreover, unlike the previous DEAP systems, some particularities might be highlighted. First of all, three different classifiers are applied. Two of them are the same classifiers employed for the DEAP systems, SVM and KNN. The third classifier is actually a set of classifiers following an ensemble learning approach. For this purpose, an AdaBoost algorithm is used. Note that the latter has been also reviewed in Section 3.1.7.

In this case and leading towards a better tuning, the hyperparameter optimisation is done through Bayesian optimisation. This technique is intended to minimise the miss-classification rate over iterations, supported by a CV strategy. Specifically, a SMBO technique is included. Thus, the generation of new hyperparameters to evaluate is subjected to Gaussian processes, which approximate the distribution of the cost function $f(x) \sim GP$ (Gaussian Process). This distribution is updated as it is iterated with the new known values for the new hyperparameters. In this way, the final distribution function $p(f(x)|f(x^*))$ is built where x^* refers to the historical values. With this estimation, the point that could be a potential candidate is calculated in the next step. For this, a function $\alpha(\cdot)$ called acquisition is used. For the definition of this acquisition function, there are different options. In this case, the probability of improvement strategy has been used, which tries to estimate the probability of an improvement with the next sample.

Regarding the validation procedure, the subject-dependent and subject-independent models were validated based on a stratified $k-fold$ CV schema ($k = 5$). On the one hand, for the subject-dependent models, the mean of all metrics for all volunteers and the mean absolute deviation (MAD) were calculated based on the obtained CV values. On the other hand, the subject-independent models were divided into train-

ing, validation, and testing sets, employing a LOSO strategy. The latter allowed us to study the performance of various subject-independent systems trained with different subject combinations and tested with a single volunteer about whom the system had no information.

The presented results and system were published in [187]. Note that in this case, we did not implement a feature selection or reduction process nor applied any miss-classification cost methodology (cost-sensitive learning). The rationale of the latter decision was based on obtaining baseline results to be compared against future system improvements when adding and applying different techniques.

4.2.4.1 User-dependent results

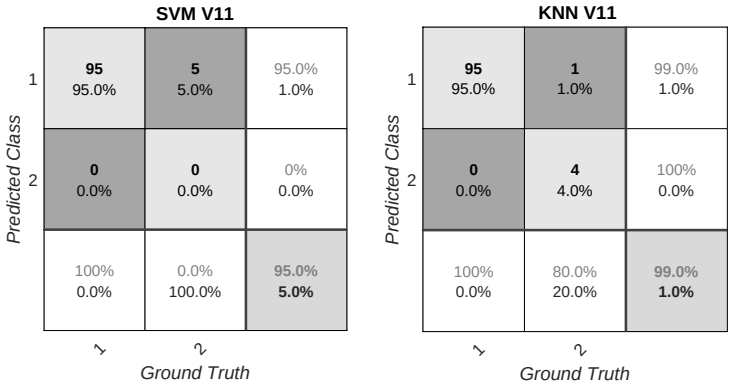
Table 4.15 shows the validation performance metrics and dispersion for the different light-weight classification algorithms selected for the generation of each subject-dependent model for all volunteers. After analysing the results, it can be observed that there was no strict dependence relationship between the class distribution and performance. Nonetheless, the performance of the models was directly affected by the type of classifier used. Moreover, another key factor that could have influenced performance was related to the alignment of subject-dependent physiological patterns and the binary fear mapped labels obtained. Furthermore, the usage of Gmean and F1 scores allowed us to distinguish the low-performance models from the higher-performing models more robustly.

Table 4.15: Performance metrics for each generated subject-dependent model and average performance metrics and dispersion for each classification algorithm.

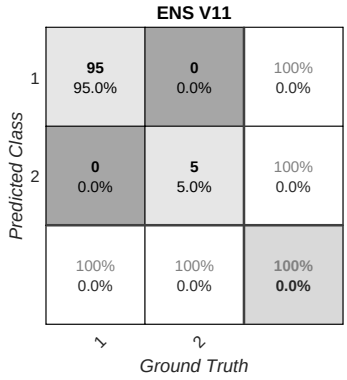
		SVM				**KNN**				**ENS**			
Training Type	**Trained Volunteers**	**ACC (MAD)**	**AUC (MAD)**	**Gmean (MAD)**	**F1 (MAD)**	**ACC (MAD)**	**AUC (MAD)**	**Gmean (MAD)**	**F1 (MAD)**	**ACC (MAD)**	**AUC (MAD)**	**Gmean (MAD)**	**F1 (MAD)**
Subject dependent	V1	89.00%	90.30%	87.73%	85.71%	88.00%	88.67%	87.90%	85.37%	88.00%	79.32%	85.12%	83.33%
	V2	88.00%	92.43%	76.41%	71.43%	99.00%	99.89%	99.23%	87.72%	91.00%	97.47%	85.41%	80.85%
	V3	91.00%	94.44%	71.20%	74.29%	94.00%	96.19%	90.47%	85.00%	97.00%	95.31%	98.13%	93.02%
	V4	93.00%	95.29%	84.06%	75.86%	99.00%	96.67%	96.59%	96.55%	96.00%	99.69%	85.62%	84.62%
	V5	76.00%	84.97%	75.01%	72.09%	81.00%	91.47%	85.62%	84.62%	98.00%	99.88%	97.95%	97.78%
	V6	90.00%	93.67%	87.92%	83.33%	98.00%	98.60%	84.08%	82.86%	99.00%	99.90%	99.23%	98.36%
	V7	93.00%	98.54%	92.47%	91.14%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	V8	85.00%	90.57%	81.22%	74.58%	94.00%	92.24%	90.58%	89.29%	93.00%	86.05%	92.12%	88.52%
	V9	96.00%	98.44%	83.16%	77.78%	99.00%	99.44%	99.40%	95.24%	100.00%	100.00%	100.00%	100.00%
	V10	89.00%	91.31%	87.73%	85.71%	94.00%	94.15%	93.24%	92.31%	100.00%	100.00%	100.00%	100.00%
	V11	95.00%	50.00%	00.00%	00.00%	99.00%	90.00%	89.44%	88.89%	100.00%	100.00%	100.00%	100.00%
	V12	77.00%	83.48%	62.91%	53.06%	91.00%	85.95%	84.97%	83.02%	94.00%	93.33%	90.58%	89.29%
		88.50% (4.66%)	88.62% (7.90%)	74.15% (14.72%)	70.42% 14.62%	94.66% (4.33%)	94.44% (4.02%)	91.80% (4.92%)	89.24% (4.53%)	96.33% (3.28%)	95.91% (4.94%)	95.51% (5.62%)	92.98% (6.38%)

However, the presented results could be biased due to the reduced amount of data available (100 samples per volunteer, five windows on average per video), as well as due to the asymmetry detected (imbalanced data). Focusing on asymmetry, this problem is especially relevant in V11. The effect on performance due to asymmetry for this volunteer is shown in Figure 4-22, which provides the confusion matrices for V11 after applying all three algorithms. Conversely, the confusion matrices of the volunteer V7 are also shown in Figure 4-23. This volunteer showed the best performance overall; i.e., considering the different metrics for the three classifiers applied. In these figures, the positive class (fear) is represented by the number two, and the negative class (no fear) is represented by the number one. The rows correspond to the predicted class and the columns correspond to the true class or ground truth. From left to right and from top to bottom, each confusion matrix shows the true-negatives, false-positives and false omission rates. The next row shows the false-negatives, true-positives and precision rate. The last row shows the false-negative rate, specificity and overall accuracy. Note that the rest of the confusion matrices for each subject-dependent model generated are shown in [187].

After analysing these confusion matrices, the performance of the algorithms for V11 was also found to be asymmetric. Thus, for instance, SVM provided a high accuracy, at up to 95.00%, but this metric was biased by the reduced number of samples of this volunteer within the positive class (only five samples). In this case, the calculated Gmean and F1 metrics results were 0.00% due to the zero positive predicted rate, and the AUC was 50.00%, showing that this classification model performed no better than random guessing. The behaviour shown by SVM in this case matched the usual unreliable performance of this algorithm for extremely imbalanced distributions; that is, SVM is oriented towards the majority class to optimise the error rate during the training stage. On the contrary, boosting algorithms usually provide a better behaviour for imbalanced distributions, as shown by ENS for this case. Nevertheless, this imbalanced situation should be avoided during the database generation, and the quality and diversity of the stimuli considered should be improved. In the case that this situation were not addressed during the database generation, the bias generated in performance could be partially solved by selecting an adequate classification technique, as discussed above. However, the lack of information from

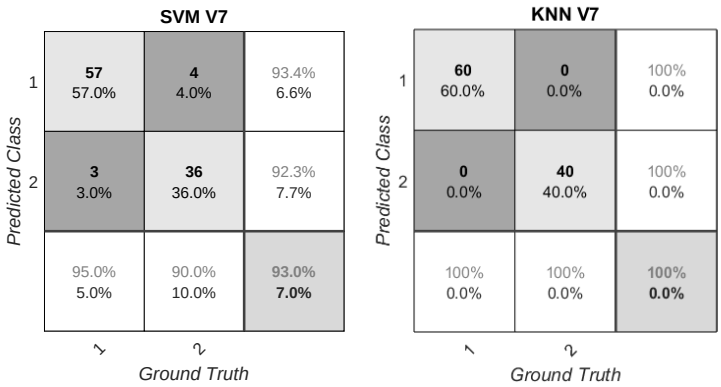


(a) SVM classifier (b) KNN classifier

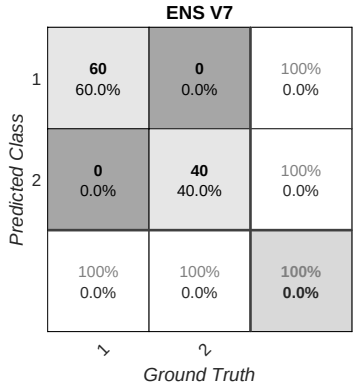


(c) ENS classifier

Figure 4-22: Confusion matrices for a subject-dependent model in V11, detected as a problem in asymmetry.



(a) SVM classifier (b) KNN classifier



(c) ENS classifier

Figure 4-23: Confusion matrices for a subject-dependent model in V7.

one of the two classes cannot be solved, resulting in a possible incorrect classification for future samples [209]. Another possible approach to deal with this problem is based on the application of data augmentation techniques or weighted classes, as previously applied for DEAP-b1 and DEAP-b2. Conversely, in the case of V7, the system showed 40.00% positive class information, which translates into a better SVM performance. KNN and ENS continued to outperform SVM due to the reasons stated above for the error rate optimisation of this classifier.

4.2.4.2 User-independent results

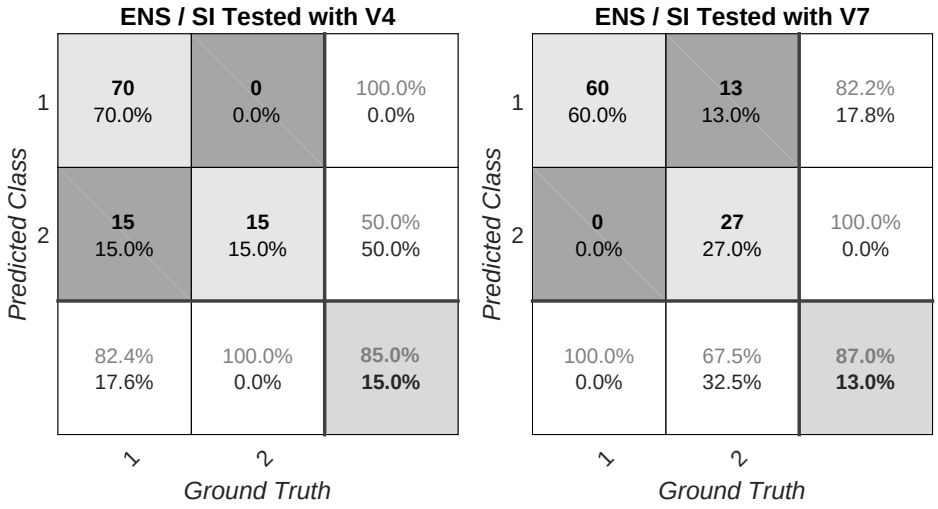
Focusing on the subject-independent use case, the combination of all individual samples resulted in a bigger dataset with 1200 samples (100 samples per volunteer $\times$ 12 volunteers). The physiological signal ranges differed for different individuals due to the nature of each individual and the differences in the measurement set-up (e.g., ambient temperature). Therefore, the data (features) from every volunteer should be normalised. To this end, we considered the Z-score method. Once the database was normalised, the binary-fear recognition system was generated using a $k - fold$ CV schema for the validation partition and a LOSO testing methodology.

Table 4.16 shows the performance metrics for each classification algorithm in the generation of the subject-independent model. Note that the training of these models was performed using all volunteers except the one used for testing in each iteration (unseen test data); i.e., a total of 12 subject-independent models were generated and tested.

Table 4.16: Performance metrics for each generated subject-independent model and average performance metrics and dispersion for each classification algorithm. The training stage is performed using all the volunteers except the tested volunteer in each model generated (unseen test data).

		SVM				**KNN**				**ENS**			
Training Type	**Tested Volunteers**	**ACC (MAD)**	**AUC (MAD)**	**Gmean (MAD)**	**F1 (MAD)**	**ACC (MAD)**	**AUC (MAD)**	**Gmean (MAD)**	**F1 (MAD)**	**ACC (MAD)**	**AUC (MAD)**	**Gmean (MAD)**	**F1 (MAD)**
Subject independent	V1	65.00%	60.83%	57.15%	47.76%	75.00%	71.25%	68.74%	62.69%	71.00%	65.83%	60.55%	52.46%
	V2	70.00%	61.33%	58.83%	42.31%	81.00%	74.00%	72.66%	61.22%	82.00%	86.72%	71.26%	60.87%
	V3	64.00%	66.00%	62.44%	40.00%	72.00%	61.88%	59.53%	39.13%	62.00%	61.19%	45.82%	24.00%
	V4	82.00%	71.01%	83.88%	59.09%	84.00%	87.84%	87.67%	63.64%	85.00%	91.61%	90.75%	66.67%
	V5	64.00%	70.55%	61.10%	55.00%	70.00%	71.74%	65.32%	59.46%	73.00%	75.58%	68.16%	63.01%
	V6	84.00%	88.57%	85.61%	77.14%	71.00%	68.81%	68.59%	56.72%	79.00%	87.86%	76.16%	66.67%
	V7	75.00%	90.54%	65.38%	59.02%	76.00%	91.83%	69.37%	63.63%	87.00%	99.46%	82.16%	80.60%
	V8	76.00%	81.90%	70.51%	60.00%	78.00%	72.86%	71.71%	62.07%	80.00%	85.00%	75.59%	66.67%
	V9	67.00%	69.67%	63.77%	21.82%	67.00%	59.44%	58.69%	18.87%	78.00%	84.78%	78.88%	42.11%
	V10	76.00%	79.63%	65.95%	60.00%	78.00%	72.92%	68.34%	63.33%	77.00%	82.30%	72.80%	67.61%
	V11	74.00%	90.53%	76.78%	23.53%	80.00%	89.47%	88.85%	40.00%	74.00%	86.32%	85.22%	27.78%
	V12	70.00%	72.05%	64.14%	51.61%	71.00%	66.90%	66.12%	53.97%	72.00%	67.72%	66.73%	54.84%
		72.25% (5.58)	75.22% (9.18)	67.96% (7.48)	49.77% (12.24)	75.25% (4.25)	74.07% (7.82)	70.47% (6.50)	53.73% (10.53)	76.67% (5.22)	81.20% (9.07)	72.84% (8.62)	56.11% (13.22)

After analysing this table, the best results were also provided by ENS, with the highest averaged performance metrics (81.20%, 72.84%, 56.11%) for the AUC, Gmean and F1-score. On the contrary, SVM also provided the worst performance in general. The differences between all the subject-independent models generated should be highlighted. For instance, the best model achieved a Gmean of up to 90.75% when testing with V4 and training with the rest of the volunteers, and the worst model provided a Gmean of up to 45.82% when testing with V3 and training with the rest of the volunteers. This fact emphasises the need for a larger and more balanced data set to deal with these problems. Regarding the F1-score, a high variability can be observed among the different models. By definition, this score is a weighted harmonic mean between precision and recall, which leaves true-negatives out of the equation. This fact is key when presented with a very low positive incidence, but a high F1-score does not necessarily imply a better performance of the system. For instance, the confusion matrices of two subject-independent tested models for the ENS classifiers are shown in Figure 4-24 for V4 and V7 with F1-scores of up to 66.67% and 80.60% respectively. Based on the pursued fear recognition application, it could be more convenient to have a miss-classification for the false-positive than over the false-negative. Therefore, comparing the F1-score for different subject-independent models should be accompanied by the requirements and needs of the application. Note that both of the explained examples did not show a perfect classification performance. The rest of the confusion matrices for each subject-independent model generated are provided in [187].



(a) (b)

Figure 4-24: Confusion matrices for ENS classifiers and tested volunteers (unseen data) over their respective subject-independent models: (**a**) tested V4, (**b**) tested V7.

Regarding the time and space complexity of the employed models, SVM and KNN were already discussed in previous sections. The remaining model, ENS, is actually an AdaBoost classifier, which is based on a single composite strong learner. The latter is made up of different weak learners that, in this case, are shallow trees. Thus, two parameters are essential to estimate the time and space complexity: the number of trees, and the maximum number of splits per tree. On the one hand, the time complexity used to be defined by $\mathcal{O}(feats * n_{trees})$ for this type of classifier, where $feats$ is the number of features and n_{trees} is the total number of trees. Note that the time complexity of the trees is not included within the total time complexity of AdaBoost, as it is negligible in comparison to the total amount of time. On the other hand, the space complexity is determined by the amount of the trained shallow trees and the maximum number of allowed splits within each of those. Additionally, the trained weights for the weak learners must be also stored. For the worst case scenario, when dealing with the subject-independent models which are more complex than the subject-dependent, the number of trees used to be one quarter of the training data set, i.e., approximately 300 trees on average (1,200 training instances), and the maximum number of allowed splits per weak learner on average is ten. Considering these values and the total set of 48 features, the estimated time complexity achieves up to 14400 operations, while the space complexity reaches approximately 13 kB (300 trees $\times$ 10 splits maximum + 300 trained weights). Note that the mem-

ory used in kB is based on a 32-bit integer data type for all the parameters to be stored.

It is noteworthy that further AdaBoost and tree algorithmic optimisations, as well as other feature selection alternatives, can be implemented and applied to achieve a smaller time complexity, which can even lead in some cases to better recognition performance. Although redundant information does not affect AdaBoost as negatively as for other classifiers, such as SVM, the elimination of irrelevant information does affect and can yield up to less computational time. In fact, Table 4.17 shows the results for the same training configuration, validation and testing as the subject-independent case employing the AdaBoost classifier, but changing the feature selection method to mrMR with $K = 10$. Note that this technique was also applied to DEAP-b2 system. It can be observed that the obtained metrics are similar to the ones presented without feature selection. However, the reduction from 48 up to 10 features directly affects the inference time complexity. Thus, with this configuration and considering the same number of trees on average, just 3000 operations need to be done.

Table 4.17: Performance metrics for each generated subject-independent model and average performance metrics and dispersion for ENS after mrMR feature selection. The training stage is performed using all the volunteers except the tested volunteer in each model generated (unseen test data).

		ENS			
Training Type	**Tested Volunteers**	**ACC (MAD)**	**AUC (MAD)**	**Gmean (MAD)**	**F1 (MAD)**
Subject independent (mrMR)	V1	90.00%	68.42%	59.37%	50.70%
	V2	87.27%	87.25%	76.18%	64.15%
	V3	87.73%	65.63%	45.00%	23.08%
	V4	88.64%	84.55%	84.48%	57.14%
	V5	90.45%	86.83%	61.50%	54.55%
	V6	92.27%	90.57%	85.22%	78.13%
	V7	87.27%	95.71%	78.58%	87.02%
	V8	90.91%	94.90%	92.38%	87.50%
	V9	88.18%	91.56%	81.10%	47.06%
	V10	85.00%	97.17%	79.06%	88.89%
	V11	85.00%	99.79%	88.85%	33.33%
	V12	93.64%	69.62%	53.67%	39.29%
		88.86% (2.16)	86.00% (9.30)	73.78% (12.60)	59.24% (18.25)

4.3 Discussion and Conclusion

This Chapter presented the work realised towards a fear detection system by using publicly available datasets. Part of the presented work is also contained in published articles [182, 184, 187]. Throughout the design of the different systems, the essential processes to be considered critical for an embedded implementation have been identified and initially addressed. For instance, the main focus of discussion is the time and space complexity of the resultant models in comparison with their performance metrics. As stated at the beginning of the Chapter, the design of a fully subject-independent model would allow the first generation of a technological tool able to detect any emotion based on machine learning. This tool can be customised during the operation with subject data collected. For example, in the UC3M4Safety team, the fear detection under Gender-based Violence situations has been the seed of this research work. Subject-dependent models requires having enough data to make different training, validation and testing sets statistically significant. In case of having sufficient information of a particular subject, then a subject-dependent model can be generated and even pursued, as it archives better performance than a subject-independent model. However, in most of the cases, when dealing with real-life applications, in which during the first deployment moment there is no or little amount of data of that particular subject, then it is necessary to implement a subject-independent model.

Table 4.18 summarises the best results obtained along this part of the work for the fear binary emotion recognition when dealing with a subject-independent model. As it can be observed, different hyper-parameter optimisation techniques, classifiers and system configuration (with and without feature selection) were explored. First of all, the DEAP-b1 system used 21 subjects, employed a KNN by using a specific extreme Hold-Out strategy, and achieved a Gmean of up to 62.00%. The particularity of such system was that the filtered signals were considered as inputs, as no feature extraction was applied. Thus, as already commented in previous sections, that fact could lead to overoptimistic results. Note that no hyper-parameter optimisation as such was applied to this system, as the values obtained during the parameter sweeping for the subject-dependent models were used. Due to the observed limitations for DEAP-b1 in terms of space complexity, DEAP-b2 was developed targeting

a lighter classifier. The latter considered the whole set of volunteers from DEAP at the expense of omitting one of the physiological signals (SKT). In this case, the singularity was based on the fear binary mapping origin, which was directly obtained from the PA space rather than using the PAD space. Towards increasing the performance of the DEAP-b2 system, the elimination of the redundant features and the maximisation of the relevant ones throughout mrMR led to the DEAP-b2+ system, which provided a better performance than DEAP-b2 and DEAP-b1. Thus, in this case, the application of feature selection techniques resulted into a vital step for the improvement of the system. However, the space complexity remained the same. Finally, the limitations faced with DEAP were fixed by using the MAHNOB database. Focusing on the tree-based classifier, we developed two systems, with and without feature selection. In this case, the testing CV technique applied was LOSO, which offers a non-overoptimistic view of the system performance. Thus, due to the specific characteristics of the classifier, when applying feature selection we achieved similar metrics for Gmean and AUC, and obtained the smallest storage for the model.

Table 4.18: The best results obtained along Chapter 4 for the fear binary emotion recognition when dealing with a subject-independent model.

System	DEAP-b1	DEAP-b2	DEAP-b2+	MAHNOB-fear	MAHNOB-fear+
Subjects	21	32	32	12	12
Signals	PPG, GSR, SKT	PPG, GSR	PPG, GSR	ECG, GSR, SKT	ECG, GSR, SKT
Hyp.Opt.	-	Grid Search	Grid Search	SMBO	SMBO
Classifier	KNN	SVM-RBF	SVM-RBF	ENS-AdaBoost	ENS-AdaBoost
CV Technique	$Hold-Out$	$k-fold$	$k-fold$	LOSO	LOSO
Space (kB)	48.59	39.06	39.06	13	13
AUC (MAD)	-	62.79 (4.72)%	81.60 (8.70)%	81.20 (9.07)%	86.00 (9.30)%
Gmean (MAD)	62.00%	62.62 (4.73)%	81.55 (10.21)%	72.84 (8.62)%	73.78 (12.60)%

Focusing on the last proposed systems, MAHNOB-fear and MAHNOB-fear+, certain limitations must be considered. On the one hand, the data segmentation approach used presents some disadvantages when dealing with slow-changing physiological signals. Different techniques should be applied to take into account all the different physiological particularities without wasting information. For instance, specifically for the GSR, the use of dynamic data segmentation and overlapping could be a valid solution. However, when dealing with resource-constrained devices,

a better solution might be to keep track of the onsets of the ERSCRs and, when detecting the offsets for the successive processing windows, calculate all the ERSCRs metrics. The main advantage of this latter method is the independence of the processing window length at the expense of storing the ERSCR tracking information until the completion of the ERSCR (offset). On the other hand, despite using a specific normalisation technique (Z-score), other approaches might be exploited. For instance, we are already working on applying different normalisation techniques, such as using recovery time-slots to normalise the data of the emotion-related stimulus and study the effect for the analysed fear use case. Finally, it should be noted that the results shown are limited by the size of the dataset considered (12 subjects), which is the weakest point of such models. As no other dataset exists that fits our use case, a larger and better dataset is required to create a more reliable system. Therefore, the limitations identified while developing these systems confirm the relevance of creating a novel dataset focused on fear detection. This dataset should include some key facts, such as the usage of emotional immersive technology, the modification of the labelling methodology to consider the gender perspective, a properly balanced stimuli distribution regarding the target emotions and a greater number of participants. More details regarding the latter fact and the new UC3M4Safety database are given in Chapter 6.

Regarding the comparison with other research works, the wide casuistry of the emotion recognition problem is a challenging task. This is due to the high amount of different techniques that can be applied within the data processing chain and the generation of the machine learning model. However, we can make a clear distinction by using five factors: a) CV used for validation and/or testing, b) the number of subjects accounted for, c) emotion classification paradigm (binary, discrete and /or dimensional multi-emotion detection), d) the amount and type of used signals, and e) usage of publicly available datasets. The latter is of tremendous importance, as those works based on open databases can be further directly compared without digging into experimental methodology differences discussions. Table 4.19 lists the above factors with respect to some of the main state-of-the-art works that are directly linked to and have influenced this one. At first glance, we observe a wide variety of techniques, which makes it difficult to compare. First of all, only research works

that are directly related to the fear detection or to emotion classification have been selected. In fact, two works are based on discrete emotion classification, seven of them are focused on arousal and valence classification (different levels) using the PA model, and three research works classify emotions by means of the PAD model. From the latter, two of them [185,186] are the ones already reviewed in Section 4.1 that use our proposed fear binarization paradigm. Secondly, only six of the works employed a Leave-One-Out (subject or trial) CV technique. The others applied $k - fold$ and Hold-Out, which, based on the data arrangement, can lead to overoptimistic results. Moreover, regardless of the emotion classification paradigm and the applied CV technique, most of the works did not report many machine learning performance metrics, apart from accuracy. Finally, considering these contextualisation aspects, we can conclude that the obtained metrics are in line with the state-of-the-art.

Table 4.19: Most recent and main state-of-the-art works that are directly linked to and have influenced this research in terms of affective computing using physiological information.

	Subjects	Signals	Classifier	CV	Emotion	Dataset	Metrics
Lisetti and Nasoz [210]	14	ECG,GSR,SKT	KNN	LOO	Sadness, anger, fear, surprise, frustration, amusement	own	ACC(fear): 85.6%
Chanel et al. [211]	10	BP,EEG,GSR, PPG,RESP	SVM	LOSO	PA space calm-neutral vs. positive-excited	own	ACC: 66.00%
Valenza et al. [167]	35	ECG,GSR,RESP	QDA	40-fold CV	Five arousal and valence levels	own	ACC > 90%
Valenza et al. [161]	30	ECG	SVM-RBF	LOO	Two levels arousal and valence	own	ACC(V):79.00% ACC(A): 84%
Abadi et al. [145]	30	ECG,EOG,EMG	SVM	LOTO	Two levels arousal, valence, dominance	DECAF	ACC(A,V,D):50-60%
Rubin et al. [160]	10	ECG	SVM	$k-fold$	Binary Panic detection	own	ACC:73-97%
Rathod et al. [212]	6	GSR,PPG	SVM	Hold-Out	Normal, happy, sad, fear, anger	own	ACC < 87.00%
Zhao et al. [213]	15	PPG,GSR,SKT	NB,RF,SVM	LOSO	Four PA quadrant	own	ACC:76.00%
Marín Morales et al. [79]	60	EEG,ECG	SVM	LOSO	Two levels arousal and valence	own	ACC:75-82%
Santa Maria Granados et al. [163]	40	ECG,GSR	CNN	Hold-Out	Two levels arousal and valence	AMIGOS	ACC:71-75%
Miranda et al. [184]	15	PPG,GSR,SKT	RF	Hold-Out	Fear (PAD binarized)	DEAP	ACC:54.00%
Amani Albraikan et al. [196]	25	GSR,ECG,EEG, RESP,SKT	ENS	$k-fold$	Three levels arousal and valence	MAHNOB	ACC:94.00%
Miranda et al. [182]	32	PPG,GSR	SVM	$k-fold$	Fear (PA binarized)	DEAP	ACC:62.80%
Oana Balan et al. [185]	32	EEG and peripheral	RF	$k-fold$	Fear (our paradigm)	DEAP	ACC:89.96%
Miranda et al. [187]	12	ECG,GSR,SKT	ENS	LOSO	Fear (PAD binarized)	MAHNOB	ACC:76.67%
Oana Balan et al. [186]	32	PPG,GSR	Boosting	$k-fold$	Fear (our paradigm)	DEAP	ACC:91.70%

Part III

Towards a new fear detection paradigm for gender-based violence situations

Chapter 5

A new autonomous system for emotion recognition: Bindi

As stated in Chapter 1, one of the main goals of this research is focused on providing a smart technological solution to prevent and fight against the Gender-based Violence. On this basis, the Bindi system is proposed, Figure 5-1. This is an autonomous multimodal system that considers Internet of Things (IoT) technologies towards the detection of risky situations under Gender-based Violence contexts. Specifically, the edge-computing part of the system is conceived as a smart cyber-physical network able to detect fear-related emotions. This is accomplished by means of physiological and physical (audio and/or speech) smart sensors continuously monitoring the user. Such task is completed by a fog-based multimodal data fusion within an ad-hoc smartphone application. Finally, in case of confirming a risky situation, an alarm is triggered to a predefined protection network. Moreover, the information is sent to specific computing servers in the cloud, which are responsible to store the collected data for further legal actions. The design of such a system boosts the generation of new mechanisms for the prevention and fight against Gender-based Violence.

In this Chapter, in first place, we carry out a detailed study regarding the current systems and tools to prevent gender-based violent aggresions. This is done considering different perspectives such as commercially available devices, research-grade systems, and institutional tools. Note that the latter focuses on Spanish institutions, due to Spain's global leadership in this respect, as detailed in Chapter 1.

Figure 5-1: Simplified Bindi system architecture based upon the different IoT technologies.

Moreover, the different technological competitive advantages of Bindi are compared and highlighted. This analysis is followed by a comprehensive description of the Bindi system. Thus, we tackle-down the different hardware and software designs within Bindi's bracelet. First of all, the system architecture is detailed in terms of both design and integration. This is accompanied with different physiological wearable integration recommendations to be considered for the following versions of the system. Secondly, the current embedded implementation is reported and explained. Note that the results provided in this Chapter have been presented in different publications [9, 159, 184, 214].

5.1 Current technology to fight against Gender-based Violence

The development of technology over the years has made the generation and application of new tools to prevent Gender-based Violence a reality [7, 20, 21]. The advantages of using technological tools to help combat this problem are manifold:

- Protection accessibility. Technology can make access to victim protection easier and closer.
- Information centralisation. Different institutions and/or forces can cooperate towards a joint monitoring of the circumstances surrounding the Gender-based Violence Victims.
- Multi-modal information gathering. The collection of diverse sources of information can be further used for prediction and prevention analysis. Moreover, this allows for a better understanding of the specific situation of the victim.

- Action response times. The previous points directly affect to the time involved within the decision making regarding the activation of the respective institutional mechanisms.
- Security reinforcement. From the user perspective, the inclusion of reliable and robust technology can provide an stronger security feeling in the Gender-based Violence Victims.

However, such advantages are also accompanied by different requirements, considerations and open questions, which can be summarised in:

- The pseudo-anonymisation of the stored data is crucial. Any technology must ensure that all sensitive or identifiable data is protected and secured. The management and ownership of such information needs to be carefully considered. Thus, strict compliance with data protection laws must be ensured. Moreover, any technological solution should ensure the chain of custody of the information collected so that it can be used later in any judicial process.
- The technological tools candidates are expected to directly connect the victims with specialised professionals. This advocate for: 1) the need for more trained professionals to deal adequately with Gender-based Violence Victims, and 2) the elaboration of new protocols aiming to avoid re-victimisation.
- An alignment amongst the proposed technological solutions, the government, and private stakeholders is of paramount importance. Note that the latter play a key role in terms of technologically-based solution developing and integration.
- Technological personalisation must be considered as an essential aspect, as there is a strong need for the technological solution to be tailored and personalised to each person. This is due to the adaptation to different contexts and heterogeneous settings. However, this might face the current limitations of technology to achieve such an adaptation.
- Accessibility to the proposed technologically-based solutions. It is known that there exists an approximately 7% mobile gender gap in mobile ownership in low- and middle-income countries [215]. This fact, accompanied by the less perceived income by women, makes the target price and the technology platform critical factors. The former is related to the affordability of the solution, while the latter refers to the fact that solutions with no need for mobile-phone

technology would help make the solution more inclusive.

All these stated points justify and foster the multidisciplinary approach claimed in Chapter 1 that is needed towards the design, development, and integration of technology that deals with Gender-based Violence contexts. Although this is a challenging task, one of the main goals of this research is to provide the necessary technological basis to start solving these stated problems and open questions. It is noteworthy to highlight that any technologically-based solution designed and oriented towards the Gender-based Violence casuistry might help prevent and combat, but it will never solve the whole problem. That, is an educational matter.

One of the most common employed technologies are the mobile-phone based applications. Nowadays, this technology is one of the most widely accepted despite the considerations described above. Recently, the authors in [216] provided a systematic review of up to 171 applications whose goal was to address Gender-based Violence throughout different mechanisms. Regardless of the specific application, the authors concluded that most of them were mainly focused and designed for short-term or one-time emergency solutions. This fact leaves aside the prevention perspective and provides the possibility of identifying solely isolated events of Gender-based Violence rather than offering a continuous monitoring and the self-empowerment of Gender-based Violence Victims, which should be one of the main goals. Although the authors stated that educational features are being increasingly included in recent applications, further investigations related to the data security, personal safety and efficacy of such solutions need to be carried out. An example of one of these applications is AlertCops [217]. This application is specifically promoted by the Spanish Home Office and it allows instant notifications of any type of incident with the law enforcement agencies. As a differentiating feature with respect to other existing applications, in the last year, the "SOS Button" has been added to this application, which allows reinforced protection for vulnerable groups. This button sends an urgent alert to the nearest police centre along with its Global Positioning System (GPS) location and a 10-second audio recording of what is happening. Moreover, this application also includes the "Guardian" feature, which has been lately included by many other applications. Specifically, it allows to share the real-time location with user-selected contacts. Although these applications can suc-

cessfully exploit the diverse mobile technology capabilities, the decision making in any case is based solely environmental or relative measures, but never to measures of the user herself.

Within this context, private stakeholders have also developed technological tools that could potentially be used to deal with the discussed use case. However, most of these solutions are included in the category of panic buttons. Even in some countries, such as India, a directive was issued related to the mandatory inclusion of a panic button on every mobile phone sold from 2017 onward. One of the highlighted panic button solutions that is specifically intended to deal with Gender-based Violence situations is SaferPro by LeafWearables, an Indian company. This is a bracelet device that comes with a low power subscriber identity module card, which makes it independent from the mobile-phone. Specifically, once the button is pressed by the user, an alarm is sent to a selected circle of responders and an audio recording is started. However, panic buttons present significant limitations regarding women's safety: 1) the requirement of an active role in their self-protection, which is certainly not possible under some types of aggression and/or blocking emotional reactions, 2) their lack of inconspicuous design that can lead to users' stigmas, and 3) the lack of infrastructure support [218]. Despite the technological efforts, this type of approach is questioned by several Gender-based Violence experts [19], who demand, among other things, more advanced research and technology in these solutions that are regarded as outdated and a higher degree of attention to the role of the victims. Apart from panic buttons, there are also commercial available devices that, although they are not exactly oriented to the use case of Gender-based Violence, allow to generate alarms automatically regarding internal and external detected abnormal events to the user. For instance, the Apple® Watch Series 4 and on-wards provide fall detection and send a SOS to predefined emergency contacts in case no action is performed by the user. The Embrace2 bracelet from Empatica is the only FDA-cleared wrist-worn wearable in epilepsy, which triggers an alarm in case of seizure detection. This is done by means of GSR monitoring. Moreover, it is equipped with other three sensors (SKT, accelerometer, and gyroscope) that can be also acquired and stored for medical purposes. This latter system also opens up the possibility to consider the use of similar sensing technology to tackle Gender-based Violence.

Amongst the recently launched advanced physiological sensing devices, Fitbit® with FitbitSense and Oura with OuraRing stand out. The former is the only commercial smart-bracelet offering more than two integrated physiological sensors: GSR, ECG, PPG, and SKT. However, the current electromechanical integration of some of these sensors hampers the application of this device to other use cases. This is mainly due to the fact that acquiring a measurement from GSR and/or ECG requires the free-hand to be on top of the bracelet, as this provides a close loop circuit. The latter system is based on a smart-ring and provides PPG and SKT acquisition with a relatively high accuracy. However, note that the niche market for these devices is focused on generic wellness, rather than any other specific use case. Up to my knowledge, the only commercially available device oriented towards providing a tool for preventing a specific physiological-related condition is the mentioned Embrace2. Nevertheless, the proliferation of commercial wearable devices with physiological sensing capabilities has been booming in recent years and it could benefit the design and development of tools oriented towards the target application of this research.

The public sector has not been oblivious to technological developments. When dealing with electronic monitoring to help prevent Gender-based Violence, Spain turns out to be one of the pioneering countries in the world promoting this type of technology. In fact, as already reviewed in Chapter 1, in 2013 an agreement was signed between the Spanish Home Office, Justice, Health and Social Services and Equality General Council of the Judiciary and the State Attorney General's Office approving the "Protocol of action of the monitoring system by telematic means of the measures and sentences of restraint in matters of gender violence". These measures compel the aggressor and the victim to carry different devices, Figure 5-2. Moreover, all the different alarms generated by the system are monitored and centralised by a specialised centre called Cometa, which is run by a private company (Securitas Direct) subcontracted by the Spanish government. The considered stakeholders as well as the information centralisation that this system provides is in line with the previous stated advantages and requirements. However, particularly for this system, the employed technology is outdated and solely based on GPS monitoring, which in some cases results in the aggressor harassing the victim even more. Although this protocol and technological solution have aided in the combat of Gender-based

Violence during the last years, its GPS-based sensing technology added to its low battery and intermittent reported failures, makes this solution very limited.

Figure 5-2: Devices considered for the electronic monitoring system within the "Protocol of action of the monitoring system by telematic means of the measures and sentences of restraint in matters of gender violence". DLI: Device worn by the aggressor; DLV: Device worn by the victim [7].

In case of looking for proposals in the academia regarding the design of systems and tools towards the avoidance of Gender-based Violence, there is literature [219]. However, as for most of the commercially devices, these are focused on the design and optimisation of panic button based systems. It is worth mentioning that when leaving aside the wearable physical final output, there is also literature that directly deals with the Gender-based Violence issue by means of machine learning applied to centralised information. For instance, the authors in [220] used machine learning to design models that accurately predict the recidivism risk of a gender-violence offender. They employed 40,000 reports of gender violence extracted from VioGen and outperformed the preexisting risk assessment algorithm based on classical statistical techniques. Apart from that, there is a lack of system proposals in the literature targeting the prevention and combat against Gender-based Violence.

From this analysis, we can conclude that none of the public, research, or private technological solutions to combat Gender-based Violence benefit from key current state-of-the-art and consumer electronics progress, such as physiological and physical analytics and affective computing. These advancements can be exploited towards a better, autonomous, and more inconspicuous technological Gender-based Violence preventing tool, which is the goal of the UC3M4Safety team by means of the Bindi system. Moreover, the design of such a tool towards women's safety requires them

to be co-creators of the solution, which this team is strongly considering by a close collaboration with different women's associations and focus groups of professional women experts in the field. Up to my knowledge, Bindi is the only system that proposes a technological tool to help prevent and combat Gender-based Violence by means of affective computing [9, 184, 221, 222].

5.2 Bindi

Chronologically, Bindi has gone through several phases of design and development, Figure 5-3. The first proof of concept was the *iGlove*, which was a co-supervised Master thesis [223]. The idea behind this system was to design and implement an initial continuous physiological wearable monitoring system. Specifically, it was equipped with three physiological sensors (BVP, GSR, and SKT) and allowed continuous data transmission using Bluetooth Low Energy (BLE) to a mobile phone. This device was based in [224]. Moreover, the integrated SoC within the *iGlove* was an ARM®Cortex-M0 32-bit with 32KB RAM and 256KB Flash. This device successfully fulfilled its goal to create the first tool to boost affective computing research within the UC3M4Safety team. Thereafter, the first formal version of Bindi, Bindi 1.0, was designed using most of the *iGlove* hardware as a solid starting point. As already introduced at the beginning of this Chapter, Bindi 1.0 is a personal-area-network system formed by three devices: a bracelet, a pendant, and a smartphone application. In particular, the integrated SoC within Bindi 1.0 was an ARM®Cortex-M4 32-bit with 64KB RAM and 256KB Flash. Specifically for this system, I was responsible for some of the main tasks related to the bracelet such as: 1) the supervision of the schematics, Printed Circuit Board (PCB), and layout design, 2) the firmware design and system integration, and 3) the coordination of the different validation and test-bench application to assure functionality. It should be highlighted that, regardless of these specific tasks, most of the design, integration, implementation, and validation were jointly carried out in an orderly and organised manner by a group of people belonging to the UC3M4Safety team. This first version of Bindi is the one addressed in this Chapter. Moreover, Bindi 1.0 is one of the sensory systems employed during the recording of the WEMAC dataset, which is explained in Chapter 6. Following the different limitations identified during the

development and usage of Bindi 1.0, the UC3M4Safety designed Bindi 2.0 during the last two years. This new system suffered a drastic miniaturisation process which leveraged the hardware integration of Bindi 1.0. Additionally, new sensors and different hardware improvements have been included in this new version, as well as new firmware functionalities. In this case, the integrated SoC within Bindi 2.0 was an ARM®Cortex-M4 32-bit with 256KB RAM and 1MB Flash. It should be noted that, as Bindi's technology has been improved, its computational needs have also increased, which has led to more storage capacity requirements in particular. Nevertheless, this fact has not led to a considerable increase in power consumption when comparing specifically Bindi 1.0 and Bindi 2.0 [9, 225].

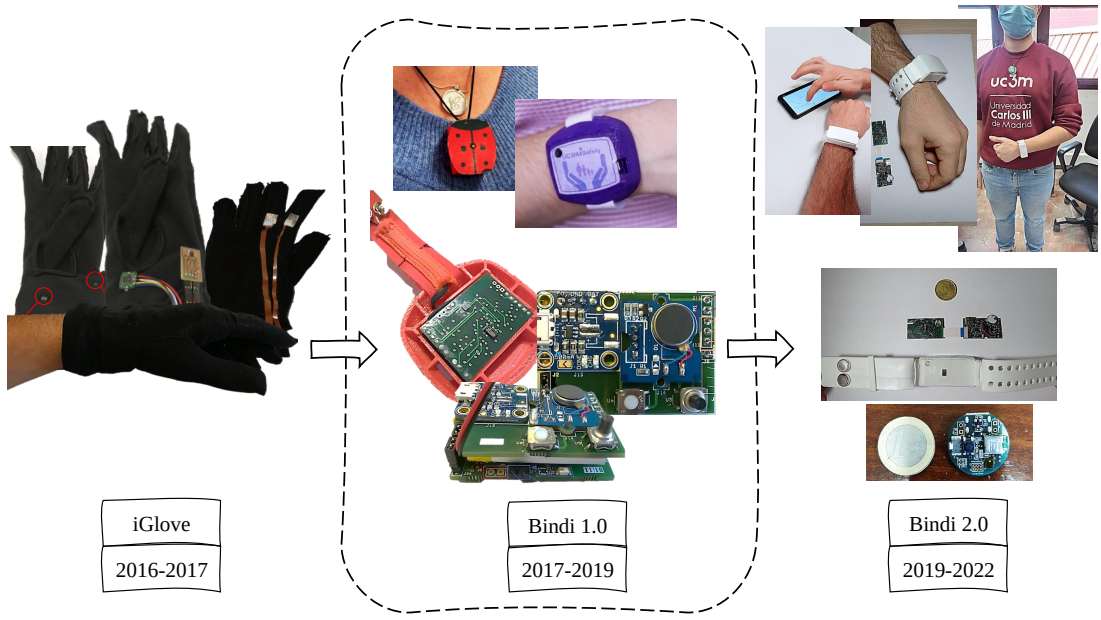


Figure 5-3: Bindi technology evolution since 2016 until 2022.

Once the state-of-the-art for technology applied towards combating Gender-based Violence and Bindi's technological context have been addressed, the following sections focus on providing an embedded perspective regarding the different digital signal processes, techniques and methods designed and implemented into Bindi's bracelet during the evolution of this research. First of all, a detailed analysis of the bracelet architecture, from both perspectives hardware and software, is presented. Secondly, different digital embedded filtering architectures are evaluated and analysed keeping a trade-off between resource requirements and physiological information preservation. Thirdly, a novel proposed SQA system for PPG signals, implemented and evaluated by using public and own datasets is detailed. This SQA

system also reports time and power consumption metrics for different extracted features. Afterwards, a complete embedded feature extraction design space exploration for a HRV use case is presented. Here, frequency and temporal data processing techniques are analysed and discussed. This is done to provide an in-depth perspective regarding the feature extraction design considerations and limitations. Moreover, a comparison between the HRV-based features obtained with the bracelet and the ones obtained with a research toolkit is reported and discussed. Finally, power consumption metrics are reported, which provides a comprehensive analysis regarding the battery lifetime of the Bracelet.

5.2.1 System architecture

As shown in Figure 5-4, the bracelet is made up of different hardware and software elements. These can be classified into four groups: the SoC, actuators, power management elements, and physiological sensors. They are described as following:

- Microprocessor Unit. Bindi 1.0 is equipped with the nRF52832 SoC that includes ARM® Cortex®-M4, an ultra-low power consumption microcontroller unit with 512KB memory flash and 64KB RAM, single-precision floating-point unit, Thumb®-2 instruction set, 64MHz clock, and some integrated peripherals (USB, UART, SPI, I2C, I2S, ADC, PDM, and AES) [226]. Note that the radio-frequency module through Bluetooth Low Energy® (BLE) communication is also integrated within this host unit. Moreover, the different employed digital signal processes were embedded into this SoC.
- Actuators. The Bracelet is equipped with a conventional electro-mechanical button for manual user activation, acting as the panic button. Additionally, a buzzer is also included to provide a physical response for the different alarms of the system [227].
- Power Management Elements. In this case, the BQ2019 and MCP73831 components by Texas Instruments® and Microchip® are used [228, 229]. These two integrated circuits are responsible for monitoring and charging the battery, respectively. For Bindi 1.0 a 500 mAh Lythium Ion Polymer Battery of 3.7V was employed.
- Physiological sensors. Three different physiological sensors are present in the Bracelet: PPG, GSR, and SKT. Specific details are provided in the following

Section, together with found limitations regarding the hardware implementation. Note that the latter were addressed in the following versions of Bindi (Bindi 2.0).

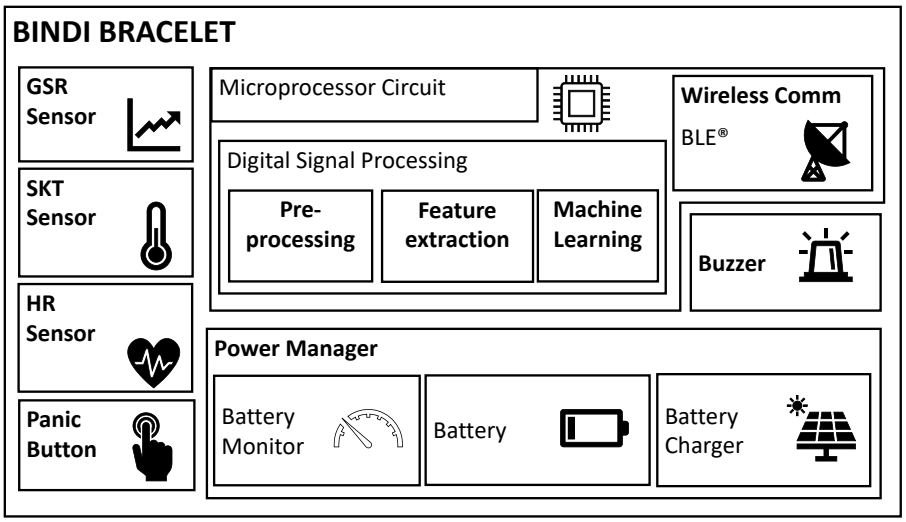


Figure 5-4: Simplified Bracelet architecture.

From a hardware perspective, most of the different elements within the Bracelet are based on commercially available smart sensors, microcontrollers, and actuators. This decision was based on three main facts: 1) easing all the design and integration processes, 2) creating the first wearable version of Bindi with commercially available parts whenever possible, and 3) reducing costs by not having to design many of the elements from scratch. Afterwards, this design decision allowed us to identify current drawbacks and limitations of the employed Commercial-Off-The-Shell (COTS). These will appear along the following subsections.

5.2.1.1 Physiological sensors design and integration

This Section provides an in-depth analysis of the sensors integrated within Bindi 1.0, as well as the limitations encountered during this process. Note that the body-locations of the sensors were directly affected by the factor form of the Bracelet, as well as by previous literature that tested physiological differences [102, 230, 231].

Heart-Rate sensor

The integrated heart-rate sensor is based on a photoplethysmographic sensor that detects BVP changes by measuring the absorption of light emitted through the skin, as studied in Chapter 2. This sensor is the MAX30101 High-Sensitivity reflective pulse-oximeter, with 18-bit ADC, I2C communication, digital noise cancellation, and

different integrated LEDs (red -660nm-, green -527nm-, and infrared -880nm-), [8]. Considering the quantum efficiency of the photodiode of the sensor, Figure 5-5, and the forward voltage required by the different LEDs, the red LED was finally selected. Note that the quantum efficiency of any photodiode or photodetector refers to the percentage or fraction of absorbed or incident photons that contribute to the actual photocurrent, i.e. the photodiode expected sensitivity divided by the maximum photosensitivity in case every photon generates an electron. Additionally, we decided to use just one of the LEDs due to reduce the power consumption and to open a new research line regarding PPG motion artefacts removal by means of blind source separation techniques. The latter resulted into a supervised Master Thesis [232], in which the foundations for the usage of motion artefact removal algorithms were established. Note that, although the latter is not within the scope of this document, it will serve as the basis to future research. Amongst the reviewed capabilities of this smart sensor, it also offers configurable sampling frequency from 50 Hz up to 3.24 kHz, and programmable LED current control. In our case, for the embedded implementations presented in this Chapter, we employed the maximum LED current (50mA with 411μs pulse-width) and 100 Hz sampling frequency. The former was decided to provide a deeper penetration, which derived into a stronger cycle difference between the systolic and diastolic phases. The sampling frequency was chosen as it is the one available in the sensor that allows proper temporal resolution to further extract the wanted features [233]. One of the main limitations of this sensor is actually its main advantage, as it provides an end-to-end solution by integration different LEDs, but this does not provide flexibility in testing other LED configurations. This led the UC3M4Safety team to investigate in different LED settings by modifying the skin-sensor air-gap or even testing multi-wavelength set-ups [234]. Regardless of such latter research, this sensor has been maintained for the integration of Bindi 2.0.

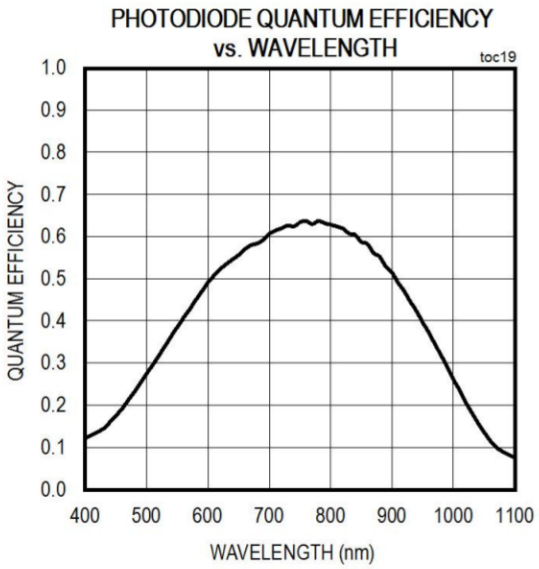


Figure 5-5: MAX30101 photodiode quantum efficiency [8]

GSR sensor

For the GSR sensor, a DC exosomatic measurement applying a constant voltage across the skin has been implemented. Moreover, dry steel electrodes were used. Note that, in this case, this is the only sensor, out of the three integrated in the Bracelet, that was designed rather than acquiring a COTS analog-front-end or smart-sensor. The design of this sensor was based upon the first circuit integrated within the *iGlove* [223]. Figure 5-6 shows the analog-front-end schematic for the current GSR sensor in Bindi 1.0. Specifically, the electrodes are connected to $J7$, by which a skin-potential measurement is performed. This is realised due to the voltage divider between the skin and $R14$. Towards the avoidance of endosomatic disturbances, a reference common to output and input is considered to make the voltage difference independent of the reference electrode position. Based on the output voltage of the sensor to be measured, a reference voltage is applied to avoid saturation by using a variable resistor ($R7$). Note that voltage followers are applied on both branches as buffers to avoid impedance related issues. Finally, a differential amplifier is employed to get the difference between the known voltage reference and the skin voltage divider. The amplification between these two voltages is given by equation 5.1:

$$V_{OUT3} = \frac{((R_{skin} - R7) * 2 * VCC_{1.8B} * 2 * 10^5)}{((R7 + 2 * 10^5) * (R_{skin} + 2 * 10^5))}. \tag{5.1}$$

This output voltage is followed by a low pass filter ($R11$ and $C11$) to avoid high-frequency noise with a cut-off frequency of up to 1.5 Hz. Note that the GSR information remains bellow such frequency, as studied in Chapter 2. Regarding power consumption, the sensor itself consumes around 0.7mA.

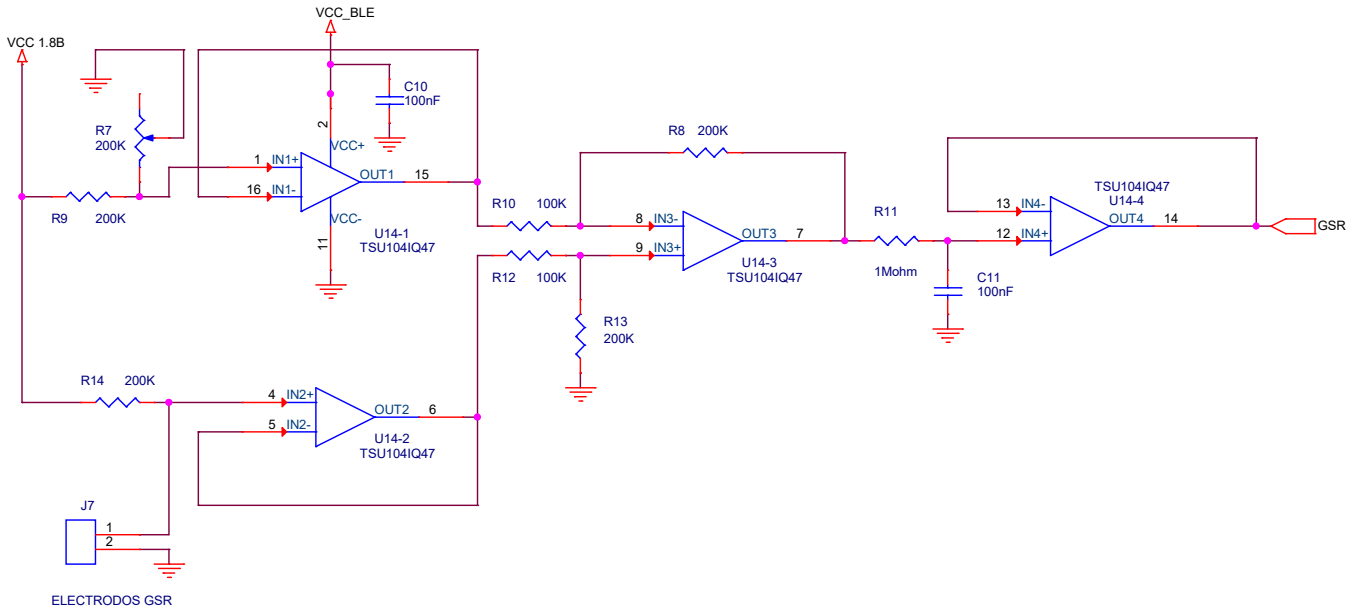


Figure 5-6: GSR analog-front-end implementation in Bindi's 1.0 bracelet.

One of the main limitations of the implemented GSR acquisition circuitry is the nonlinear behaviour. Figure 5-7 and 5-8 show the output voltage (V_{OUT3}) and the injected current into the skin, respectively. Note that the voltage is depicted using different $R7$ values, and $R14$ is fixed to 200 $k\Omega$. The latter was set to that value in order to limit the injected current bellow the recommended limits of $10\mu A/cm^2$ for safety requirements [120]. Following a trade-off between sensitivity and a desired range of up to 0-20 μS, we decided to fix the variable resistor to $50k\Omega$. Thus, considering a sampling resolution of 14 bits (ADC), the LSB is up to $219\mu V$, and the worst conductance resolution of the sensor is $0.007\mu S$. This resolution is enough to capture 0.01 μS changes to properly record all the SCRs. Note that, assuming a maximum quantization error of LSB/2, for this case that leads up to $\pm$ $0.003\mu S$.

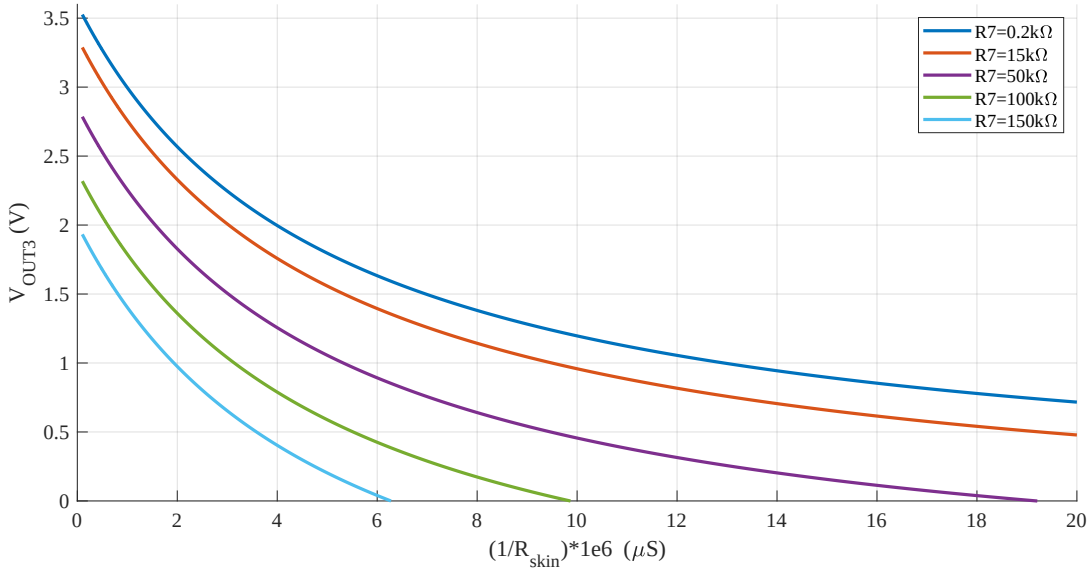


Figure 5-7: Bindi 1.0 GSR response considering different skin resistances.

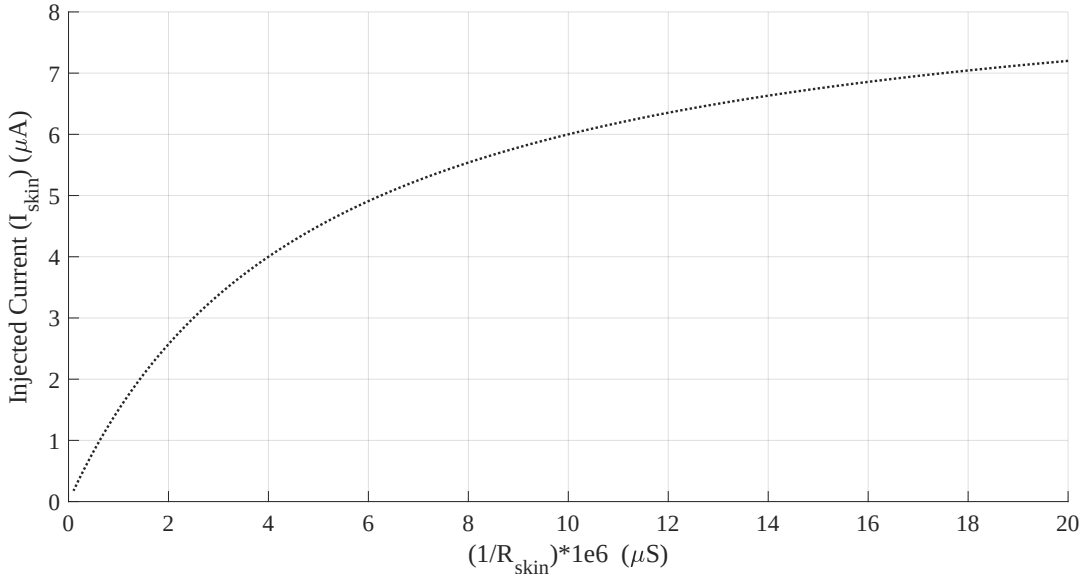


Figure 5-8: Non-linear response of the skin current given by voltage divider between $R14$ and R_{skin}.

This sensor was successfully empirically validated using passive components (resistances) in [223]. Moreover, due to the complexity towards the generation of a proper skin model [235], we decided to use a research-grade GSR sensor and a reduced set of volunteers to validate an actual GSR measurement [236]. In these experiments, our GSR sensor was placed on the distal forearm due to the Bracelet form factor, whereas the validation GSR sensor was located on the palm. The latter location is known to have the highest sweat gland density in the body [112], which implies a more affective-sensitive signal. Moreover, Bindi was working based on dry electrodes, whereas the validation sensor was using hydrogel electrodes, see Figure

2-15. This fact is key when comparing the signals, as the hydrogel improves the signal quality by lowering the impedance that exists at the electrode-skin interface. Figure 5-9 shows the normalized raw GSR signals obtained by both devices for a volunteer during two different trials. The vertical dash line in the figure marks the stimuli separation, where the first and second stimuli are joy and fear, respectively. Analysing the correlation for the signals acquired by both sensors, a Pearson metric of 0.85 is obtained, which denotes a strong direct positive correlation. Similar correlation coefficients were obtained for the rest of the volunteers. Differences between both signals are appreciated, which can be due to sensor motion artefacts, hydrogel effects, and sensor location. Moreover, most of the SCRs captured by the validation sensor are present in the signal of Bindi. Hence, we concluded that the validation of the sensor was successful.

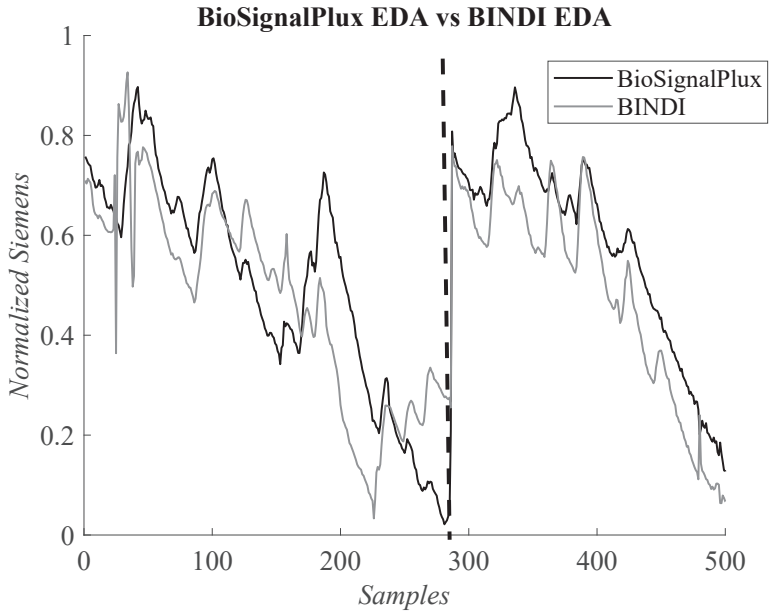


Figure 5-9: Normalized filtered GSR signals obtained by Bindi and the validation sensor for a volunteer in two stimuli. The dash vertical line denotes the stimulus separation.

The limitations found during the development and integration of this GSR sensor made the UC3M4Safety team work towards a new sensor that tackles the non-linear response and possesses adjustable and subject-independent hardware. This makes the new sensor system to be able to adjust its hardware based on the current baseline sensed or any other GSR-based individual parameter to assure the recommended sensitivity without exceeding the current density recommendation limits. This is currently being under testing and a publication is in progress [237].

SKT sensor

Finally, the MAX30205 component is proposed to acquire a reliable skin temperature measurement [238]. This integrated circuit is defined as a clinical-grade sensor for wearable applications, providing a $\pm 0.1\,^{\circ}\mathrm{C}$ accuracy over a $30\,^{\circ}\mathrm{C}$ to $50\,^{\circ}\mathrm{C}$ temperature range. It integrates I2C communication and a high resolution, sigma-delta, 16-bit ADC. Moreover, when being in active mode, it consumes around $0.6mA$.

As stated in Chapter 2, the skin temperature measurement is a robust indicator to characterise the homeostasis process of the body. Although using contact-temperature sensors is simple as long as the contact surface (skin) is available, acquiring accurate measurements of such variable is a challenging task due to the different setup variables and conditions. This is referred to considerations as the homogeneity of the skin, the thermal contact resistance, and the attachment effectiveness, amongst others [139]. Specifically, the MAX30205 measures the temperature of its own die by the thermal path between it and the PCB. Thus, the measured temperature is acquired throughout the leads and the exposed pad. Within this context, and considering the factor form of the Bracelet, we decided to integrate this sensor within the PCB, just right bellow the PPG sensor, Figure 5-10. Despite the fact that the manufacturer in the data-sheet states that temperature errors due to self-heating are low because of the minimal low supply current, it is also specified that a sampling period $\geq$ 10-seconds is required to avoid such effects completely. Thus, the measurement principle of the sensor together with the PCB implementation did not result in the most accurate to acquire the skin body temperature nor the most efficient way to avoid self-heating, thermal mass, and/or thermal conductivity problems. The consequence of this problem was an initial thermal gradient that lasts around 200-seconds until the thermal-mass of the PCB is at equilibrium. For instance, Figure 5-11 shows the filtered output of the sensor after placing a finger on top of the integrated chip under controlled room temperature conditions. This problem was solved for the following versions of Bindi (Bindi 2.0), as well as for the experiments carried out and explained in Chapter 6, by considering the integration of the MAX30208 temperature sensor [239]. This sensor was the next version of the MAX30205, including the same digital capabilities, but changing the measurement principal and the power consumption in operating mode. Specifically,

it measures throughout the top package contact instead of using a thermal pad, and it consumes around $70\mu A$ when acquiring. Note that the power consumption is considerably lower in comparison with the previous sensor. Figure 5-12 shows the modification performed to the Bracelet to include the new temperature sensor and an experiment comparison for both of them. Note that we used part of the evaluation board of the MAX30208 [240]. The performed experiment consisted into three phases: 1) the sensors were left outside for 1 hour (November, 14ºC), 2) the system was switch-on and started measuring right after getting into the room, 3) skin contact was performed for both sensors after being three minutes measuring at room temperature, and 4) skin contact was released after one minute. Thus, we can observe how the MAX30208 response is faster than the MAX30205, two times faster specifically, and how the measurement principle and thermal mass of the PCB are affecting towards reaching an accurate measurement. Moreover, there can be also observed an offset between both of them, which is also due to the commented factors. Notwithstanding the encountered problems with the integration of the MAX30205 and although there is an offset with respect to the MAX30208, the measurements obtained from the former are valid once the initial PCB thermal transient is over.

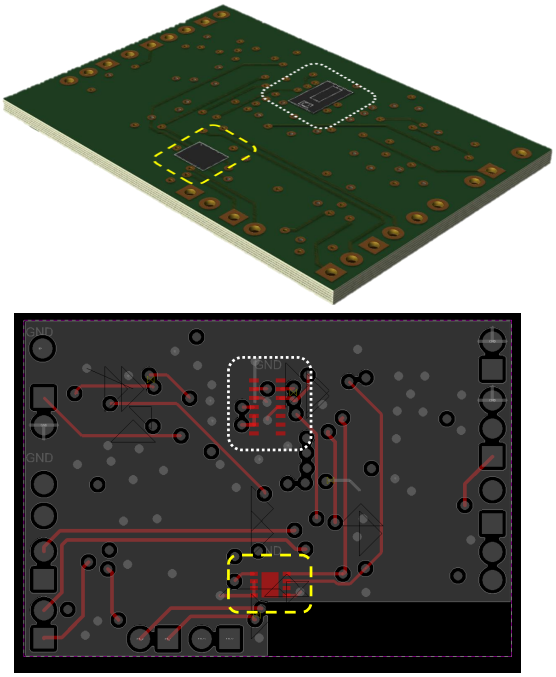


Figure 5-10: Skin temperature (yellow/bellow circle) and heart-rate sensors layouts integration into the Bracelet. The grey area determines the ground plane.

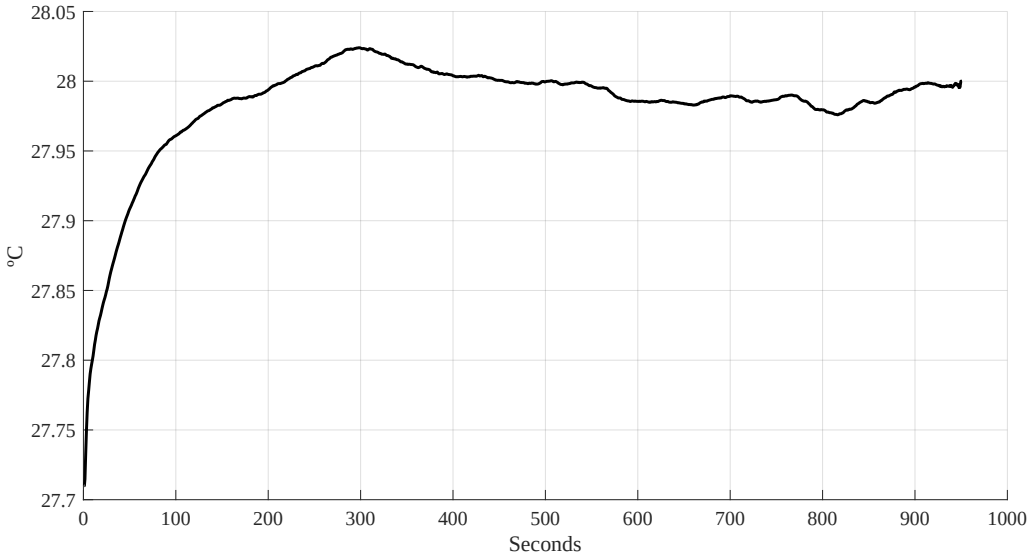


Figure 5-11: MAX30205 filtered output after placing a finger on top of the integrated chip under controlled room temperature conditions.

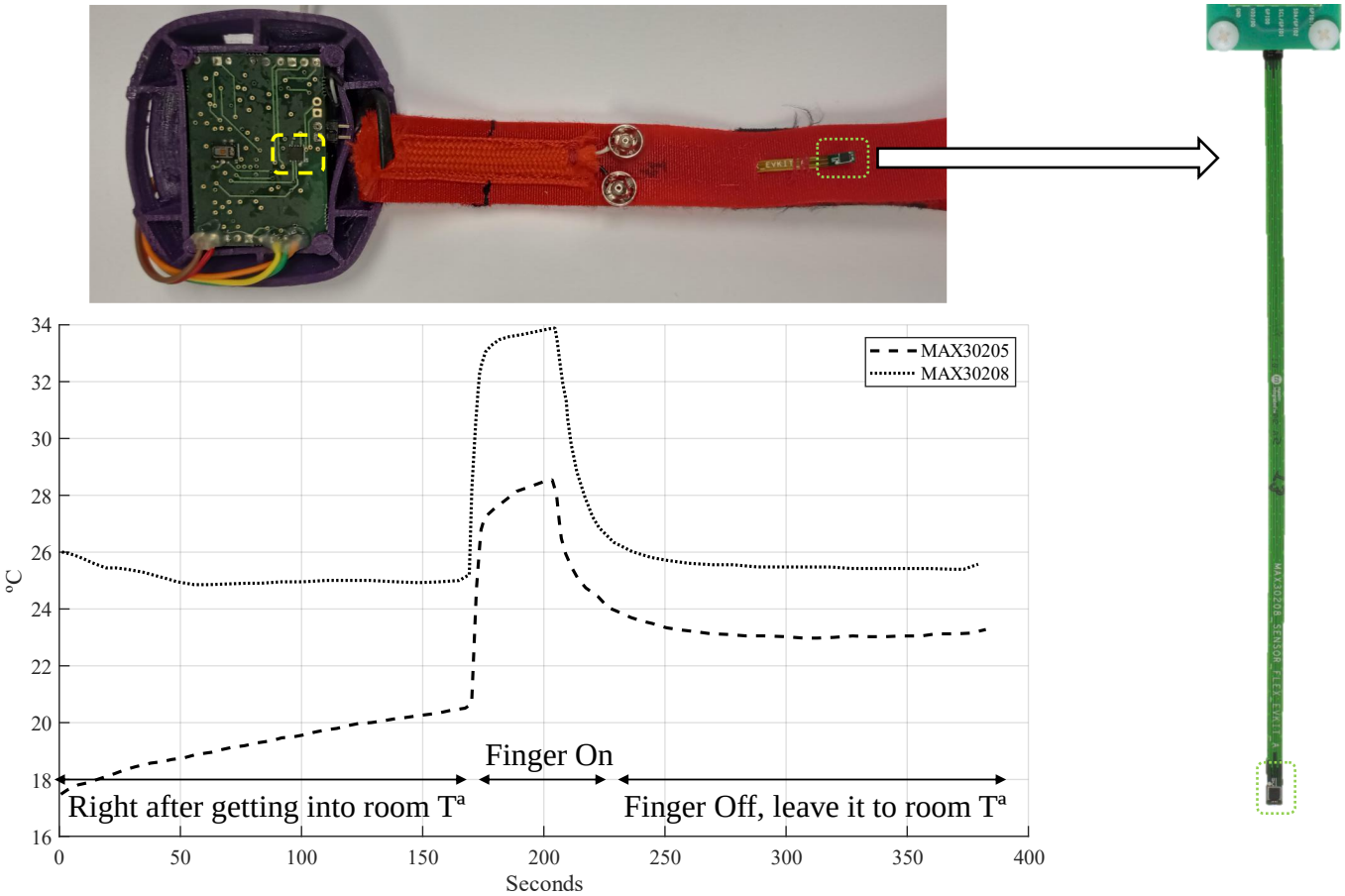


Figure 5-12: Modification performed to the Bracelet to include the MAX30208 and experiment comparison for both of the temperature sensors. On the right is part of the evaluation board of the MAX30208.

5.2.1.2 Digital signal processing design

The designed firmware for Bindi took advantage of the functionalities or Software-Development-Kit (SDK) provided by the microcontroller manufacturer, in this case Nordic Semiconductors®. Figure 5-13 depicts a simplified structure for the stack embedded into the Bracelet. Each part is described as follows:

- nRF HAL. This is part of the Nordic SDK. Specifically, it is the Hardware Abstraction Layer (HAL) for the different low-level functionalities of the system including direct interface with the ARM® core, peripherals, and radio, among others.
- BINDI BLE. This is an ad-hoc BLE manager system that handles the different radio transmission and reception queues, performs the formatting of Bindi-BLE packets, and manages the direct interaction with the softdevice. Note that the latter is the BLE stack being employed, which in Bindi 1.0 is the S132 [241] that builds upon BLE 5.1 qualified.
- SYSTEM INIT. This is the part responsible for managing all the initialisation processes regarding the set-up request for the required peripherals, as well as for general GPIO initial configuration.
- BINDI HAL. This is one of the main parts of the stack. It is an ad-hoc peripheral-level HAL, which is specifically intended to manage all the different Bindi-related interactions with the peripherals, performs the raw acquisition management, carries out the first initial filtering stages, and proceeds to segment the data and store the processed buffers to further being processed by the BINDI APP layer. Moreover, it also deals with the actuators interaction, i.e. switch on and off the vibrator motor and receiving the panic button interruptions.
- BINDI APP. This layer is in charge of the main system-level functionalities such as physiological processed data management, feature extraction, main digital signal processing (DSP), and classification.
- User Application. The previous system-level functionalities are handled and synchronised by a finite state machine (FSM) that resides in this layer and is modified accordingly to the specific user application.
- CMS Task Handler. This is a cross-functionality that can interact with the

whole stack. It is mostly used to decode all the received (BLE) packets and trigger the respective required action regarding specific parts of the stack. This tool is also used to debug when being in developing mode.

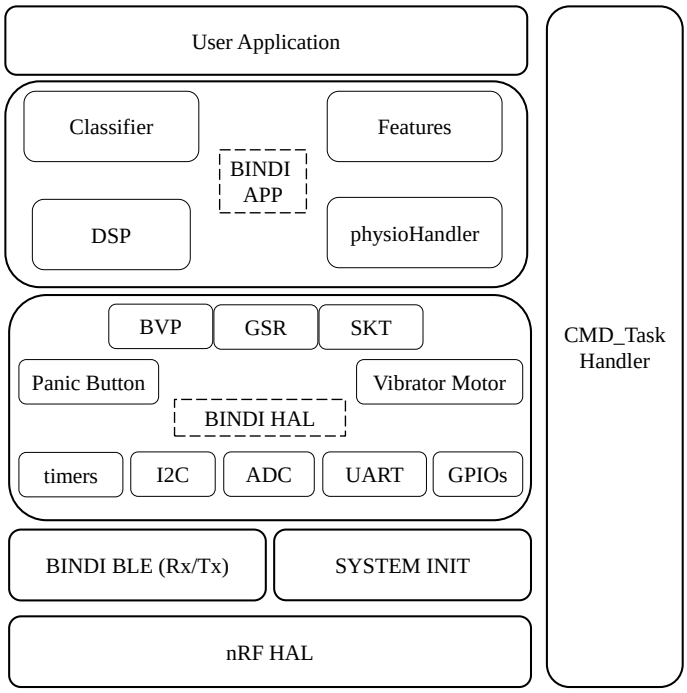


Figure 5-13: Current firmware stack of the Bracelet of Bindi.

Considering the data processing chains and following the data segmentation adopted when dealing with the last proposed fear detection system in Section 4.2, Figure 5-14 shows the different timing processes being performed within the Bracelet. Towards reducing the host operations and internal peripheral usage, rather than employing independent timers for each of the physiological signals, we make use of the timings provided by the PPG smart-sensor. This is done as this sensor is the one having the highest sampling frequency, $100Hz$, whereas the GSR and the SKT working at $10Hz$ and $5Hz$ respectively. Thus, every time that a new sample from the PPG sensor is written to the BVP buffer, we check whether it is time to sample the rest of the sensor in a synchronised manner. This physiological acquisition schema is repeated every second and allows to avoid any complex timing or temporal drift calculation. Moreover, the acquired samples for every signal are evenly separated. Note that the latter is crucial to properly apply different DSP processes such as FFTs. The data being acquired is filtered and stored in 20 seconds buffers, which are further fed to the feature extraction and classification modules. Figure 5-14 also shows the overlapping process outline. One of the main limitations of this acquisition schema is

the fully dependence on the PPG smart-sensor, since, in the event of sensor failure, the entire system is compromised. Different works are currently being performed to implement and provide a flexible measuring schema able to deal with malfunctioning events. Additionally, research regarding the integration of embedding online testing within the Bracelet to assess such cases is being in progress [242].

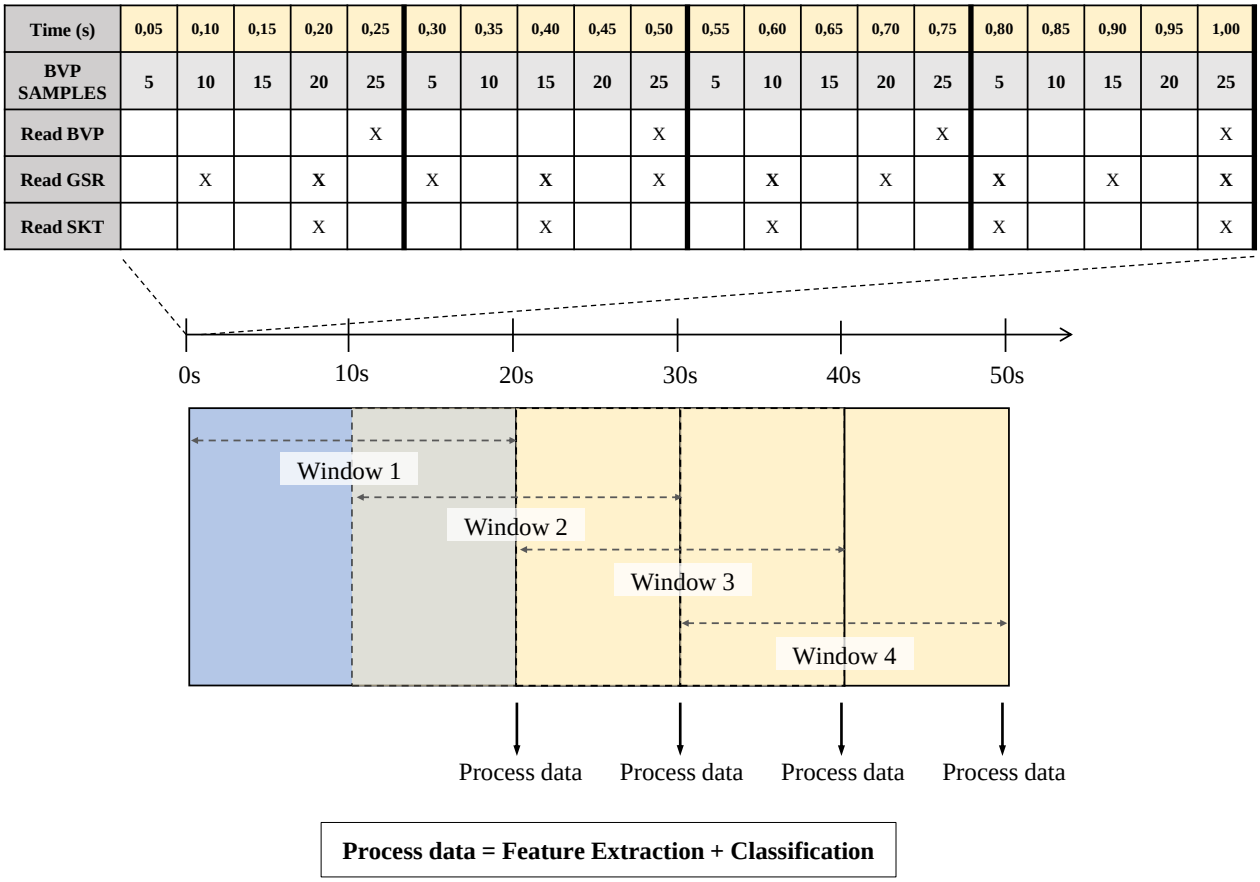


Figure 5-14: Current physiological synchronisation and data processing timings in the Bracelet.

In case of specifying each of the different digital embedded processes being done within every temporal window (20 second) and for every sensor, Figure 5-15 shows part of the current system architecture of the Bracelet focusing on the data-flow throughout such main processes. As already stated, every data processing chain starts by gathering the respective sensor data using I2C or ADC acquisition. After that, the data is filtered and segmentation (windowing) takes place. Some of the evaluated and implemented embedded filtering architectures are explained in Section 5.2.2. At this point, the sensors follow different paths. For instance, the current implementation regarding the BVP data is subjected to a quality assessment process

employing a SQA system, which is detailed in Section 5.2.3. At this point, different motion artefact removal algorithms are applied to recover most of the signal information if needed. Note that such algorithms are currently being developed and, although it is depicted in this architecture, it is not fully implemented yet. Thereafter, features are extracted from filtered and segmented data. Focusing on the PPG data processing chain, Section 5.2.4 details some of the processes involved during the feature extraction for BVP related metrics. Finally, the obtained features are fed to the inference engine and the resultant label is wirelesses transmitted to the Bindi APP. It should be noted here that, although the following Sections provide different in-depth analysis regarding some of these digital processes, the embedded implementation of the whole data processing chains, including the inference block, is a work currently in progress. For instance, in [184], we proposed a fully embedded data processing chain, from acquisition to embedded classification, by considering the average value of each variable for a temporary window of 10 seconds as both filtering and feature extraction stages. We implemented a lightweight KNN and applied cost-sensitive learning to train and deployed a subject-dependent system. That system was an initial embedded proof-of-concept and served as a base building block to start designing and improving the following version. For this reason, the discussion of any embedded classifier integration is out of the scope of this document and will be the subject of research arising from this work. Likewise, the motion artefact removal embedded integration is also left out of the scope of this research.

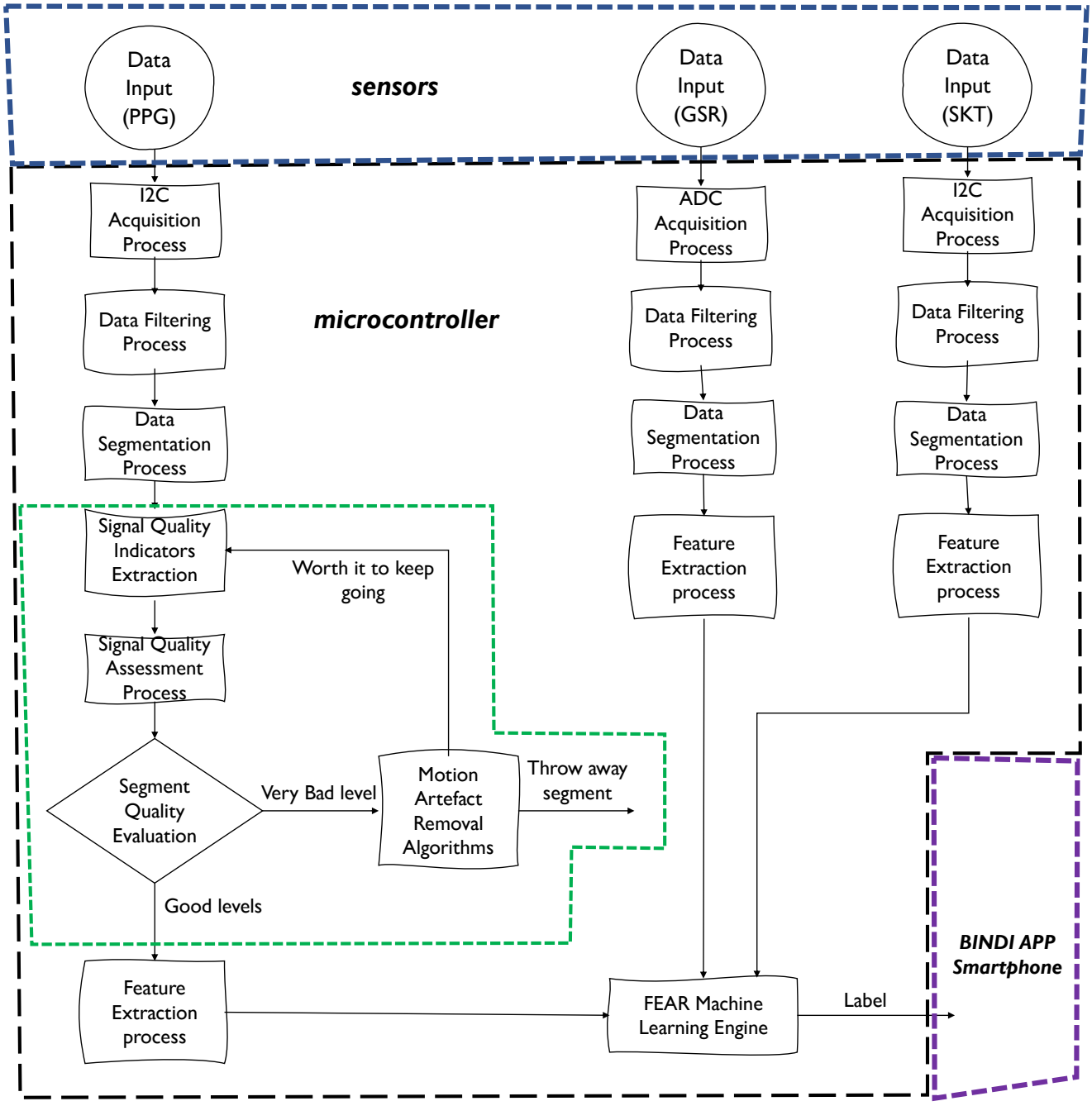


Figure 5-15: Current system architecture for the main digital processing tasks of the Bracelet.

5.2.2 Embedded Filtering Evaluation

In this Section, an embedded filtering evaluation for a PPG-based filtering stage is performed. This analysis is extracted from [159]. Considering the constrained resources of Bindi, five crucial parameters have been assessed for the different filtering architectures considered: memory usage, window computation time, settling time, stop-band mean attenuation, and bandpass ripple. The two first parameters are the ones directly related to the embedded filter implementation. The computation time is also constrained by the timing defined by the application. The rest of the parameters are related to filtering characteristics. For instance, settling time is especially relevant denoting the filter stabilisation time, which could be linked with waste in time and memory. The stop-band mean attenuation is related to the mean attenuation level with the designated rejection band, while band-pass ripple is the amount of variation in the gain within the designated bandwidth of the filter.

From an embedded or a digital perspective, as already reviewed in Chapter 4, there are two commonly applied filtering techniques: IIRs and FIRs. IIRs are computationally fast, although they do not have a linear phase response, which could lead to not preserving the wave-shape or physiological morphology. For instance, this fact can result in wrong delineated points to be identified by the BVP peak detection algorithms. Such disadvantage is alleviated by using a forward-backwards IIR filtering technique, which requires double filtering and double time-reversal of the signal. This latter technique leads up to a high computational time at the expense to obtain a zero-phase transfer function. On the contrary, FIR filters can be designed to have a linear phase response, so preserving the physiological morphology and not affecting possible patterns. However, they require more coefficients and memory than IIRs. These and other digital techniques are used to deal with out-of-band noises, such as baseline wander and high-frequency noise. In case of a BVP signal, the rejection of these noises is key to properly minimise the changes in their morphology that does not have a cardiac origin.

The four filter design options considered are: three band-pass FIR filters with different orders and a two-stage filter based on moving averaging. The design of such band-pass filters was conducted by Matlab® according to the equiripple method. On the one hand, the resulting coefficients were quantified to a 14-bit integer to reduce

memory usage and boost processing time. This number of bits is the maximum precision that ensures no overflow in our system, with 18 *bit/sample* BVP signals and 32-bit registers. The frequency response impact is minimal and the root means square deviation of the output compared with 64-bit floating-point coefficients is negligible. On the other hand, the two-stage filter is composed of two moving averaging steps. The first one is a low pass 4-sample filter, while the second one is the signal subtraction of the 100 values, centred moving average.

Table 5.1 shows the results obtained for the evaluated embedded filtering architectures. Analysing this table, we can observe that, for the band-pass filters, increasing the filter order (the number of coefficients) increases mean stop-band attenuation but also ROM memory usage, computation time, and settling time. The computation time for these band-pass filters could be reduced if coefficients are stored in RAM at the expense of memory usage. Note that the computation time for the two-stage filter is significantly lower than for the band-pass filters.

Regarding stop-band attenuation, this parameter benefits band-pass filters, providing a higher attenuation across all the stop-band. Note that the most efficient stop-band attenuation vs. ROM memory usage relationship is reached by the 400-coefficient band-pass filter, because of the constant code size effect. Focusing on the band-pass ripple, a low value is desired to avoid deformation of the signal. In the case of the proposed filtering architectures, the band-pass ripple of the filters do not provoke any distortion in the signal. Regarding settling time, the difference between two-stage and band-pass filters is large, benefiting the latter.

Table 5.1: Results obtained for the evaluated embedded filtering architectures.

Desing options	Compt. Time [ms]	RAM [bytes]	ROM [bytes]	Set. Time [samples]	Mean stopband att. [dB]	bandpass ripple [dB]
400-coef	0.2474	10	626	400	-38.8	0.09
200-coef	0.1240	10	426	200	-25	0.64
100-coef	0.0623	10	326	100	-14.9	3.09
2-stage	0.0048	20	470	4	-9.3	1.93

Overall, from all this analysis, the two-stage filter is recommended. This posses a good trade-off between computation time, attenuation, and memory usage for a wearable constrained system as Bindi. Apart from such design decisions supported by those metrics, it should be noted that from a physiological point of view, the number of coefficients associated with the settling time can negatively affect physiological monitoring. This fact is motivated by the number of samples that need to be

removed for every considered architecture, which affects negatively the final amount of physiological information from which to extract the different features. Note that, although we focus solely in this signal, some of the extracted conclusions can be extrapolated for the other two filtering stages to be addressed within the system (GSR and SKT).

5.2.3 Signal Quality Assessment

SQA is a key process for continuous and reliable physiological monitoring [243]. Specifically, this type of processes strongly benefit Bindi as they are focused on assessing the quality of the signal using different features extracted from it and decision rule. Thus, these systems provide a quality measurement of the segmented signal being processed. Note that this system does not deal with any motion artefact removal task or similar. This signal quality output can be further used by the different feature extraction algorithms or even by the fear machine learning to properly adjust or weight the quality of such temporal instance. Regarding its different stages, it is formed by up to three main processes:

- The first one is the feature or Signal Quality Indicator (SQI) extraction stage. Different SQIs are extracted from the segment of the signal to properly characterise it. Note that appropriate features or SQIs are those that change between clean and noisy segments of the signal.
- Following the previous process, the extracted features are evaluated based on different decision rules to quantify the noise level.
- The output of the latter stage is the signal quality index (SQi), which is binary-based in most of the cases. When using different sources of the same signal or even different signals, a third data fusion stage is performed. In such stage, individual SQis are combined to give the final quality metric.

Note that the filtering process and the data segmentation of the signal are not within the scope of the tasks of the SQA; however, the signal needs to be filtered and segmented before the SQA application. As for the previous embedded filtering evaluation, the presented SQA system, [214], is also focused on PPG signals due to their wearable relevance and importance within Bindi. It should be highlighted that the work presented in this Section is the result of an international collaboration with the University of Essex [214].

In the literature, most of the PPG SQA proposed embedding low-resource methods share the following characteristics:

- They are based on hard thresholded decision rules to assess the SQi. This methodology obviates the high uncertainty as a result of inter-subject differences or intra-subject ones, such as variable noise levels across time.
- They consider a high amount of training or threshold-adjustment data using a combination of different datasets. However, the actual number of public datasets containing signal quality annotations is scarce, which forces the researchers to label the used data.
- The proposed systems are tailored to the specifically labelled dataset, which results into an experiment-dependent system that hinders achieving enough generalisation to cope with different experimental settings.

Being aware that the generation of annotated datasets is a challenging task, a few-shot consideration validation or adjustment together with a posterior online self-tuning might be exploited towards the design of heterogeneous systems that can deal with the low amount of annotated data available. Note that such type of design perspective can be also applied to systems that are expected to be trained or adjusted based on into-the-wild data and daily volunteer annotations, as in such experiments the gathered annotations are expected to be sparse. Moreover, previous research that performed SQA embedded implementation and presented different trade-offs to consider at design is scarce. On this basis, in this Section, a novel embedded subject-invariant SQA system using a reduced set of features combined with an interval fuzzy rule-based system (FRBS) is presented. This system is the current SQA runing into the Bracelet. Specifically, to deal with the SQA generalisation and tailoring coming from the PPG signal wide casuistry, a type-2 fuzzy system is implemented, as it provides a better uncertainty framework for harnessing uncertainty. Moreover, an adaptive fine-tuning stage is also proposed and applied to self-adjust the FRBS in an online manner, which provides an agnostic user adaptation.

Focusing into the SQI extraction stage for PPG sensors, there is a clear division between time-domain and frequency-domain methodologies. The former represents the most common techniques used in the PPG-SQA systems in the literature. For instance, the statistical behaviour of different trend-based SQIs were studied in [244].

Specifically, seven indicators were tested (perfusion, Kurtosis, skewness, relative power, signal-to-noise ratio, zero crossings and entropy) using 160 recordings of 60 seconds each, a total of 9600 seconds. In the results presented, skewness outperformed the other SQIs by achieving an F1-score up to 87.20% on detecting acceptable and unfit pulses. This publication defined three different levels of quality rather than the usual binary classification. Regardless of the advantage given by the low computational complexity of these trend-based SQIs, designing an SQA purely and solely based on these metrics is exposed to the heuristic decision rules with hard-thresholds. Regarding SQA systems based on frequency-domain feature extraction, Krishnan et al. in [245] used the spectrum of the signal skewness (bi-spectrum) to exploit the phase relations that exist in a clean PPG signal. These methods imply high computational effort in comparison to some others time-domain based which do not require performing either Fast Fourier Transform (FFT) algorithms or any basis transformation. Moreover, the development of deep and machine learning algorithms led to classification systems that automatically detect the different anomalies within the PPG signal in a more robust way [246]. However, they are not free from empirically determined decision rules, and the deep learning approach hinders an optimal embedded implementation.

Focusing on the SQA systems proposed in the literature that went embedded, we can highlight three recent works. In [247], Vadrevu et al. proposed one of the first real-time PPG SQA systems by extracting time-domain features. They applied six heuristic predefined rules to assess the quality of the signal, and used a 32-bit ARM Cortex-M3 micro-controller. They combined two different public benchmark PPG databases with their own dataset. Such data combination was used for both threshold adjustment and performance validation. Finally, they achieved up to 95.93% for overall accuracy. Although they showed competitive power consumption data regarding the effect of data retention decreasing and SQA embedded implementation, their system was still subjected to the empirical threshold adjustment. This fact tailored the proposed system to those specific set of estimated thresholds. Moreover, they did not perform any blind testing. Similarly, in [248], Reddy et al. proposed the use of time-domain features with a set of empirical rules and thresholds. They also combined different public benchmark PPG databases, but divided them into

two datasets. One of those was used for threshold adjustment and the other for testing. They implemented the system into the same micro-controller as Vadrevu et al. and achieved up to 93.21% overall accuracy. Finally, in [249], Samiul Alam et al. employed Kurtosis and auto-correlation function with empirical thresholds as well. They followed the same dataset arrangement as Reddy et al., and achieved up to 96.50% overall accuracy. Besides being affected by the same commented empirical tailoring consideration, they used a high performance embedded platform (Quad-core ARM Cortex-A53). The latter hinders the comparison task with an extreme edge-computing context dealing with wearable devices. Amongst the commented advantages and disadvantages of these systems, two factors should be highlighted. First, the complete set of features used in these works was domain-specific, which requires some prior knowledge of the nature of the type of noise to be detected. Second, all the proposed systems were adjusted or trained using either the same dataset or part of a combination of different datasets. The latter fact is specially relevant due to the heterogeneous challenge previously detailed, as achieving a SQA system applicable for a wide range of real-life situations and activities requires to not only considering different volunteers but also performing blind testing with different databases.

After having reviewed the SQA systems for PPG monitoring, we can conclude that there is not a general common set of techniques to deal with this problem, but different domain methodologies and even combination of those. Moreover, regardless of the nature of such feature extraction techniques or classification algorithms, the systems presented in the literature fall back on hard-thresholded approaches. This produces the system to be tailored to the training dataset due to those heuristic decisions. When looking for other types of SQAs trying to overcome such limitations and dealing with generalisation, some research is found applying a type I Fuzzy Logic System (FLS) [250]. However, physiological SQA becomes challenging when having heterogeneous settings. Thus, type I FLS is limited on the amount of uncertainty it can cope with. For this reason, and leading up to heterogeneous SQA application, a reduced set of domain-specific and domain-agnostic features with an interval type II FLS, specifically a Fuzzy Rule Based Classifier (FRBC) is exploited in this research. Note that the type II technique is specifically intended to deal with the commented

uncertainty as each level of the features is fuzzified based on an interval fuzzy set called Footprint Of Uncertainty (FOU) [251].

5.2.3.1 SQA Design, Training and Validation

Figure 5-16 shows the SQA training architecture used in this research. Specifically, this architecture is composed of seven different processes. The following sub-sections provide a technical overview regarding each of the stages within this architecture.

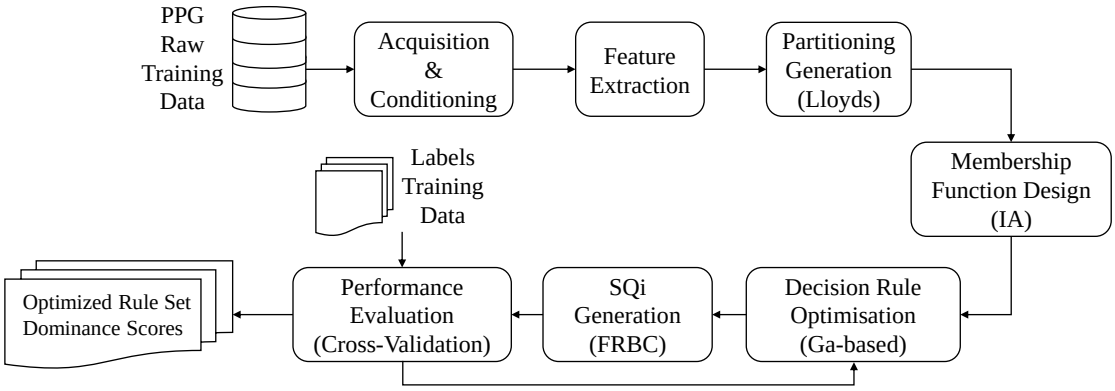


Figure 5-16: SQA training architecture proposed.

Acquisition and Conditioning

Once the signal is filtered following the previous selected embedded filtering architecture, segmentation takes place. This process is based on the fact that performing the feature extraction in small chunks of data will alleviate the different statistical processes to be done (e.g., mean or standard deviation time complexity calculations are based on the amount of data or samples, that is $\mathcal{O}(n)$). In our case, for the proposed SQA, the length of the segmented window is set to 3 s. This specific duration can provide two Heart Rate (HR) periods for a minimum of 40 beats-per-minute (BPM). Moreover, within this short period of time, we can even consider a quasi-stationary behaviour of this physiological signal. Note that as we decrease the processing window, the resource usage within an embedded system is also decreased, but the minimum BPMs at which we can assure two periods of the signal increases. This fact leads to a trade-off decision that in our case is driven by the commented physiological facts and previous works that used the same or similar temporal window lengths [249].

Feature extraction

Afterwards, different feature extraction techniques are applied to characterise the current window processing. Specifically, four features are extracted. Note that

all the implemented features are time-based. This decision is due to their lower computational complexity and to their robust performance proven in recent publications [244]. The different features extracted are detailed as following:

- Kurtosis. This is the statistical metric related to the shape of a probability distribution by measuring degree of concentration presented around the mean of the frequency distribution for a real-value variable. It is also described as the measure of the tailedness. This higher-order statistic measurement is given by equation 4.18.
- Entropy. Shannon Entropy provides a quantitative measurement with respect to the uncertainty or randomness of the signal. This feature is defined as:

$$e = -\sum_{i=1}^{N} (x_i^2) log(x_i^2), \tag{5.2}$$

  where N is the sample size, and x_i is each of the filtered data samples.
- Signal-to-noise-ratio (SNR). This is one of the most common features used in SQA systems. It compares the power of a desired signal with respect to observed noise. In this case, the following computation is performed:

$$snr = \frac{\sigma_{abs(x)}}{\sigma_x}, \tag{5.3}$$

  where $\sigma_{abs(x)}$ is the standard deviation of the absolute value of the signal, while σ_x is the standard deviation of the signal.
- Matrix Profile. Up to my knowledge, this feature has not been used for any PPG SQA system in the literature, although it is extensively used in time series anomaly detection [252, 253]. This metric offers different advantages that can provide a robust and reliable SQI, such as domain agnosticism, deterministic time, and parameter free. The working equation for the matrix profile is based on the distance profile given by the Z-score normalised euclidean distances of different sub-sequences within the time series:

$$d_{i,j} = \sqrt{2m(1 - \frac{Q_{i,j} - \mu_i \mu_j}{m\sigma_i \sigma_j})}, \tag{5.4}$$

  where $Q_{i,j}$ is the dot product of the two sub-sequences with length m ($T_{i,m}$

and $T_{j,m}$) of the time series, and μ and σ are the mean and standard deviation of the respective sub-sequence. Note that for this research work, the mean over the set of values stored in $d_{i,j}$ is calculated and assigned to every 3-second processing window. Regarding the specific algorithm, we used SCRIMP++, which offers the lowest time complexity amongst the different possible implementations [254].

After extracting the complete set of features for all the different subjects, an automatic gain controller (AGC) is applied to limit the amplitude and scale the extracted information. In this case, we used a $0 - 10$ AGC.

Quantization and partitioning generation

Following a fully data-driven approach, this research work uses data quantization and partitioning over the considered training data to generate the different fuzzy sets in an unsupervised manner. Thus, these processes are essential to assess the limits of the defined conceptual linguistic representations for every feature and properly modelled the different membership functions. This is done to assess if there exist a partition or separation of the feature values based on their distribution.

Specifically, in this case, we applied the cyclic Lloyd's algorithm [255] to optimise the different partitions using the reviewed features and targeting the extraction of three linguistic variables: Low (L), Medium (M), and High (H). The LLoyd's algorithm is executed in an iterative process for each incoming sequence or feature, $A_1, A_2, A_3, ..., A_m$, addressing a minimal mean square distortion or mean square error for the generated partitions $B_1, B_2, B_3, ..., B_m$. Due to the fuzzy logic system to be applied within the proposed system, the output of this stage must be the quantized partitions or intervals for each incoming sequence or feature. Note that, as we deal with the design of a subject-invariant SQA system, the quantization and partitioning optimisation is applied independently for each subject, which gives a set of m individual partitions or intervals (ν_i) with $m - 1$ endpoints (τ). This is outlined in Figure 5-17. The generated partitions are afterwards considered by the next stage of the proposed SQA training architecture to design the different membership functions to be implemented.

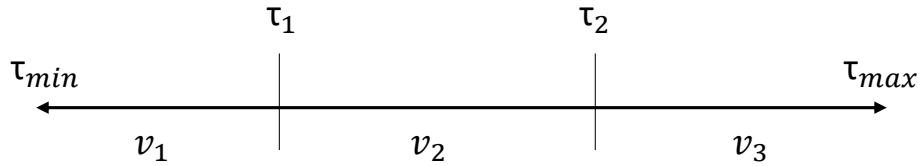


Figure 5-17: Interval representation with three (m) partitions (ν) and two found endpoints (τ). The values τ_{min} and τ_{max} are the min and max of the incoming sequence being evaluated or left and right endpoints.

Type-II membership functions design

From the generated intervals or partitions and endpoints, we used the Interval Approach (IA) methodology [256] to design the initial membership functions as well as the FOUs. This technique is applied over the complete set of feature's partitions $[\nu_1, \nu_m]$ and is based on two different processes. First of all, a preprocessing step for the complete set of intervals is applied. This step is based on four stages. The first stage applies a saturation check just to assure that every feature is in range. After that, the next two steps deal with outliers detection. On the one hand, a Box and Whisker test is used to remove possible outliers out of certain inter-quartile limit criteria, while on the other hand, a tolerance limit processing is applied to check that every point is contained within a specific range with respect to the mean and standard deviations of left, right endpoints and intervals. For the last stage, a reasonable-interval processing is performed. This is based on specific definitions or requirements that the different intervals must fulfilled. For instance, a non-valid or non-reasonable interval is that not overlap with another data interval. Finally, after the preprocessing stages, the second step of the IA technique is carried out. It comprises different stages as well, from mapping an interval to an initial type I membership function to computing a mathematical model for the final proposed FOUs. For instance, Figure 5-18 shows the final output for the membership functions and FOUs generated using the matrix profile feature training data.

Formally, every type II fuzzy set or linguistic concept is defined by a membership function that is given as following equation 5.5:

$$\widetilde{A} = \{(x, u, f_x(u)) | \forall x \in X, \forall u \in [\underline{\mu}_{\widetilde{A}}(x), \overline{\mu}_{\widetilde{A}}(x)] \subseteq [0, 1]\}, \tag{5.5}$$

where x is the universe of discourse contained within X, u is the primary membership value, $f_x(u)$ is the secondary membership value, and $\mu_{\widetilde{A}}$ represents the respective

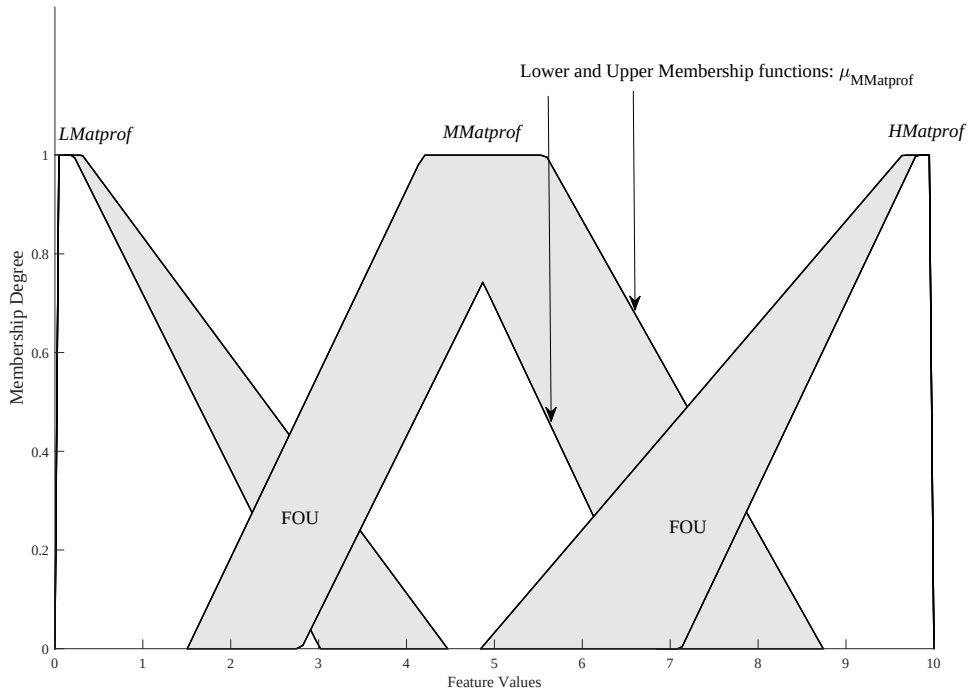


Figure 5-18: Type II membership functions generated from matrix profile feature data applying IA for all training subjects. Three linguistic variables: Low (L), Medium (M), High (H). Grey shaded area is the obtained FOU.

lower ($\underline{\mu}_{\widetilde{A}}(x)$) and upper ($\bar{\mu}_{\widetilde{A}}(x)$) membership degree functions of linguistic concept $\widetilde{A}$. Specifically for an interval type II fuzzy system, $f_x(u)$ is simplified as:

$$f_x(u) = 1, \forall x \in X, \forall u \in [\underline{\mu}_{\widetilde{A}}(x), \overline{\mu}_{\widetilde{A}}(x)] \subseteq [0,1]\}. \tag{5.6}$$

Thus, given a discrete universe of discourse and regardless of the memberships shape, eighth points for every fuzzy set are stored, i.e. four x points per membership functions (lower and upper) delimiting such universe. Note that in our case, due to the previous feature scaling factor, $X \subseteq [0, 10]$.

Decision rule optimisation

One of the ultimate goals of this training architecture is to generate the set of optimal rules to integrate into the embedded FRBC. Note that each rule is conceptualised by the following nomenclature:

$$R_j : IF\ \phi_a\ is\ \lambda_b\ and\ ...\ and\ \phi_c\ is\ \lambda_d\ then\ Y\ is\ \gamma_{n/c} \tag{5.7}$$

where $a \neq c$, ϕ are the different antecedents or features contained within rule j, λ are the activated linguistic variables for every antecedent, and γ is the respective

consequent of the rule. Note that for this research work, the implemented FRBC is based on a binary output, which leads up to two different classes or consequents, i.e. positive class for noisy segments (γ_n) and negative class for clean segments (γ_c).

To achieve the purpose of this stage, an evolutionary genetic algorithm (GA) is integrated and used to identify the rules that together give the best classification results. Note that such training optimisation has been used in previous works for different applications [257]. Specifically in this case, the GA is set to a maximum generations (i.e. maximum number of iterations before the algorithm halts) and population size (i.e. the number of feasible solutions) up to 50, uses a tournament selection function, and employs a one-point cross-over for the chromosomes combination. Note that the GA tolerance is fixed to $1 * 10^{-5}$. Thus, the structure of each phenotype is given by

$$\begin{aligned} \rho^j = \{&\phi_1^1, \phi_2^1, \phi_3^1, \phi_1^2, \phi_2^2, \phi_3^2, ..., \phi_i^j, \\ &\lambda_1^1, \lambda_2^1, \lambda_3^1, \lambda_1^2, \lambda_2^2, \lambda_3^2, ..., \lambda_i^j, \\ &\gamma_n, \gamma_c, ..., \gamma_i\}. \end{aligned} \tag{5.8}$$

Note that, initially, the rules are randomly generated, the maximum number of antecedents allowed for every rule (A_{max}) is fixed to three, and the maximum number of total rules (M) is set to ten. The latter considerations are done to assure that the final set of rules are comprehensive and interpretable enough [258].

Moreover, within this stage, a Rule Weight (RW) is assigned to every generated rule for both upper and lower memberships. This score is calculated as outlined in [259], following:

$$\begin{aligned} \overline{RW}_j &= \overline{c}_j \cdot \overline{s}_j \\ \underline{RW}_j &= \underline{c}_j \cdot \underline{s}_j \end{aligned} \tag{5.9}$$

where c_j and s_j are the rule confidence and rule support for rule j respectively. The former represents the likelihood or conditional probability of a pattern correctly classifying a data instance, while the latter is a measurement to quantify the rule

coverage over the training dataset. They are given by,

$$\begin{aligned} c_j(\phi_j \Rightarrow \gamma) &= \frac{\sum_{x_t \in \gamma} w_j^s(x_t)}{\sum_{j=1}^{M} w_j^s(x_t)} \\ s_j(\phi_j \Rightarrow \gamma) &= \frac{\sum_{x_t \in \gamma} w_j^s(x_t)}{M} \end{aligned}, \tag{5.10}$$

where x_t is every data instance contained within the training set, and w_j^s is the scaled strength of activation of such data with respect to every rule, i.e. the matching degree of rule j with input x_t. The scaled strength of activation is calculated as:

$$w_j^s(x_t) = \frac{w_m(x_t)}{\sum_{k,Y=\gamma} w_k(x_t)}, \tag{5.11}$$

where $w_m(x_t)$ is the strength of activation, and $w_k(x_t)$ is the sum of all strengths of activation that have the same class as the consequent of rule j. Finally, the strength of activation is computed as outlined in the following equation:

$$w_j(x_t) = \prod_{z=1}^{A_{max}} \mu_{\widetilde{A}}^z(x_t), \tag{5.12}$$

where $\mu_{\widetilde{A}}^z(x_t)$ represents the membership degree value of the x_t data instance for the interval type II fuzzy lower and upper membership degree functions, as denoted in equation 5.5.

SQi generation

During the evaluation of every GA iteration, the fitness is calculated based on a specific validation set. The split between training and validation sets was done using different CV techniques, hold-out and k-fold. On the one hand, different validation set percentages were employed for the hold-out validation. Specifically, the system has been trained using a random and stratified 40%, 30%, 20%, and 10% hold-out. On the other hand, a 5-fold disjoint training and validation datasets were used. These processes ensure that there is no bias in the selection of the training and validation datasets. Note that, as the signal segment acquisition and feature extraction are not subjected to any overlapping process, there is no information flow from the learning of rules from one training set or fold to others.

Regarding the specific SQi generation for each of the instances contained within the

validation set, two methods are applied. Both of them are based on the association degree computation with respect to the rule j being evaluated, which is given by

$$\overline{h}_j(x_t) = \overline{w}_j^s(x_t) \cdot \overline{RW}_j,\ \underline{h}_j(x_t) = \underline{w}_j^s(x_t) \cdot \underline{RW}_j, \tag{5.13}$$

where the strength of activation and the RW are obtained using equations 5.12 and 5.9 respectively. Thus, the overall association degree considering the contribution of the upper and lower type II membership functions for a rule j is computed as

$$h_j(x_t) = \frac{\overline{h}_j(x_t) + \underline{h}_j(x_t)}{2}. \tag{5.14}$$

Based on this final classification score, the first reasoning method (α) employed to assign the predicted class is based on the maximum matching method by selecting the consequent of the rule with the maximum association degree. The second method (β) is based on the maximum association degree from the aggregation of all association degrees having the same consequent. Note that, as the output of the system is binary, the latter is translated to the maximum between two accumulated association degrees. In the case of tie, we randomly classify the predicted class for both methods. Therefore, these processes can be expressed as:

$$Y_\alpha = \gamma_j \Rightarrow \max_{j \in [1,M]} (h_j(x_t)), \tag{5.15}$$

$$Y_\beta = \gamma_j \Rightarrow \max_{\forall k \in j} \left(\sum_{k, Y=\gamma_n} h_k(x_t), \sum_{k, Y=\gamma_c} h_k(x_t) \right), \tag{5.16}$$

where Y_α and Y_β are the predicted class obtained with the first and second reasoning method respectively. Up to my knowledge, this is the first time that the second reasoning method is proposed, validated and implemented.

Performance evaluation metrics

Finally, the performance assessment of every cross-validated iteration is done throughout the cost computed using the Mathew's Correlation Coefficient (MCC) as following:

$$cost = 1 - \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5.17}$$

where TP, TN, FP, and FN are the true positives, true negatives, false positives and false negatives obtained from the confusion matrix using the predicted labels compared with respect to the golden labels. Note that for the 5-fold cross-validation, the cost is computed as the mean of all fold validation dataset costs. After the cost is retrieved, the GA compares such value with a pre-defined tolerance criterion. If the cost is greater than the tolerance of GA, the GA then populates a new set of rules, and the process is repeated until the tolerance criterion of the GA is met. In addition to the MCC, other metrics are also used to further compare the different cross-validations. Such metrics are: sensitivity, specificity, geometric mean between sensitivity and specificity (Gmean), and accuracy (ACC).

5.2.3.2 SQA Implementation and Self-Tuning

Regarding the differences between the online (embedded) and offline architectures, it should be noted that no partitioning, membership generation nor rule optimisation is performed in the former, as these processes were already done during the training process. Figure 5-19 depicts the full embedded architecture. First of all, acquisition, conditioning and feature extraction follow the same schema as in Section 5.2.2. Secondly, with respect to the FRBC, the optimised set of rules, RWs and membership functions values, which were obtained after training and validation, are hard-coded parameters. These are specifically used for the online calculations of the scaled strength of activation and association degrees for every new coming data instance. Note that such computations follow equations 5.11 and 5.14 respectively.

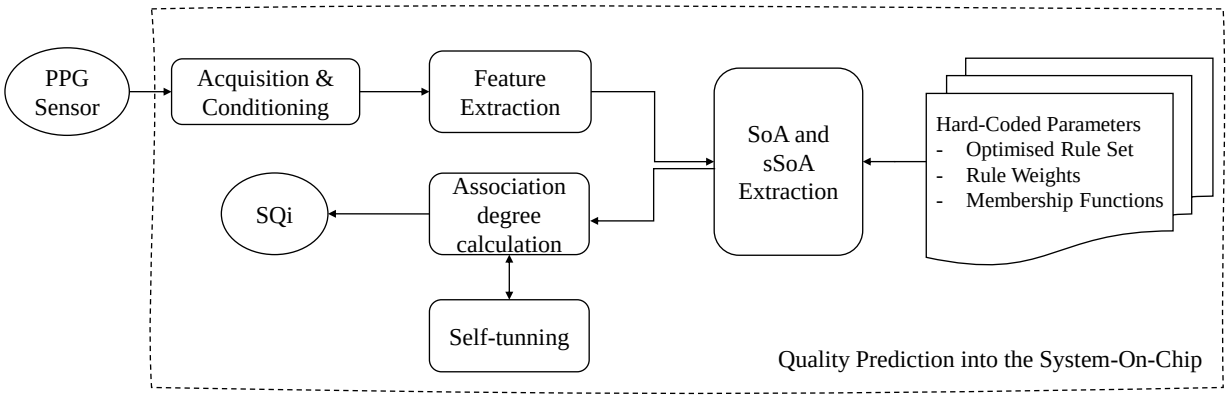


Figure 5-19: SQA embedded architecture implemented. SoA: Strength of Activation. sSoA: Scaled Strength of Activation.

The small rule base considered in this research work as a requirement to ease interpretability of the model can lead, in the long-term, to facing accumulated un-

certainties. Thus, we propose an online self-tuning type II FRBC. Specifically and considering previous works related to adaptive type II fuzzy systems [260], the implemented self-tuning is based on the online generation of new rules in case the new data instance shows zero association degree considering the initial ten rules, i.e. none of the existing rules fired. Algortihm 1 shows the implemented process for such online self-tuning. First of all, the antecedents and activated linguistic concepts conforming the new rule are estimated based on the maximum three strength of activation based on the membership degree of the new data instance, line 1. Note that the selection is limited to three due to the maximum antecedent requirement imposed, as stated in previous section. Thus, the new rule will always be formed by three antecedents with their respective activated linguistic concepts. Next, the degree of similarity is computed by using the Nguyen algorithm [261], line 2. This process does not require to run online computation of similarities, as the different degrees of similarities between the different memberships for every antecedent have been already calculated offline. There can exist rules, within the original rule base, that possess less than three antecedents, thus, to consider such particularities, the returned similarities are scaled by being multiplied by the number of antecedents of the rule being compared with. Following that, the RWs for the new rule are derived by considering the obtained array of similarities, i.e. there is a degree of similarity of the new rule with every existent rule, and the existent RWs within the rule base, lines 3-4. Finally, to estimate the consequent (γ) or class of the new rule, the consequent of the rule having the maximum similarity with the new rule is selected, lines 7-11. After running this algorithm, the rule base and RWs are updated with the new information. Moreover, the same inference process based on the association degrees, which was performed before the execution of the algorithm, is repeated assigning the corresponding label to the new data instance. To limit the impact of this process within the system, the maximum number of new rules was set to five, which can lead up to a maximum of 15 rules considering the initial 10.

5.2.3.3 Tools and Methods

A total of three different datasets were used to train, validate, and test the proposed SQA. First, we carried out our own experiment by which the few-shot training and validation data were gathered. The data was extracted from the experiment

Algorithm 1: Online self-tuning by using Nguyen similarities.

Input : New data instance (d_{new});Membership values (MFs);
Current rule base, lower and upper weights (R, RW_{low}, RW_{up});
Output: New rule base, new lower and upper weights
(R_{new}, RW_{lnew}, RW_{unew});
Data: Similarity for lower and upper membership (s_{low}, s_{up});
Mean similarity and Max position (s_{mean}, s_{maxPos});

$R_{new} \longleftarrow GetMaxSoA(d_{new}, MFs)$ based on (5.12);
$s_{low}, s_{up} \longleftarrow Nguyen(d_{new}, MFs, R, R_{new})$;
$RW_{lnew} = \sum_{j=1}^{M}(s_{low_j} * RW_{low_j})/\sum_{j=1}^{M} RW_{low_j}$;
$RW_{unew} = \sum_{j=1}^{M}(s_{up_j} * RW_{up_j})/\sum_{j=1}^{M} RW_{up_j}$;
Update $RW_{low,up}$ with $RW_{lnew,unew}$;
for $i \leftarrow 1$ **to** M **do**
  $s_{mean_i} = (s_{up_i} + s_{low_i})/2$;
end
$s_{maxPos} \longleftarrow FindMaxPos(s_{mean})$;
$R_{new}(\gamma) = R(s_{maxPos}, \gamma)$;
Update R with new rule R_{new};

explained in Chapter 6. Specifically for the proposed SQA system, a few-shot of 993 seconds of PPG signal recorded at 200 Hz from 10 different volunteers was used. Note that the stimuli were dynamic in terms of the movement of the volunteer, i.e. the volunteer could move without any restriction other than be sitting. Based on this data, manual annotations were performed by an expert who was familiar with PPG and artefacts for labelling the acceptable and unacceptable quality of the PPG segments. The labelling was assessed for every non-overlapped 3-second PPG window, which led up to 331 windows with 269 acceptable and 62 unacceptable PPG segments. Thereafter, two public benchmark datasets were used to provide a fully blind test, i.e. no information is provided about testing data during the training and validation. The first dataset is Capnobase [262], from which we obtained 9120 seconds of PPG signal, while the second dataset was the Complex Systems Laboratory (CSL) [263] with a total of 7200 seconds of PPG signal. Capnobase was the first public benchmark for respiratory and PPG quality analysis and originally contains 42 cases (volunteers) with 8 min duration PPG recordings at 300 Hz sampling frequency. However, just 19 out of the 42 cases present both acceptable and unacceptable PPG segments. This dataset provides artefact labels with no windowing temporal restriction. CSL gathers two hour PPG signals from two different volunteers, children in a pediatric intensive care unit, which was recorded at 125 Hz

sampling frequency. This dataset also provides artefact annotations with no windowing temporal restriction, which were recently released in [264]. It should be highlighted two main considerations regarding this specific database collection. On the one hand, a total of 17313 seconds of PPG signal were used, from which approximately only 5% is used for validation and 95% for blind testing. On the other hand, the specific selection of the two detailed testing datasets was based on the annotations availability. Thus, these considerations targeted the fact that the design of SQA systems needs to provide enough generalisation to deal with heterogeneous settings, as previously stated.

To adjust the PPG segment labels provided by the testing datasets to the 3-second processing window of the proposed system, the testing data was segmented into such window length and the label for each window was positive (unacceptable segment) in case of being within or overlapping with the original labels. Thus, after this process, the total amount of acceptable and unacceptable PPG segments obtained from Capnobase was 2909 and 131 respectively, and 2131 and 269 from CSL.

5.2.3.4 Results

This section presents the experimental results regarding the validation, testing, and real-time operation performance for the proposed PPG SQA system.

Validation and testing

Before performing the validation and further processes, the initial rule base was obtained by GA optimisation, as detailed in Section 5.2.3.1. Table 5.2 presents the obtained results for both reasoning classification methods and the employed validation techniques. Different considerations might be addressed before explaining such results. First of all, no self-tuning is applied during validation. Secondly, a total of 30 independent iterations are run for every cross-validation method, i.e. the training and validation partitions are randomly selected for every run. This is done to provide statistical value. Finally, the RWs are obtained for every training partition independently.

Regarding the different reasoning methods, α, as the one considering the maximum association degree amongst all the rules, and β, as the one considering the maximum association degree between the aggregated association degrees for the positive class and those for the negative class, are compared. It can be observed that β outper-

Reasoning Method	Cross-Validation Method	Validation Performance Metrics				
		Sensitivity $\mu(\sigma)$	Specificity $\mu(\sigma)$	Gmean $\mu(\sigma)$	MCC $\mu(\sigma)$	ACC $\mu(\sigma)$
α	40% Hold-Out	45.19 (7.17)	99.54 (0.47)	66.87 (5.21)	0.61 (0.06)	89.47 (1.37)
	30% Hold-Out	55.78 (16.68)	97.81 (2.29)	72.92 (11.03)	0.64 (0.11)	89.97 (2.67)
	20% Hold-Out	72.91 (11.00)	90.98 (3.98)	81.25 (7.23)	0.62 (0.13)	87.58 (4.56)
	10% Hold-Out	71.27 (14.93)	93.50 (5.10)	81.20 (9.80)	0.66 (0.18)	89.19 (5.67)
	5 k-fold	**83.71 (1.34)**	**92.63 (0.44)**	**88.05 (0.68)**	**0.73 (0.01)**	**90.95 (0.38)**
β	40% Hold-Out	86.46 (5.55)	87.57 (3.10)	86.96 (3.33)	0.66 (0.06)	87.37 (2.80)
	30% Hold-Out	87.68 (6.96)	87.59 (4.26)	87.56 (4.64)	0.67 (0.09)	87.60 (4.07)
	20% Hold-Out	87.09 (8.60)	88.24 (4.31)	87.57 (5.41)	0.68 (0.11)	88.03 (4.30)
	10% Hold-Out	88.41 (10.43)	90.10 (10.44)	89.07 (6.46)	0.72 (0.13)	89.80 (5.54)
	5 k-fold	**87.37 (1.30)**	**88.54 (0.45)**	**87.95 (0.66)**	**0.68 (0.01)**	**88.31 (0.40)**

Table 5.2: Validation performance metrics using both α and β reasoning methods and our own dataset.

Dataset	Reasoning Method	Testing Performance Metrics				
		Sensitivity	Specificity	Gmean	MCC	ACC
[262]	α w/o s-T	79.39	93.92	86.34	0.51	93.29
	α w/ s-T	**82.44**	92.05	**87.11**	0.48	91.64
	β w/o s-T	80.91	93.81	87.12	0.52	93.25
	β w/ s-T	**84.73**	90.82	**87.72**	0.47	90.55
[264]	α w/o s-T	71.75	99.48	84.48	0.81	96.38
	α w/ s-T	**75.47**	99.06	**86.46**	**0.82**	**96.41**
	β w/o s-T	73.60	99.48	85.56	0.82	96.58
	β w/ s-T	**81.41**	98.82	**89.69**	**0.84**	**96.88**

Table 5.3: Testing performance metrics for the different testing datasets using both α and β reasoning methods, and self-tuning (s-T).

forms α by reaching higher average metrics with less deviation in most of the cases. This is due to the RW balance or distribution between the rules, as in this case the rules having negative class consequent possess a higher RW in comparison with the rules having positive class consequent. Moreover, overall, it can be observed that α presents dependency over the amount of training data, whereas β provides a more robust system validation regardless of such fact. Note that, although this has been observed specifically for this train dataset, it could be applicable for other datasets as well as other problems. The best result for the α method is achieved by using 5 k-fold, which lead up to 88.05% and 0.73 of Gmean and MCC averaged values respectively. When comparing the results for the β method, it can be observed that the 10% hold-out cross-validation obtains the best averaged results. However, in case of considering the balance between the averaged metrics and their deviations, the 5 k-fold configuration shows the best performance with 87.95%, 0.68, 0.66, and 0.01 of Gmean and MCC averaged and standard deviation values respectively. This

analysis is completed by comparing the k-fold validation results for both reasoning methods. In that case, α presents slightly better metrics than β for all performance metrics except for sensitivity. The latter fact is an indication of the actual behaviour of the system for both reasoning methods over future unseen data, as β provides better sensitivity at the expense to decreasing specificity. Notwithstanding the latter differences between both methods, it can be concluded that the k-fold cross-validation outperforms the rest of the methods and, thus, the optimised RWs to be used for testing are obtained by averaging the RWs obtained during the k-fold validation considering the 5 folds and the 30 independent iterations.

After performing the validation of the system and obtaining the initial optimised rule base and their respective RWs, the test dataset collection is done as detailed in the previous Section. Table 5.3 shows the results obtained for both considered benchmark datasets and reasoning methods. In this case, we also provide the results regarding the self-tuning application. Note that in bold are those metrics that increased after the self-tuning integration. On the one hand, the β method generally achieves higher metrics than the α method for both datasets when the self-tuning is not applied. In fact, the difference between any of the metrics that are worse for the β than for the α method does not even exceed 0.2%. On the other hand, the application of the self-tuning process led to the addition of one new rule per dataset with positive class consequent. One clear difference can be highlighted between the results for such use case, as while for CapnoBase there is solely a sensitivity improvement, for CSL there can be observed an increase for most of the performance metrics except for specificity, whose worsening does not exceed 0.7%. This difference in the trend of the results for the different datasets can be attributed to the nature of the datasets itself. Note that the heterogeneity is present since these datasets contain data from different volunteers but also that the artefacts within the extracted PPG segments can have different characterised dynamics. In fact, although CapnoBase is bigger than CSL, the latter contains the double of unacceptable labelled PPG segments. Thus, the results and also the effects of the self-tuning application will vary based on how well is characterised the target to be detected (unacceptable PPG segments). In short, Table 5.4 presents the averaged testing results after combining both testing datasets. The best results are obtained using the β reasoning method

Reasoning Method	Averaged Testing Performance Metrics				
	Sensitivity	Specificity	Gmean	MCC	ACC
α w/o s-T	75.57	96.70	85.41	0.66	94.84
α w/ s-T	**78.96**	95.56	**86.79**	0.66	94.03
β w/o s-T	77.25	96.64	86.40	0.66	94.92
β w/ s-T	**83.07**	94.82	**88.75**	0.66	93.72

Table 5.4: Averaged testing performance metrics using both α and β reasoning methods, and self-tuning.

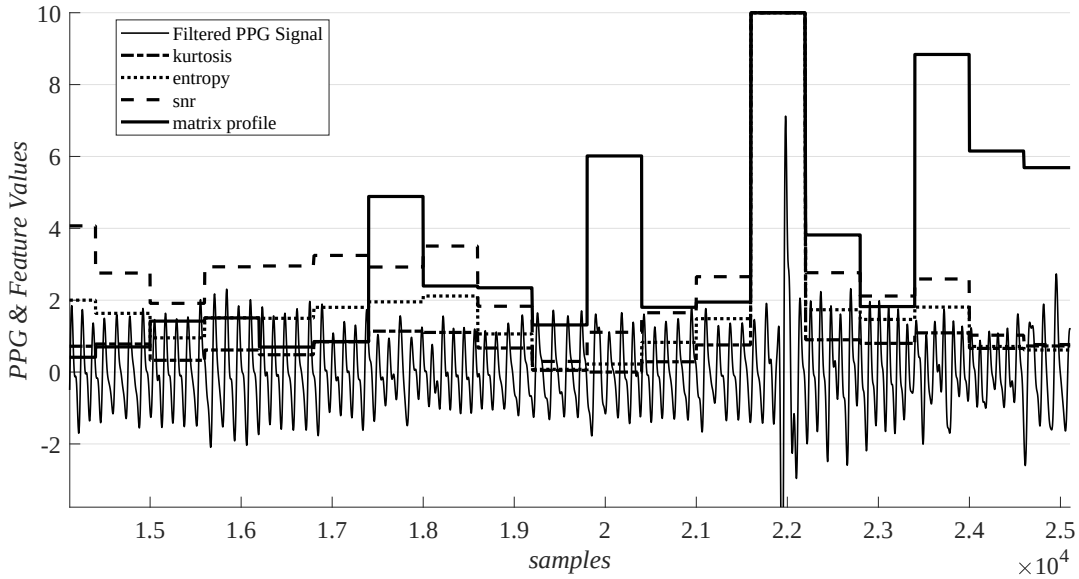


Figure 5-20: Real-time capture for the embedded SQA implementation showing the different feature values every processing window (3-sec).

and the self-tuning adjustment. This leads up to 88.75% and 0.66 Gmean and MCC averaged, which is comparable to the obtained validation results.

Real-Time operation performance

For the embedded implementation of the proposed SQA, the initial optimised rule base, RWs and membership functions are hard-coded into the SoC. This is done by quantifying such parameters using 32-bit registers. To validate the embedded integration, 33 PPG segments (24 acceptable and 9 unacceptable 3-second PPG segments) obtained from the first patient of the validation dataset are run into the SoC. For instance, Figure 5-20 depicts an excerpt of the filtered PPG signal as well as the different feature values for every 3-second processing window evaluated. Note that the embedded reasoning method is β, as it achieved better results during offline validation and testing.

Table 5.5 reports the performance metrics comparison between the embedded and

Platform	Sensitivity	Specificity	Gmean	MCC	ACC
MATLAB®	88.89	95.83	92.29	0.85	93.94
SoC	88.89	91.67	90.27	0.78	90.90

Table 5.5: Comparison between the embedded and MATLAB® performance metrics obtained for the 33 PPG segments evaluated into the SoC using the β reasoning method.

	SoA	Kurtosis	Entropy	SNR	MP	−AD	+AD
R^2	0.99	0.99	0.98	0.99	0.97	0.94	0.80

Table 5.6: Coefficient of determination (R^2) for the main processes performed within the SoC. SoA: Strength of Activation. MP: Matrix Profile. − and + AD: Negative Class Association Degrees.

the offline implementation. As expected, the latter achieves better performance. Specifically, it reaches up to 92.29% and 0.85 of Gmean and MCC values respectively, while the embedded implementation makes those results worse by lowering them to 91.67% and 0.78. This performance difference is due to the loss of precision for the different processes within the SoC. In fact, Table 5.6 shows the coefficient of determination (R^2) between the embedded and the MATLAB® results for the main processes of the system. Considering that this is the first implementation of the proposed system, most of the processes deal with 32-bit floating-point numbers within their operations (IEEE 754). Thus, the less computationally demanding and most simple stages such as the strength of activation, Kurtosis, and SNR calculations get a R^2 of 0.99. However, when facing more complex mathematical operations such as the logarithmic within the Entropy calculation and the multiple FFT algorithms performed during the matrix profile computation, the R^2 drops to 0.98 and 0.97 respectively. Although these precision errors are low, their accumulation along the entire prediction make the final process, that is the association degree for negative and positive classes, to have a R^2 of 0.94 and 0.80 respectively. This is the reason for the performance metrics drop within the SoC. Note that the MATLAB® implementation operates with double data types, which corresponds to 64-bit floating point numbers.

Regarding the energy saving analysis with and without SQA method, Table 5.7 reports four different test signal scenarios. The latter were chosen to ease the comparison with [247] and [248]. Regardless of applying the proposed SQA method,

Test Signal Scenarios		System without SQA			System with SQA			Overall Energy (with SQA) saving/extra
	EC_{Sensor} (mJ)	EC_{SQA} (mJ)	EC_{TR} (mJ)	Total (mJ)	EC_{SQA} (mJ)	EC_{TR} (mJ)	Total (mJ)	
60-sec Noise Free Signal	1053.36	NE	318.60	1371.96	59.40	318.60	1431.36	4.3% Extra
60-sec Noisy Signal	1053.36	NE	318.60	1371.96	59.40	NE	1112.76	20.7% Saving
6-sec Noisy out of 60-sec Signal	1053.36	NE	318.60	1371.96	59.40	292.05	1404.81	2.3% Extra
12-sec Noisy out of 60-sec Signal	1053.36	NE	318.60	1371.96	59.40	238.95	1351.71	1.5% Saving

Table 5.7: Real-time energy saving analysis with and without SQA method. EC_{SQA}: Energy consumption for the SQA implemented system. EC_{Sensor}: Energy consumption from the PPG sensor. EC_{TR}: Energy consumption for BLE Transmission. NE: Not executed.

the energy consumption of the PPG sensor is unavoidable for a continuous physiological monitoring system or application. Thus, the energy consumption baseline of the system is 1053.36 mJ, which corresponds to the normal functioning of the sensor as well as the I2C communication involved to gather the data from it. For the transmission of every 3-second processing window, the energy consumed by the BLE is around 15.50 mJ, which lead up to 318.60 mJ consumed for 60-sec noise free signal transmission. The energy consumption due to the execution of all stages involved into the proposed SQA system is 59.40 mJ. Thus, considering the different test signal scenarios, we can conclude that the proposed SQA method can save an overall energy power consumption from 1.5% to 20.7% for noisy PPG signals with duration from 12 to 60 seconds. Conversely, the extra energy consumption due to the SQA execution reaches up to 4.3% for an entire 60-second noise free signal. Finally, time and memory complexities were also quantified for the proposed SQA. The obtained averaged time to execute the proposed SQA method was 53.07 ms. The total memory required to handle global and temporary variables, as well as acquisition and processing buffers is 15kB.

To contextualised some of the obtained results with respect to other reported works on SQA, Table 5.8 presents the key metrics for the proposed system and three recent works [247–249] that were also reviewed. The proposed work provides comparable performance metrics to the state-of-the-art. It should be noted that the other works did not use the artefact labels provided by the benchmark datasets. Instead, they labelled such data again. Moreover, none of them used exactly the same datasets, whether they were using them for threshold adjustment or testing. In terms of

Work	Validation Metrics	Experiment Independent	Few Shot	Training Observations	ACC (%)	Clock (MHz)	Memory (kB)	Energy (mJ)	Platform
Vadrevu (2019) [247]	Absolute amplitude, crossing rate and autocorrelation	×	×	38620	95.93	84	13	210	SAM3X8E ARM Cortex-M3
G.N.K. Reddy (2020) [248]	FOPC-DC feature	×	×	15000	93.21	84	29.56	–	SAM3X8E ARM Cortex-M3
Samiul Alam (2021) [249]	Kurtosis, and autocorrelation, empirical thresholds	×	×	8000	96.50	1200	88	63.1	Quad-core ARM Cortex-A53
Proposed	Type-2 Fuzzy Subject-Invariant (β w/ Self-Tuning)	✓	✓	331	93.72	64	15	59.40	nrf52832 ARM Cortex-M4

Table 5.8: Comparison with reported work on SQA.

memory and energy consumption, we provide one of the lowest metrics. Specifically for the energy consumption, [247] and [248] also provided real-time energy saving analysis comparable to the one reported in Table 5.7. They obtained an overall energy saving above 90.00% for the second test signal scenario. However, they did not considered nor reported the power consumption of the sensor, which is supposed to be continuously working. It should be also noted the platform difference, as while [247] and [248] used a comparable embedded device (microcontroller) to the one used in this research work, [249] employed a Cortex-A53 which is far from being able to be properly compared to ours. Finally, the few-shot and experiment-independent approaches are highlighted towards the application and adaptation of the system to heterogeneous settings as well as into-the-wild usability.

The proposed SQA provides a simplified low-complexity fuzzy rule-base Mamdani inference model deployed in low resource edge devices. The main novelty of this research is the non-heuristic, adaptive, wearable oriented, and subject-invariant aspects of the proposed SQA system. First, the non-heuristic feature is obtained by using a novel unsupervised method for generating interval type-II fuzzy sets from PPG signals based on quantization. Secondly, the adaptation of the system is achieved by defining and implementing a novel online unsupervised fine-tuning based on scaled similarity between interval type-II fuzzy sets for model self-adaptive updates. Finally, the subject-invariant aspect and heterogeneity are accomplished since all the datasets employed contain data from different volunteers. This fact makes the artefacts within the extracted PPG segments have different dynamics. To demonstrate the online PPG SQA implementation, a detailed analysis of the embedded performance of the proposed methods in the Bracelet is performed, out-

lined and compared against the state-of-the-art. Overall, the proposed work provides comparable metrics to the compared state-of-the-art. It achieved an overall blind testing accuracy of up to 93.72%. The real-time evaluation showed an energy consumption up to 59.40 mJ for the proposed SQA, which led up to 20.7% overall energy savings. Within this context and comparison, certain limitations of the proposed system must be also considered. First, further digital signal processing optimisations can be applied, such as smaller integer computations scaling and single instruction multiple data. Second, further experimentation and data gathering are being performed to increase the training data and explore the design space. Some of the advantages and limitations identified while performing this system confirm the need for SQA systems focused on providing enough generalisation to deal with heterogeneous settings as well as SQA embedded implementations into extreme edge devices. Note that, although we focus solely in the PPG signal, further research can be investigated towards similar approaches for the other signals considering this system as reference.

5.2.4 Feature Extraction Design Space Exploration

This section presents a feature extraction design space exploration, which is focused on extracting BVP related information. The presented analysis is divided into the different stages in the software architecture, as Fig. 5-21 shows. Thus, it is assumed that the considered signals are properly filtered and segmented before the application of any feature extraction technique. In each explored stage, parameters of interest are evaluated and recommended. First of all, the morphological delineation (Peak detection block) is discussed by means of a detailed comparison of different peak detection algorithms. Secondly, the common applied techniques to extract frequency information when dealing with the obtained unevenly or non-uniformed delineated points, i.e. interpolation or beats counting blocks, are presented. For the sake of the Bindi application, the interpolation technique is employed, implemented and discussed. Finally, specific recommendations of the different trade-offs discussed on these Sections are applied in Section 5.2.4.3 for a particular 4-second stress detection use case, in which a validation of the embedded feature extraction processes into the Bracelet are validated against a research-grade tool.All metrics considered in this section to characterise the different stages are obtained from the

embedded implementation within the design space exploration into the Bracelet. Note that the final selection of the different evaluated parameters will depend upon the requirements and needs of the application. As for Section 5.2.2, this analysis is extracted from [159]. Thus, although the evaluated algorithms are focused on a PPG use case, some of the feature extraction techniques, such as the FFT, are common to the rest of the signals.

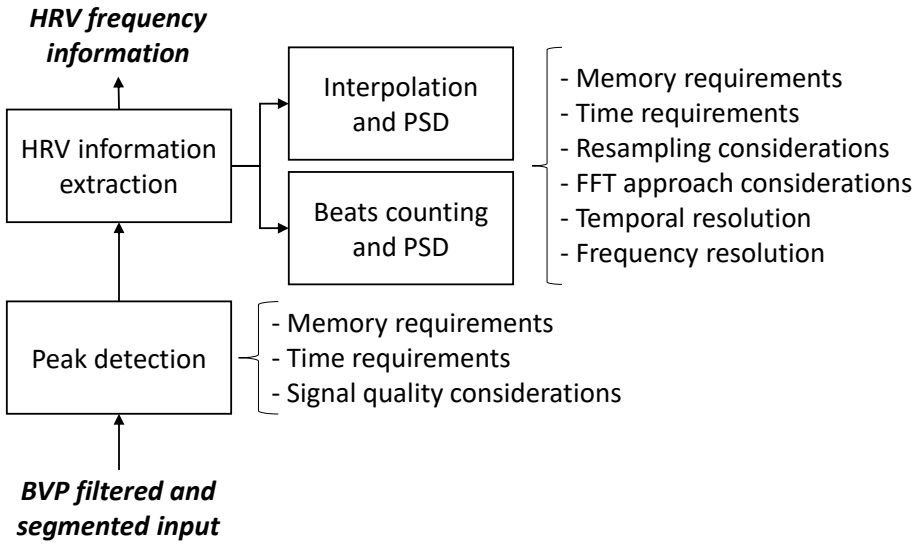


Figure 5-21: Parameters and processes involved in the BVP-based DSE.

5.2.4.1 Feature Extraction: Peak Detection

Focusing specifically into the PPG casuistry, different approaches can be used to delineate PPG time series. The robustness of this delineation process is key to properly detect the desired morphological PPG parameters. This fact is determined not only by the previous filtering steps but also by the different PPG waves morphology, which can be directly affected by factors such as age and emotions [265,266]. Figure 5-22 shows the morphological difference between three different age groups measured with our PPG sensor. The observed differences are in line with the ones published in the literature [267]. For instance, the dicrotic part of the wave is the most affected. This fact is mainly due to the vascular tone variation with age, which is translated directly into more or less vasoconstriction and vasodilation. This situation produces differences in arterial pressure leading to distorting diastolic run-off. Within this physiological variable context, different delineation algorithms could provide different results, i.e. different identified morphological points from the PPG signal. This fact can even get worse when constraining the initial stage filtering architectures applied, which used to happen in physiological monitoring embedded applications.

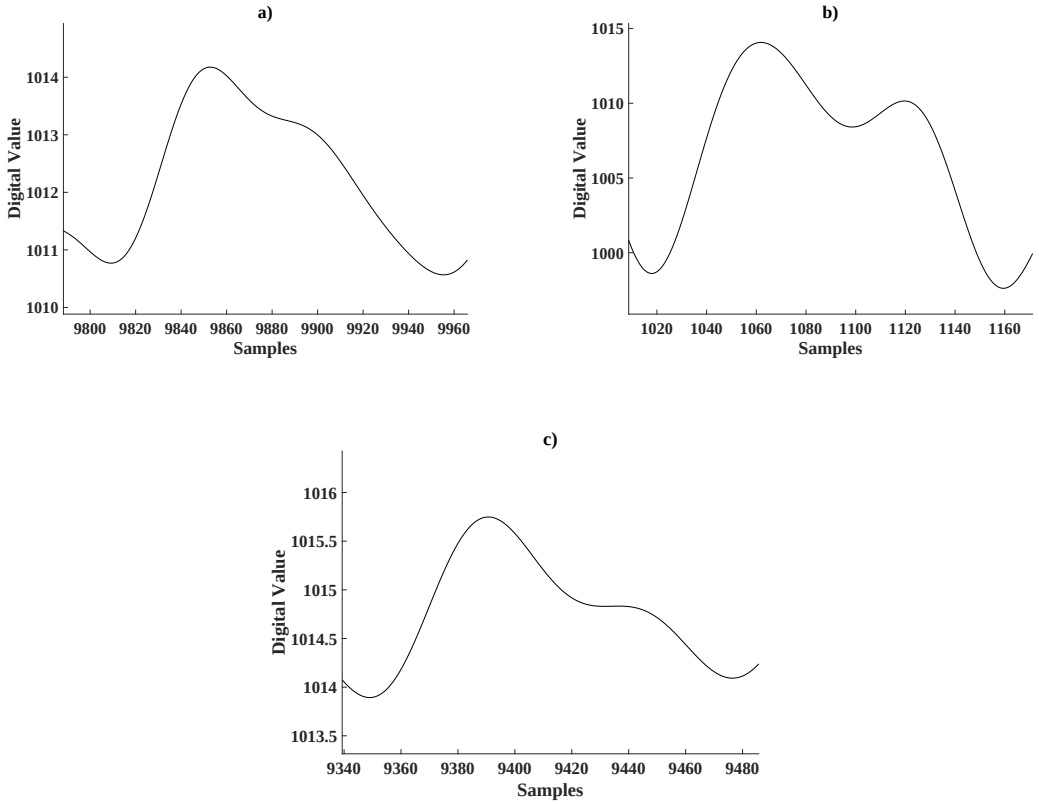


Figure 5-22: PPG morphological differences between three age groups. (a) 18-24-year-old person. (b) 35-44-year-old person. (c) 55-65-year-old person. The signals shown was acquired by the Bindi bracelet.

For instance, if the application does not use any baseline wander removal filter, e.g. notch filter below 0.5Hz, the different points extracted by the delineation algorithm employed should be robust enough to be unaffected by the low-frequency out-of-band trends. Different techniques can be applied to assure robust peak-to-peak detection. However, some of them require the implementation of zero-crossing throughout the first and second derivatives of the signal [268, 269]. This fact directly affects the computational time within the data processing chain.

Considering the commented particularities of this stage for a PPG use case, a comparison between two well-known lightweight approaches is presented. On the one hand, the first is based on a local maximum/minimum method (LCM) developed by the UC3M4Safety team using the local slope and mean evolution over short periods of samples along with the data processing window [159]. LCM methods are well known within PPG peak detection algorithms as they used to be less computationally demanding at the expense to lower performance. On the other hand, the second algorithm is taken from [270], which is based on an adaptive threshold detection

method (ADT) using a varying slope calculated iteratively based on the standard deviation of the signal. Note that the second algorithm was validated against publicly available datasets and outperformed LCM techniques without requiring first and/or second derivative signal operations.

For the first method, Algorithm 2 describes the operations performed for each BVP window. Specifically, the algorithm starts assuming that the maximum value is the first sample of the signal. After that, mean and slope values over a specific number of samples to be compared (*stc*) are calculated and evaluated. This step is performed if one of the two threshold signal level conditions (max or min) are met. Thus, a balanced trade-off to adjust *stc* is essential. For instance, if the BVP signal is sampled at 100Hz, the mean evaluation over ten samples supposes -6dB attenuation for 6Hz and -3dB attenuation for 4.5Hz, being the latter close to 4Hz which is a frequency of interest for targeting maximum cardiac frequencies (BPM$\geq$ 240). Therefore, based on the expected residual high noise frequencies of the signal filtered, this parameter can be adjusted. For this particular algorithm, and guided by the trade-off taken on the filtering stage in Section 5.2.2, a *stc* equal to ten can be chosen, which increases peak detection capabilities at the expense of time complexity. Another key parameter included within the algorithm is $dist_{min}$, which is initially assigned to a specific number of samples k. This variable is referred as to the minimum permitted distance between two identified systolic peaks and it is fixed to the number of samples for the highest frequency within the BVP bandwidth. As for the previous example, if the BVP signal is sampled at 100Hz and the highest acceptable HR frequency is 3.5 Hz (210 BPM), then k is set to 28 (samples). This parameter does not affect the algorithm time complexity, but rather provides robust handling for possible transients that are still in the signal and could be affecting the peak detection. Moreover, we introduced different conditions into the developed algorithm to cover special morphological cases. On the one hand, in the case of dealing with wide systolic crests, just the last point of such is considered the systolic peak, lines 11-17 of Algorithm 2. On the other hand, in case of having short systolic crests, it could happen that a potential peak is left behind, which is taken into account in lines 19-21 of Algorithm 2. Although the algorithm is used to extract the systolic peaks information, valleys processing is also performed within it. The latter

is performed through the opposite operations than for the peaks processing. Note that, in our case, the signal is centred without any DC drift or tendency before the application of this algorithm thanks to the previous filtering stages.

Algorithm 2: BVP Peak Detection Algorithm

```
1  function getPeaks (bvp_signal, bvp_len);
   Input :
   Clean BVP signal bvp_signal;
   Total number of samples bvp_len;
   Output:
   Detected peaks position peaks_index;
   Total number of peaks peaks_total;
   Data:
   Max and Min for each search peaks_max, peaks_min
   Counter for detected peaks peaks_count
   Minimum separation between detected peaks dist_min
2  peaks_max ⟵ bvp_signal(0);
3  peaks_min, peaks_count ⟵ 0; dist_min ⟵ k;
4  for i ← 1 to (bvp_len − stc) do
5      if bvp_signal(i)>peak_max then
6          peak_max ← bvp_signal(i);
7          peaks_min ← peaks_max − stc;
8          Get vtc_mean for [i,stc];
9          Get vtc_slope for [i,stc];
10         if vtc_mean≥peak_max then
11             if peaks_count&& i − peaks_index(peak_count − 1) < dist_min then
12                 peaks_index(peak_count − 1) ← i;
13             else
14                 peaks_index(peaks_count) ← i;
15                 peaks_count ← peaks_count + 1;
16                 peaks_min ← 0;
17             end
18         else
19             if peaks_min && vtc_slope<0 then
20                 Update counter, index and peaks_min;
21             end
22         end
23     end
24     if bvp_signal(i)<peak_min then
25         Perform opposite operation to detect valleys;
26     end
27 end
28 peaks_total ← peaks_count
```

Figure 5-23 shows an analysis on the time impact for the two different peak detection algorithms considering the number of samples in the processing window. Related to this time complexity, a linear performance can be observed for our LCM method, due to the neighbour evaluation done with every sample. Regarding the ADT algorithm, an increase in computational time between 30 % and 50 % can be observed compared to LCM with the highest *stc*. This difference is mainly due to the mandatory calculation of the standard deviation for the whole processing window

signal, which is needed to obtain the varying slope to be used by the ADT algorithm. Related to memory storage considerations in this stage, the implemented LCM algorithm needs 2KB of ROM and 32B of RAM. In the case of the ADT implemented, 2.5KB of ROM and 64B of RAM are used by this algorithm.

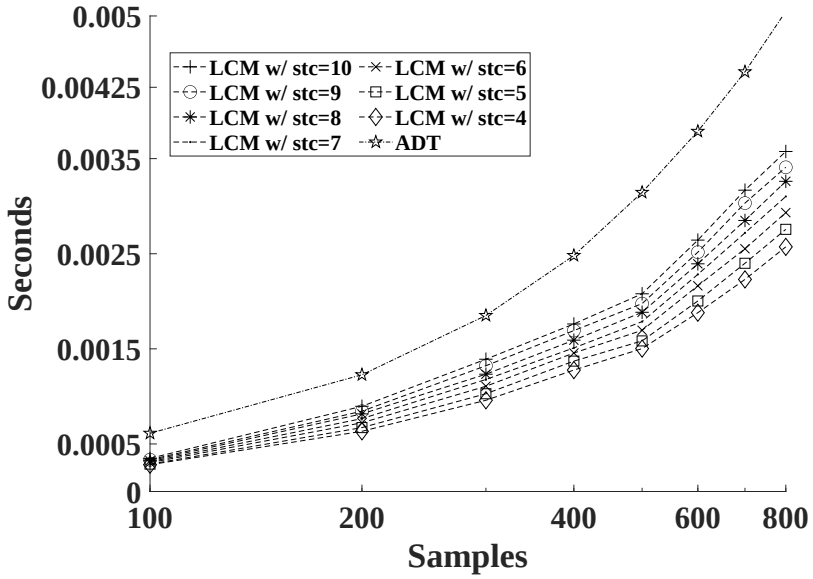


Figure 5-23: Time impact analysis for the peak detection algorithms considered.

From this stage and considering the time complexity and resource usage results by the evaluated algorithms, we conclude that, although the LCM methods are one of the simplest, they can fit the requirements of many applications. Thus, the use of such lightweight algorithms is recommended. For the sake of providing a real application validation following these feature extraction trade-offs, Section 5.2.4.3 presents the results from an actual HRV use case activation monitoring.

5.2.4.2 Feature Extraction: HRV Information

In digital constrained embedded systems, frequency analyses are performed by DFT. One of the usual algorithms is the Fast Fourier Transform (FFT). However, this algorithm is based on the assumption of an equidistant sampled input. At this point, two possibilities arise based on the application needs. If the application is not limited by any inference time restriction, the system can wait until enough HRV points are extracted and the desired frequency resolution is possible. On the contrary, when continuous rapid inference is needed within a fixed temporal window, interpolation between the HRV samples is applied to reestablish the temporal coherence.

Focusing on continuous rapid use cases to boost the response time of Bindi, we considered two main parameters for the evaluation of this feature extraction stage: the type of interpolation and the length of the FFT. Fig. 5-24 shows a time impact analysis for both the interpolation techniques (linear and polynomial) and the FFT implemented and considered for different window processing lengths. As expected, polynomial methods have a higher time complexity, although producing more precise results if spectral accuracy is needed subsequently. Note that Lagrange polynomial quadratic interpolation is considered for this comparison. Regarding the FFT, a fixed-point 32-bit radix-2 FFT algorithm is used, which provides one of the lowest computational complexities ($\mathcal{O}(n \log n)$) and is then adequate for the embedded device. It is noteworthy that for all the FFT lengths evaluated, applying polynomial interpolation implies doubling the processing time with respect to linear interpolation. Considering this fact and that quadratic interpolations within the time-domain are preferred for HRV [271], such temporal complexity difference can be taken.

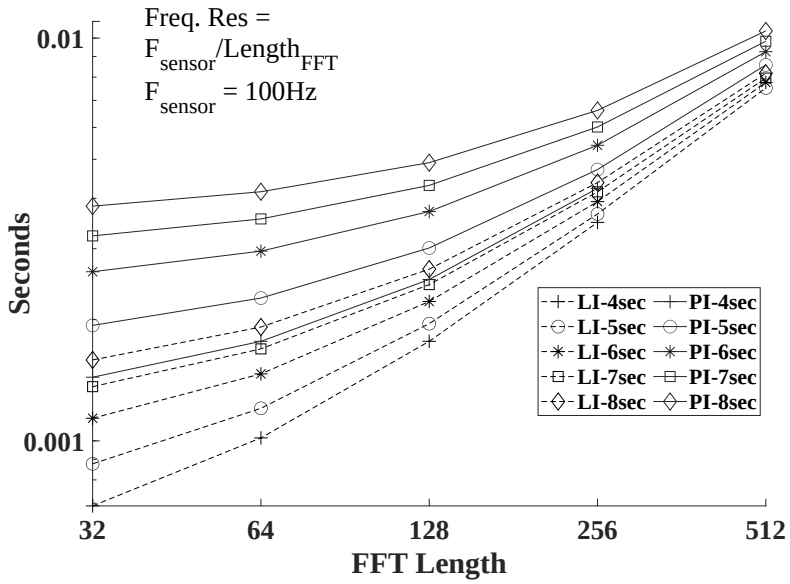


Figure 5-24: Time impact analysis based on different interpolation methods and the FFT implemented and considered.

In this stage, a trade-off between temporal and frequency resolution must be considered. Note that independently of the f_{sensor}, if the window processing length is fixed, the frequency bin resolution for the chosen FFT_{len} will not change. Thus, to improve frequency resolution for a fixed temporal window, resampling techniques are applied after interpolation in these situations. For instance, if the HRV is in-

terpolated at 100Hz for a fixed four-second time window, it results in a 0.39Hz/bin resolution. However, after applying a 1Hz resampling, frequency resolution increases up to 0.25Hz/bin. Note that for the latter resolution, only 256 available points are taken. In case of taking more points than the window length, zero-padding must be applied. Thus, time and frequency resolution, as well as interpolation and resampling techniques, depend on the application. This is a key aspect when dealing with applications that require HRV frequency information extraction, as the lowest band of interest is located from 0.01Hz to 0.04Hz. Therefore, to achieve full HRV frequency band detection capability a minimum of 0.04Hz/bin should be assured. A frequency resolution value higher than that will decrease such detection capability or spectral bands separability. Note that frequency bin resolution is given by equation 4.7. Regarding temporal resolution, there must be considered that the duration of the processing window must be selected to assure the presence of at least two HRV points. Otherwise, interpolation is not possible.

Related to memory storage considerations in this stage, special care should be taken for FFT resource requirements by implementing in-place properties and non-recursive behaviour. Note that resources used during the resampling operation are considered negligible. For the interpolations, both consumes up to 698B of ROM and 10B of RAM, while the FFT needs 3KB of ROM and 548B of RAM.

This last step is especially sensitive. For instance, linear interpolation can even introduce deformations in the resulting power spectra. In this case, a quality-based design decision is recommended to prevail the physiological information. Thus, assuming the same resource usage for both interpolation methods and considering all the physiological advantages that polynomial interpolation provides, the latter is recommended over linear.

5.2.4.3 HRV Use Case Implementation

To give a real use case and implement all the different recommendations concluded for the latter feature extraction exploration, a specific four second physiological activation rapid-inference application is presented. Bindi bracelet is programmed with all the detailed signal processing architecture and taken trade-offs. In this case, an experiment with six volunteers and ten different stressed and non-stressed one-minute audiovisual stimuli was used. These stimuli were previously labelled and

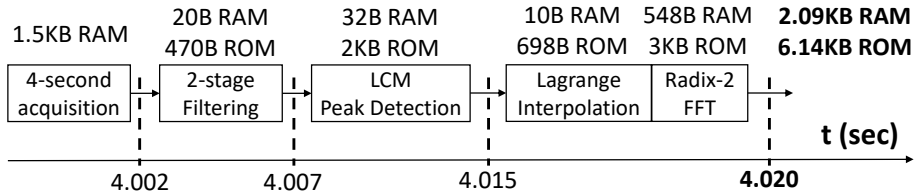


Figure 5-25: Complete data chain for a 4-second window processing given the trade-offs discussed.

selected by the authors. After each stimulus, the volunteers self reported their own level of arousal or excitement felt when watching the video. To provide a validation tool or a golden measure against the signals acquired by our platform, a research-grade sensory system[1] was considered.

For this experiment, a 100Hz f_{sensor} was used and a fixed four second temporal processing window was employed, which required a buffer of 400 samples (1.5KB). Note that for HRs bellow 45BPMs, this window is not applicable, as only one HRV point could be found. Every four seconds, the HRV points are extracted and interpolated, which is followed by a FFT calculation and a PSD estimate given by

$$PSD_i = \frac{2 * |fft|}{s}, \tag{5.18}$$

where PSD_i is the power spectral density for one specific frequency bin i, $|fft|$ is the squared spectrum magnitude and s is the sum of squared samples of the window function used. Specifically, to deal with scalloping loss and picket fence effects, a flat top window is applied. A fix FFT_{len} of 256 points is used, leading to a 0.39Hz/bin resolution. This resolution is enough to observe the activity of lower frequency bands (up to 0.4Hz) and higher ones (from 0.4Hz up to 1Hz). Take into consideration that in case of having less than 256 points after interpolation, zero padding is applied. The same digital procedure is applied for the validation tool and Bindi. Considering all the data provided in previous sections, the final implementation of all the different stages for this specific application requires up to 2KB RAM, 6KB ROM and takes about 20 milliseconds to provide a valid HRV estimation from the completion of one processing window, Figure 5-25.

Table 5.9 shows the collected results obtained for two arbitrary selected stress (H) and non-stress (L) stimuli for the six different volunteers. $\overline{P}_{Gf1}$ is the averaged

[1]https://www.biosignalsplux.com/index.php/researcher

Table 5.9: Measurement result of specific HRV stress detector use case.

Type	$\overline{P}_{Gf1}$	$\overline{P}_{Gf2}$	$\overline{P}_{Bf1}$	$\overline{P}_{Bf2}$	ε [%($\varepsilon_{f1}, \varepsilon_{f2}$)]
1_H	5.28	0.15	5.27	0.16	(0.18,6.66)
1_L	5.05	0.16	4.81	0.17	(4.75,6.25)
Δ	-0.23	+0.01	-0.46	+0.01	
2_H	4.09	0.24	4.29	0.20	(4.88,16.66)
2_L	3.83	0.27	3.45	0.29	(9.92,7.41)
Δ	-0.26	+0.03	-0.84	+0.09	
3_H	5.18	0.16	5.17	0.16	(0.19,0.00)
3_L	4.40	0.19	4.28	0.21	(2.72,10.52)
Δ	-0.78	+0.03	-0.90	+0.05	
4_H	5.27	0.15	5.32	0.15	(0.09,0.00)
4_L	5.16	0.16	5.12	0.17	(0.7,6.25)
Δ	-0.11	+0.01	-0.20	+0.03	
5_H	5.07	0.16	4.82	0.17	(4.93,6.25)
5_L	4.64	0.17	4.63	0.18	(0.21,5.88)
Δ	-0.43	+0.01	-0.19	+0.01	
6_H	4.98	0.16	4.96	0.17	(4.03,6.25)
6_L	4.84	0.17	4.46	0.20	(7.85,17.64)
Δ	-0.14	+0.01	-0.50	+0.03	

quotient between the first frequency bin (0.39Hz) and the second frequency bin (0.78Hz) during the stimulus using the signal from the validation tool, while $\overline{P}_{Gf2}$ is the one observed for the averaged quotient between the second frequency bin and the sum of the first and the second. The fourth and the fifth columns are the analogue values taken from Bindi. These results show a decrease on the first factor for all the patients from the stress to the non-stress stimulus. Conversely, there is an increment in the second factor. This is in line to the theory of the ANS. As commented in Chapters 2 and 4, the lower frequency bands are dominated by the SNS which is in charge the *fight-or-flight* response of the body, while PNS is related with the higher bands and responsible of controlling relaxed (*rest-and-digest*) conditions. The errors between the validation results and Bindi results are also provided in Table 5.9 . These errors are low ($\varepsilon < 10\%$), except for cases such as 2_H or 6_L, in which strong motion artefacts presented in the signal of Bindi were not cleaned as expected, resulting into locally contaminated signals segments, which affects directly to the peak detection process and, therefore, to the HRV extraction, see Figure 5-26.

By performing this particular use case, the different detailed trade-offs applied to rapid-inference applications have been successfully implemented. Notwithstanding

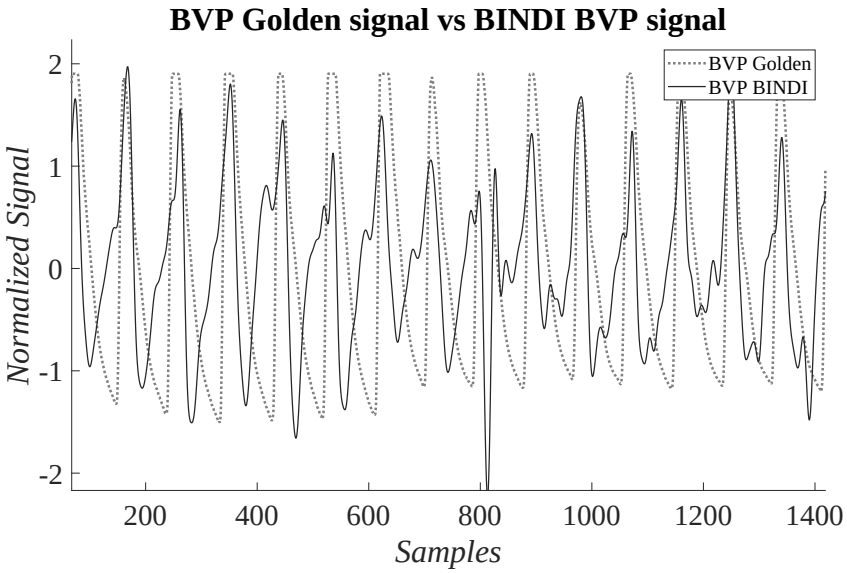


Figure 5-26: Motion artefacts effects displayed in one segment of the stress audiovisual stimulus of volunteer 2.

that the presented use case does not reach the full HRV frequency band detection capability, the goal of rapid stress detection is accomplished using low amount of resources at the expense of frequency resolution.

5.2.5 Power Consumption Analysis

Power consumption management is a requirement in the design of a wearable system. In Bindi, an accurate measure of the state of battery charge and autonomy of the two wearable devices is essential to ensure that the system works when needed. This section provides a quantitative current consumption analysis for the Bracelet. This analysis is performed by measuring the most energy-demanding actions through the monitoring part of the device. Thus, the electric current consumed by acquiring data through each physiological sensor is measured separately. Moreover, the power consumption incurred by making use of the buzzer in soft, medium, and strong intensities is also measured. Thus, we chose to measure the power consumption due to sensor data communication and acquisition, which are essential for the system and are intrinsically related to the specific hardware design of the devices.

The results obtained in the current consumption analysis for Bindi 1.0 appear in Figure 5-27. The vibration modes are the most current consuming actions, where the higher the vibration produced, the higher the current required, as expected. However, the buzzer impact on autonomy is reduced because it is activated for a

short time in risky-related situations, meaning that its activation may usually be sporadic. The SKT and GSR sensors also produce a small increment from the idle state. However, the PPG sensor has a higher impact than the other sensors. Thus, we can conclude that the current bottleneck of the system, in terms of power consumption and operating time, is the PPG sensor. Notwistanding such fact, the low power consumption in the idle state makes the Bracelet battery life to be approximately 40 hours when using a 500 mAh battery. Note that these calculations are based on no-alarm situations.

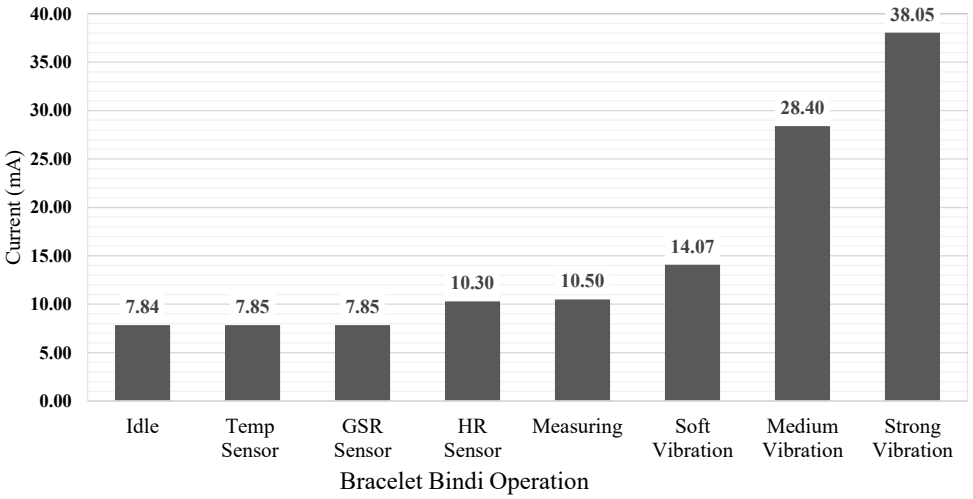


Figure 5-27: Average current consumption in the Bracelet [9].

5.3 Conclusion

This Chapter reported the main contributions of this research regarding the design of wearable systems oriented towards continuous physiological acquisition and monitoring. It has become clear that the design of wearable systems for the continuous monitoring of physiological signals and the design and embedded integration of affective computing related processes is a challenging task that requires a careful balance between embedded resources, power consumption, and system response time. Moreover, we presented different works contributing to the embedded implementation of machine learning related processes on the edge, i.e. into our Bracelet. It should be highlighted that the complete data processing chain for all the sensors, from acquisition to classification, is currently being improved and expanded towards considering and sizing the embedded impact as well as its limitations.

The Section 5.1 provided the necessary technical background to contextualise the

	BINDI	COMETA	ATENPRO	AlertCops	Safer Pro	EMPATIC A Embrace
Smart-watch/bracelet factor form	**Yes**	No	No	NA (App)	Yes	Yes
Automatic alarms	**Sí**	No (just GPS)	No	No	No	Sí
Technology	**Cutting-edge**	Obsolete	Obsolete	Mobile-based	Improvable	Cutting-edge
LEAs warning	**Yes**	Yes	Yes	Yes	No	No
Sensor Monitoring Goal	**Emotional state**	Distance (GPS)	No	No	No	Seizures
Use-Case	**GBV**	GBV	GBV	General (including GBV)	GBV	Epilepsy

Figure 5-28: Bindi's competitive advantage over its main and most direct competitors.

different highlights that Bindi's technology can offer. Figure 5-28 presents a compact summary regarding Bindi's competitive advantages over its main and most direct competitors. Thus, we can conclude that the reviewed available technological solutions to combat Gender-based Violence, the ones oriented for a general use-case, or even solutions with different goals but being technologically comparable to Bindi neither offer the same functionalities nor take advantage from the cutting-edge technology. Bindi possess great potential to be an effective technological tool to prevent and combat Gender-based Violence.

Throughout Section 5.2, the Bracelet hardware and firmware architectures have been carefully dissected to show in detail every digital processing block, both those that are completely closed at the implementation and integration level and those that are still in the design and development phase. Within this context and focusing on the wearable aspect, this Section presented a PPG SQA system able to identify segments of physiological information with poor quality, however, it did not present any work related to the motion artefact correction or removal of such segments. When looking for such type of systems in the scientific community and regardless of the PPG SQA system proliferation [243], there are plenty of Motion artefact

Removal (MAR) systems proposed in the literature. This fact is based on the craving for recovering the whole PPG signal, disregarding the type and amount of noise present. Even different initiatives and challenges were created along the last decade to foster the development of new MAR algorithms and methods. For instance, the IEEE Signal Processing Cup in 2015 was based on a laboratory captured PPG dataset using a treadmill to generate different types of movement artefacts and intended to deliver a general framework to deal with MAR in heart rate monitoring [272]. On this basis, there are two different perspectives. On the one hand, when focusing on offline use cases, the application of MAR techniques can be feasible by considering an available high amount of computing resources. However, these algorithms consider the whole signal for their processing pipelines, and in some cases the reconstruction or extraction of a valid measurement from a noisy signal is not achievable. On the other hand, wearable applications targeting continuous PPG monitoring are subjected to the requirement for low resources usage and low power consumption. This leaves a thin gap to implement some of the best and heavier computational MAR algorithms, such as independent component analysis, empirical mode decomposition, and deep learning based methods [273]. Therefore, the signal quality assessment through SQA methodologies prior to the application of any MAR algorithm is essential when aiming at continuous monitoring of PPG and other cardiac-based related signals [274]. Notwithstanding the latter fact, research related to the proposal of novel embedded MAR techniques has been initiated within our research group [232], intending to contribute to this particular research topic.

Regarding the detailed feature extraction processes, certain limitations of the proposed system must be considered.First, different signal processing techniques can be applied. For instance, to deal with the unevenly spaced HRV data, Lomb-Scargle periodrogram method could be applied [275] instead of FFT. Second, specific power consumption for every of the feature extraction techniques needs to be properly analysed towards the identification of possible bottlenecks within the digital signal processing architecture. The latter is currently being performed and further publications are on preparation.

Chapter 6

A new dataset for emotion recognition: WEMAC

In Chapter 4, we presented the work realised towards the design of a fear detection system by using publicly available databases. Different limitations were identified while developing these systems and it was confirmed that in order to come up with an optimal fear recognition system, a novel database focused on fear detection is required. Such database should include key factors already highlighted in previous chapters, such as:

- The usage of emotional immersive technology.
- The labelling methodology modification to consider the gender perspective.
- A properly balanced stimuli distribution regarding the target emotions.
- A greater number of participants.
- The integration of a recovery process based on the physiological signals of the volunteers to quantify and isolate the emotional activation between stimuli.

Moreover, targeting one of the main goals of this research, i.e. the generation of new prevention and combating Gender-based Violence mechanisms, this database must be conceived by considering the necessary particularities related to this specific profile in order to carry out a proper methodology design. Within this context, this Chapter presents the UC3M4Safety database, whose final objective is the unravelling of the activation mechanisms of Gender-based Violence Victims under violence situations. This goal is intended to be accomplished by the generation and performance of different experiments:

- Pre-labelling experiment. This generated the first two datasets, which are published in [276] and [277]. The aim was to study and validate the effectiveness of a set of audiovisual stimuli when it comes to generating discrete, concrete and unique emotions. This experiment focused into finding stimuli that were able to provoke the same emotional reaction to the largest number of people as possible. In addition, this study allowed us to analyse the methods for classifying these emotional states, the understanding of critical aspects about by the participants, and the influence of gender on the detection of fear [82].
- Laboratory experiments with Non Gender-based Violence Victims. These experiments generated four datasets. The first release is denoted as "Women and Emotion Multi-modal Affective Computing dataset" (WEMAC). They consist of experiments performed in a laboratory environment with only women volunteers who never experienced Gender-based Violence. Specifically, a reduced set of stimuli, which were extracted from the first datasets, are used together with physiological and physical (voice and audio) information acquisition. Apart from the undoubtedly value, for the affective computing area, of generating a dataset with emotional reactions in women while recording their physiological and physical variables, the fear-like emotions disentanglement by means of monitoring physiological and physical reactions from women that are not Gender-based Violence Victims is also necessary to understand these variations and patterns under non-specific population profiles[1]. This research deals with these datasets focusing on the physiological and multi-modal data.
- Laboratory experiments with Gender-based Violence Victims. At the moment of this PhD report, these experiments are still under development. They will generate four additional datasets. They are based onto the same experimental methodology followed with the Non-Gender-based Violence Victims. It should be highlighted that special attention has been paid to avoid re-victimisation of Gender-based Violence Victims, including psychological monitoring and working in their emotional recovery from violence. In these experiments, the goal is to compare the physiological and multi-modal responses to the previous experiments with Non-Gender-based Violence Victims.

[1]Here, non-specific population profiles are referred to as those that have not suffered gender-based violence nor are under post-traumatic stress conditions.

- Into-the-Wild experiments with both profiles, Gender-based Violence Victims and Non-Gender-based Violence Victims. To date, these experiments are under development. They will generate at least ten more datasets. These experiments are intended to be performed during the daily life of some of the volunteers that were involved in the laboratory experiments. The goal is to get a real physiological and multi-modal behaviour to further study and characterise the labelled emotions along the different days.

Regarding the structure of this Chapter, we start by providing a comprehensive explanation of the methodology and development followed during the generation of the laboratory experiments for the Non-Gender-based Violence Victims. As in Chapter 4, the analysis of the self-reported labelling distribution is also presented for this dataset. This analysis is followed by a physiological response exploration to give a proper insight into the physiological patterns, recoveries, and other particularities observed during the experiments. This exploration is concluded by presenting the first fear detection results based on such information. The obtained metrics are used and fused together with the speech results, which gives a multi-modal perspective of the problem. The latter was done in a multidisciplinary research work with the members of UC3M4Safety experts in Signal Theory and Communications. Note that these are the first fear recognition results using multi-modal information recollected from our database and they have already been presented in [9].

6.1 Methods, Tools and Stimuli

As already mentioned, this Chapter uses the information recollected in the WEMAC dataset. The whole dataset contains a total of 104 women volunteers that were exposed to 14 validated audiovisual emotion-related stimuli. This dataset is intended to be publicly available throughout different releases. In particular, this research uses the data contained within the first release, accounting for a total of 47 out of the 104 volunteers. A total number of 123 experiments were done, from which 104 recordings were considered valid (no sensors malfunctioning). All of them were performed between October 2020 to July 2021.

Regarding the methodology designed for the experimentation, Figure 6-1 shows a simplified diagram for every volunteer and stimulus displayed. The Ethics Com-

mittee of Universidad Carlos III de Madrid approved this protocol regarding Ethics aspects and Data Protection aspects. Prior to the experiment, the recruited volunteers are explained all the different steps to be followed and given a set of documents, such as an informed consent, personal data processing, and a general questionnaire. As specified in Section 2.3.3, this questionnaire can provide additional information related to cognition, appraisal, attention, personality traits, gender, and age. The collected data were: age group, recent physical activity or medication that can alter the physiological response of the participant, self-identified emotional burdens due to work, economic and personal situation, and mood bias (fears, phobias, traumatic experiences). Note that the multivariate analysis of these factors is out of the scope of this research.

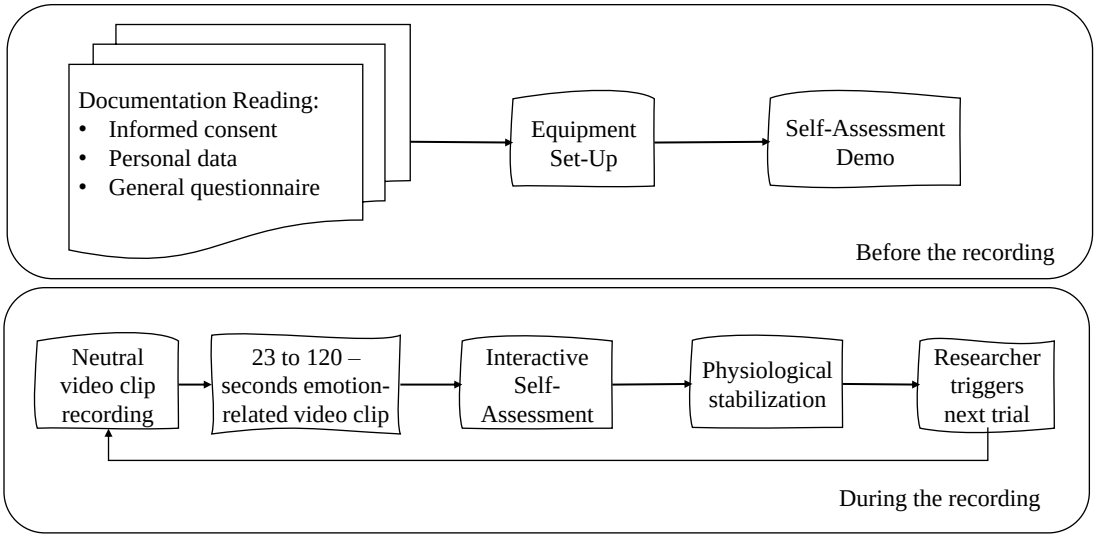


Figure 6-1: Experimental methodology followed during the development of the WEMAC dataset. Prior and during the experimentation.

Following the documentation reading, the equipment set-up is performed and consists of:

- Oculus® Rift-S Headset[2] used to maximise the immersive experience and, consequently, get a better emotion elicitation. This is the platform through which the different stimuli are displayed to the volunteers both in 2D and 3D format.
- BioSignalPlux®[3] standard measurement system employed to acquired different physiological signals. In particular, these were: finger PPG, ventral wrist GSR, forearm SKT, trapezoidal electromiography, chest respiration, and wrist

[2]https://www.oculus.com/rift-s/

[3]`https://biosignalsplux.com/products/kits/researcher.html`

inertial movement by means of an accelerometer. It provides a golden measure to be compared with the rest of the sensors', such as the bracelet of Bindi. In fact, the PPG and GSR signals obtained from BioSignalPlux and Bindi have been previously compared and successfully correlated in publications [159] and [236].

- Bindi's bracelet with dorsal wrist PPG, ventral wrist GSR and SKT. The hardware and software architecture and detailed design of this element are detailed in Chapter 5.
- An additional GSR sensor to be integrated in the following iteration of the bracelet (Bindi 2.0). The hardware and software design of this new sensor are out of the scope of this document, although a publication is in progress [237].

The synchronisation of all the different sensors acquisition together with the stages of the experiment is performed by a laptop running a Unity® framework based program. This work was done by the UC3M4Safety team. Note that all the devices sensing physiological information were running at a 200 Hz sampling frequency.

Finally, the last step in the preparation of the experiment is a self-assessment labelling demo, in which the volunteers get used to the virtual reality environment and know better the different labelling categories and particularities. The elements compiled during the labelling self report process are, in order of appearance:

- Audio sample recorded throughout the microphone in the Oculus® Headset, right after the emotion-related video-clip visualisation. The volunteers are requested to relive the emotions felt during the emotion-related stimulus visualisation. For this research, it is assumed that the correspondence is solid enough between both instants. Note that, although this last assumption can be considered as a simplification to be applied for a first data manipulation, it will need further validation in future works.
- Modified SAM manikins for PAD affective dimensions mapping, as detailed in Section 2.4 of Chapter 2.
- Familiarity level with respect to the felt emotion and to the displayed situation in the video-clip. Both were asked using a 9-point Likert scale as for the SAMs.
- Liking of the video with three possibilities: yes, neutral, no.
- Selection of one discrete emotion out of a total of twelve. They were obtained

from the pre-labelling study performed by the UC3M4Safety team, which used the first two datasets and was published and detailed in [82], [278] and [279].

During the experiment, as already detailed, every volunteer visualised a total of 14 2D or 360º audiovisual stimuli. These stimuli were obtained considering 28 audiovisual stimuli from a larger stimuli pool that contains a total of 42 stimuli validated by more than 1332 people (811 females, 521 males) during the first detailed experiment of the UC3M4Safety database [57, 82]. Note that these 28 audiovisual stimuli were selected based on three main premises: the highest emotional discrete labelling agreement observed in women during the pre-labelling experiment, targeting for an adequate laboratory experiment duration, and a balanced distribution of fear vs no-fear by considering a PA model as performed for the stimuli selection of the MAHNOB database explained in Section 4.2.1 of Chapter 4. Thus, two different batches can be applied, with 14 stimuli per batch. Such amount of audiovisual stimuli together with the documentation reading, equipment set-up, and self-assessment demo, used to take from 1 to 1.5 hours per volunteer, while data processing implies from 3 to 8 hours. Table 6.1 reports the ordered structure of these batches.

Stimulus	**Emotion**	**Quadrant (PA)**	**Length**	**Format**	**Batch**
1	Joy	1	1'26"	2D	1
2	Fear	2	1'20"	3D	1
3	Sadness	3	1'59"	2D	1
4	Anger	2	1'03"	3D	1
5	Fear	2	1'35"	2D	1
6	Calm	4	1'	3D	1
7	Anger	2	1'	2D	1
8	Fear	2	23"	2D	1
9	Disgust	3	40"	2D	1
10	Fear	2	2'	3D	1
11	Joy	1	1'41"	2D	1
12	Fear	2	1'20"	2D	1
13	Gratitude	4	1'40"	2D	1
14	Fear	2	1'27"	2D	1
15	Fear	2	1'52"	2D	2
16	Joy	1	1'28"	2D	2
17	Fear	2	46"	2D	2
18	Sadness	3	45"	2D	2
19	Fear	2	1'33"	3D	2
20	Calm	4	1'	2D	2
21	Anger	2	1'59"	2D	2
22	Fear	2	1'14"	2D	2
23	Disgust	3	1'36"	2D	2
24	Fear	2	2'	3D	2
25	Surprise	1	1'41"	2D	2
26	Fear	2	1'06"	2D	2
27	Gratitude	4	1'30"	2D	2
28	Fear	2	1'59"	3D	2

Table 6.1: List of audiovisual stimuli used within the WEMAC Dataset.

For the first batch, the stimuli duration are in 1'32"±46", while for the second batch the duration are in 1'46"±44". It can be observed that both batches have 8 stimuli belonging to the second PA quadrant, which was done on purpose to maintain a proper balance between fear-like and non-fear-like emotions. Note that the balance premise considers the PA model, rather than the PAD, to ease and simplify such task. Due to this fact, the stimuli pre-labelled as anger were considered as well within the second quadrant, and so being within the positive class for the dimensional ground truth labelling. Note that all the volunteers in the same batch visualised them in the same order.

Before the presentation of every emotion-related stimulus, a specific neutral clip is also used to ease the emotional recovery. These are randomly selected from a larger pool provided by the Stanford psycho-physiology laboratory [197]. In the same way but at the end of the self-assessment, 360º recovery scenes are also presented. These are selected by unanimous consensus of the research team. The main difference between the neutral and the recovery clips is that while the former are totally passive, i.e. there is not a recovery monitoring, the latter actually implements a physiological monitoring. This allows for the online assessment of the three measured variables stabilisation. Such process is performed by using the physiological measurements acquired by Bindi's bracelet. Specifically for these first experiments, we implement a physiological recovery stabilisation controller into the SoC of the bracelet, which worked based on segmented temporal data processing windows. Such system performs an online basic filtering process of the signals, extracted the BPMs from the calculated heart rate, and verified the stabilisation of the signals for more than four consecutive processing windows. Once the stabilisation has been achieved at least by two out of the three variables, the bracelet notified via BLE to the host computer running the virtual reality framework. Note that the physiological recovery implementation was also followed by a Bachelor Thesis under my supervision [280]. Its goal was to implement new recovery mechanisms towards the improvement of the current one during the experiments. The new implemented features and improvements were even applied to other projects being developed within the research group.

6.2 Self-Reported labelling response exploration

As detailed in the previous Section, the first release of the WEMAC dataset contains data from 47 volunteers. Specifically, 32 and 15 volunteers visualised the first and second batch, respectively. Due to the different labelling methodologies considered and based on the previous works using the public benchmark data explained in Chapter 4, both discrete and dimensional labelling has been employed. Note that both are binarized to provide a fear-like binary classification problem. Thus, all the discrete labels that were not identified as fear are codified as the negative class, while the ones assessed as fear are set as the positive class. The same process is performed for the dimensional labelling methodology, but following the proposed fear binarization method, see Section 2.3.4 of Chapter 2, as done for the public benchmark databases, see Chapter 4.

Figures 6-2 and 6-3 show the binarized discrete and dimensional class balance for the self-reported labels of the 47 volunteers. Note that in these figures, the ground truth class balance per batch is also represented. On average, for the ground truth, both batches possess 53.57% and 46.43% of negative and positive classes, respectively. The average balance for the dimensional self-assessment labels is 55.80% and 44.20%, while for the discrete self-assessment labels is 60.47% and 39.53% for the negative and positive classes, respectively. However, the main difference is obtained when comparing the standard deviation, which is up to 15.22% for the dimensional labels and 7.84% for the discrete labels. Although the averaged class balance of the dimensional self-reports is closer to the golden class balance, its deviation is two times the averaged class balance of the discrete self-reports, which is directly related to the labelling agreement of the different 47 volunteers. Within this context, we also defined a 25.00% threshold to identify the volunteers whose class balance was affected by a 1.5 times (equal or higher) the golden class balance, which we identify as labelling outliers. Note that this is a first approach simplification, as further physiological data analysis might be performed with such outliers to properly characterise their emotional reactions. Thus, volunteers that reached such threshold are marked in brackets. For the discrete labelling were identified up to five volunteers (5, 6, 15, 33 and 40), and for the dimensional labelling were identified up to nine volunteers (3, 5, 6, 13, 20, 21, 22, 40, 42). It should be highlighted the class bal-

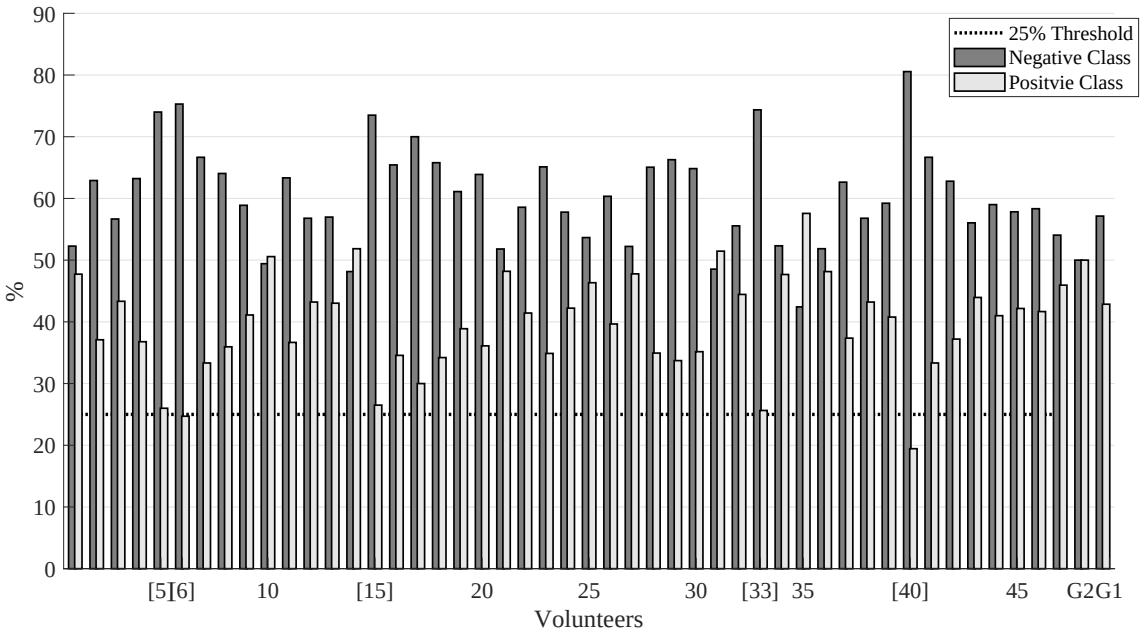


Figure 6-2: Class distribution for binary fear mapping over the discrete subjective self-reports in WEMAC for all the 47 considered female volunteers, and the original intended class distribution of the experiment: G2 and G1 for the second and first batch, respectively.

ance differences observed per volunteer when considering both methodologies. For instance, the third volunteer shows a 57/43% balance approximately for the discrete labelling, and a 85/25% balance for the dimensional assessment. This fact suggests a different comprehension and understanding of each of these labelling methodologies, which can lead to different machine learning systems when using one or the other. This is the main reason that led us to work with both approaches.

Following the same schema analysis for this dataset as the one applied to the public benchmark databases in Chapter 4, the label inter-individual correlations have been assessed to check if all the volunteers are labelling every emotion-related stimulus. In this case, the results obtained after a Levene's test and a Kruskal-Wallis test for the binarized discrete labelling provided different results. The former rejected the null hypothesis that the variances are equal across all volunteers ($p<0.001$), while the later failed to reject it ($p>0.001$). Conversely, the same methods rejected the null hypothesis for the binarized dimensional labels ($p<0.001$). This difference is a consequence of the final conclusion extracted from the previous figures by indicating that, at least for the discrete labelling methodology, there is not enough evidence to claim the variances are different across volunteers. Thus, this fact suggests that each of the methodologies is characterising different aspects of the emotions, which is in

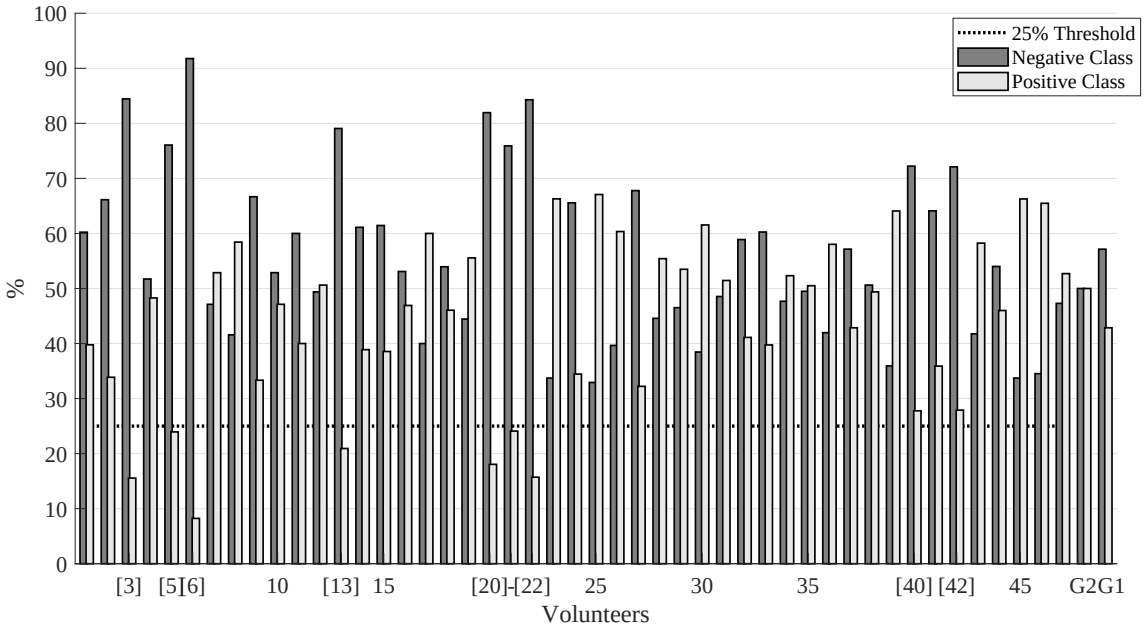
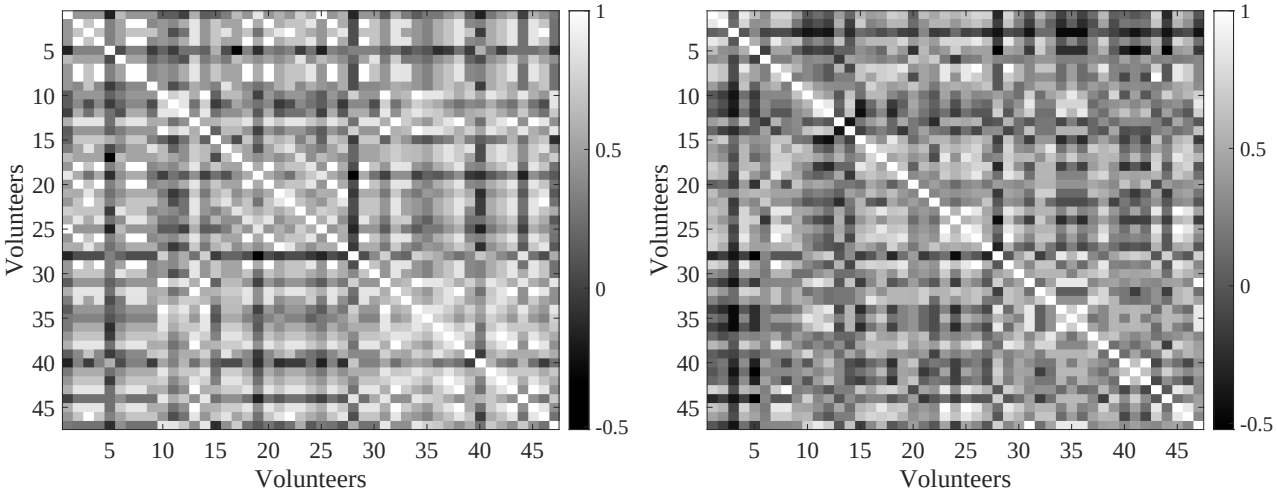


Figure 6-3: Class distribution for binary fear mapping over the dimensional PAD subjective self-reports in WEMAC for all the 47 considered female volunteers, and the original intended class distribution of the experiment: G2 and G1 for the second and first batch, respectively.

line with the information provided in Chapter 2 when stating that both models exist but each is intended to explain different features of emotions [48]. Note that the set of binarized labels exhibit a non-normal distribution and that the significance level was set at $p<0.05$.

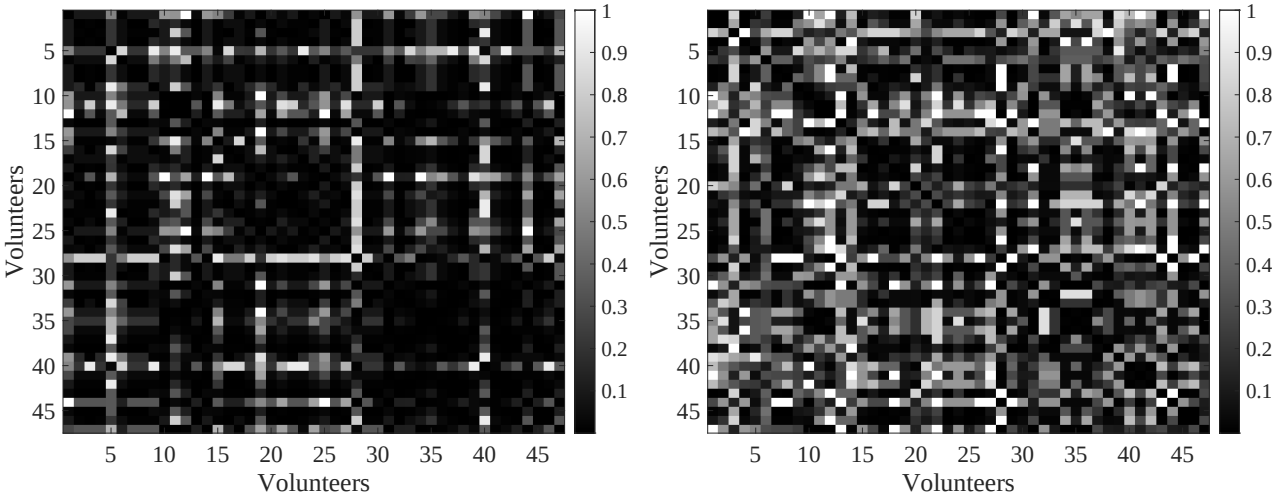
After the variance analysis, the Spearman correlation is also applied for this dataset. However, due to the previous observed difference, the non-averaged matrices are showed to graphically demonstrate the effect and actual inter-individual consequence. Specifically, Figure 6-4 presents the inter-correlation across the 47 volunteers for both methodologies. These matrices provide a one-to-one subject information regarding the labelling differences. When comparing the inter-correlation matrices, we can spot some common regions within both of them. Although it can be observed that the discrete inter-correlation matrix possess a lighter grey colour, which indicates that the correlations are slightly more positive, there is not a clear differential conclusion by just analysing solely these matrices. Therefore, the p-values corresponding to such correlation matrices are plotted in Figure 6-5. In this case, there is a clear distinction between both methodologies. The discrete labelling shows a black colour identification for the majority of the volunteers, which indicates a p-value lower than 0.1. Conversely, the dimensional self-assessment does not reports



(a) (b)

Figure 6-4: Spearman one-to-one subject inter-correlation across the 47 volunteers for both labelling methodologies: a) discrete, and b) dimensional (PAD).

such behaviour. This fact supports the previous conclusions and suggests that the association or agreement between the fear binary labels of the volunteers within the discrete case is stronger than with the dimensional methodology. Please, to contextualise this analysis, note the following two considerations: a) the volunteers from both batches were used indistinctly, and b) this analysis serves as a preliminary study to assess the agreement within the same methodology and the differences with respect to both of them, however, it can be further continued by digging into specific one-to-one volunteers differences and/or even by applying different statistical methods.



(a) (b)

Figure 6-5: P-values obtained from the Spearman one-to-one subject inter-correlation across the 47 volunteers for both labelling methodologies: a) discrete, and b) dimensional (PAD).

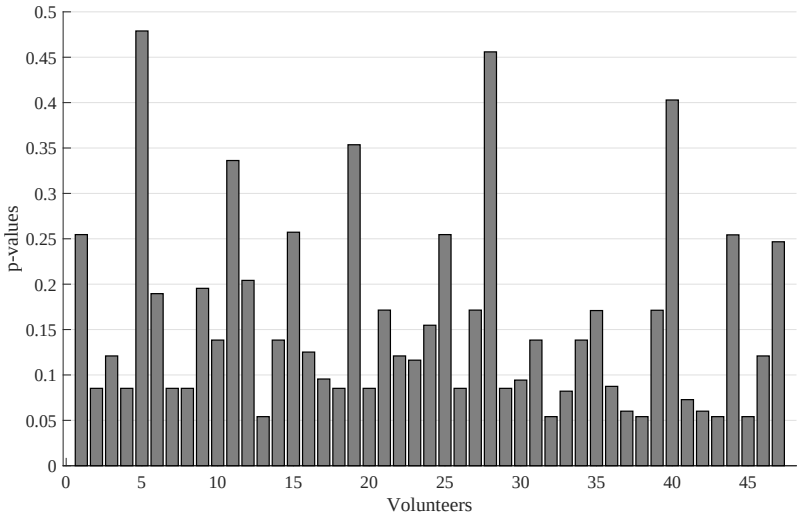


Figure 6-6: Averaged $p-values$ for all considered volunteers and their labels applying the Spearman correlation for their PAD-based fear binary mapping labels.

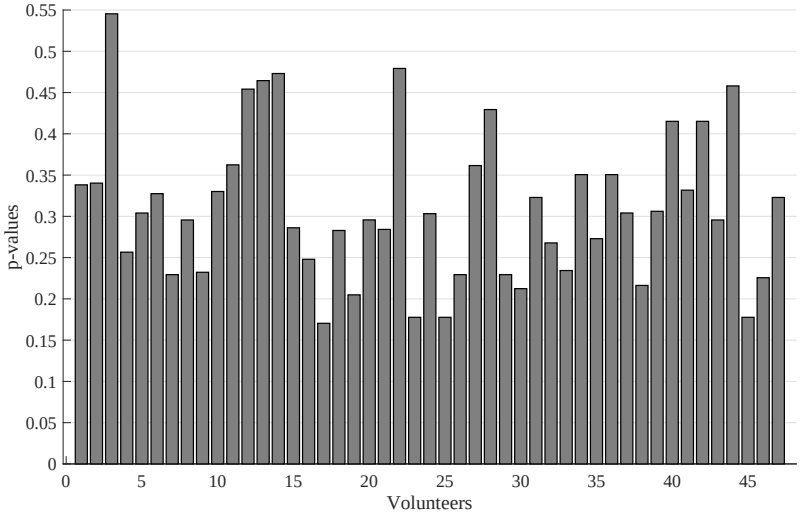


Figure 6-7: Averaged $p-values$ for all considered volunteers and their labels applying the Spearman correlation for their discrete-based fear binary mapping labels.

In order to provide an individual averaged perspective for the agreement, Figures 6-6 and 6-7 are shown. The obtained results fail to reject the null hypothesis on average for each of the 47 volunteers, which indicates that there is not enough strong evidence to suggest that an association between the fear binary labels of the volunteers unequivocally exist. For instance, the mean p-values are 0.15 and 0.31 for the discrete and dimensional cases, respectively. Thus, although it should be noted that this is an averaged result, the extracted conclusion is in line with the previous ones. Note that the volunteers showing the highest averaged p-values are generally those that possessed the labelling inconsistencies reported in Figures 6-2 and 6-3.

The analysis provided in this section suggests that the discrete labelling methodology outperforms, in terms of agreement, the dimensional assessment. This con-

clusion does not strictly mean that a system trained separately with both labels is going to have the same difference in performance, as the self-reported labels are always affected and biased by cognitive processes, unlike the physiological responses as detailed in Chapter 2. Thus, the different results gathered from this stimuli balance and label consideration study need to be further contextualised also when evaluating the obtained results for the machine learning systems based also on both labelling methodologies.

6.3 Physiological response exploration

In this section, we carry out a physiological response exploration to give a proper insight into the physiological patterns, recoveries, and other particularities observed during the experiments. Moreover, those are concluded by presenting the first fear detection results based on such information using the discrete and dimensional self-reported labels for the 47 volunteers. The obtained metrics are used and fused together with the audio results in the next section.

The signals used along this response exploration were properly filtered and denoised. The BVP signals were subjected to the same filtering strategy as described in Section 4.1.2, and also to a forward-backwards low-pass Butterworth IIR filter to deal with the baseline wander residual. For the GSR and SKT signals, a basic FIR filtering with 2Hz cut-off frequency was applied. After that, this filtered output was down-sampled to 10Hz and also processed with both a moving mean and a moving median filters. The former used a 1-second window and helped reducing the high noise residual after the initial FIR, while the latter employed a 0.5-second window and dealt with the rapid-transients.

All the presented physiological analysis and results in this section are done by using the signals acquired by the BioSignalPlux® research toolkit system. This decision is considered towards obtaining comparable results for the different physiological analysis and machine learning systems proposed with respect to literature. This fact is essential to further be able to replicate the same analysis for the other sensor systems employed in the experiments and evaluate the differences. Although the latter task in not within the scope of this research work, the physiological signal acquisition verification and validation with Bindi and the BioSignalPlux® research

toolkit system has been already performed and published in [159, 236].

6.3.1 Physiological patterns and recoveries

The work presented in Chapters 4 and 5 that dealt with all the digital signal processing stages, such as filtering and feature extraction, allowed us to provide a deeper analysis into the physiological response within our own dataset (WEMAC). The physiological exploration is a challenging task when considering this type of experiments. This fact is mainly affected by the complexity of emotions and by the uncertainties or intrinsic physiological variations due to intra and inter individual differences. Thus, for the sake of simplicity, we perform a preliminary physiological exploration in this section by considering some of the reviewed signals and a reduced set of features. Specifically, for the physiological pattern analysis, we used the GSR signal extracted during the emotion-related stimuli visualisation to determine the degree of similarity considering signals between the same and other volunteers. Regarding the physiological recovery analysis, specific features extracted from the GSR and BVP signals were used to provide a detailed comparison between the recovery and the emotion-related stimuli visualisation stages. These analysis provided useful insights concerning the expected and actual physiological responses of the volunteers. Moreover, they can be also expanded by studying the complete set of physiological signals and features.

6.3.1.1 Pattern analysis

For the physiological pattern exploration, as already commented, the GSR signal behaviour was analysed. The selection of this physiological signal was based on the direct relation that it has with emotional responses, as reviewed and studied previously in Chapters 2 and 4.

In this study, the pattern exploration was done for all the 47 volunteers by using a common time series pattern analysis technique called Dynamic-Time-Warping (DTW) [196,281–283]. This technique allows quantifying the similarity between two time series with equivalent features even when having different velocities or phase space trajectories. For instance, the GSR signals amongst the different volunteers exhibit such behaviour. Figure 6-8 shows the GSR signals for volunteers 4, 15, and 27, extracted during the visualisation of one of the fear stimuli. There can be seen

different phasic peaks locations, some within the same temporal interval for the three volunteers and some in totally different instants. Note that the behaviour and dynamics of the signal depends mainly on the type of emotion-related stimuli and the volunteer (intra-individual factors as detailed in Chapter 2). Each of these signals, which are acquired at regular intervals, can be defined as:

$$S^{i,j} = (s_1^{i,j}, s_2^{i,j}, s_3^{i,j}, ..., s_N^{i,j}), \tag{6.1}$$

where i and j are the volunteer and the stimulus, respectively; and $s_k^{i,j}$, with $k \subseteq [1, N]$, are the different acquired samples for the entire emotion-related stimulus duration. Thus, DTW finds the optimal signal-to-signal distance measure, following some restriction rules, highlighting the similarities between the signals and providing a measurement of their similarity regardless of non-linear variations. Specifically, a cost function is used to assess the dissimilarity between all the samples of both time series being compared. In our case, the cost function is given by the Euclidean distance following

$$d_{mn}(S^{i,j}, S^{q,p}) = \sqrt{\sum_{m,n=1}^{K} (s_m^{i,j} - s_n^{q,p}) * (s_m^{i,j} - s_n^{q,p})}, \tag{6.2}$$

where m and n are the specific samples for each time series. Note that j and q can be the same or different stimuli. The results obtained with this operation are arranged into a cost matrix, which is used to find the warping or optimal path. Once such path is found, the final result is the total cost or distance, which is directly related to the similarity between both sequences, given by

$$d_{min}(S^{i,j}, S^{q,p}) = \sum_{m,n \in K} d_{mn}(S^{i,j}, S^{q,p}). \tag{6.3}$$

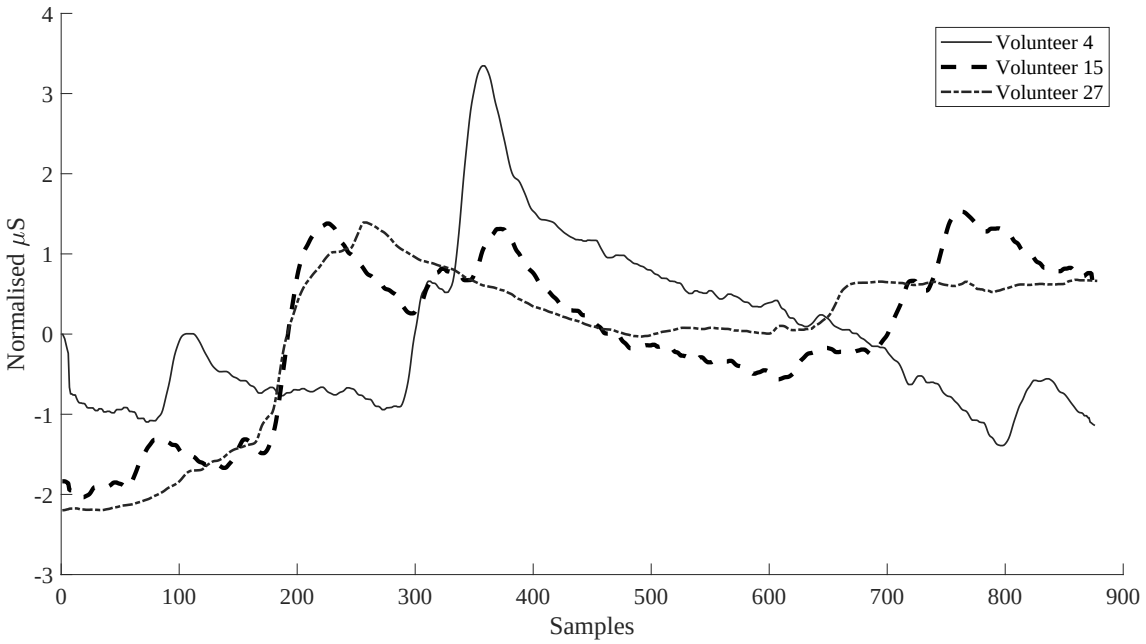


Figure 6-8: GSR signals extracted from the whole visualisation of the sixth stimulus from the first batch (last one stimulus) of volunteer 4, 15, and 27.

Therefore, in case of the GSR signals depicted for the three volunteers in this example, the obtained distances are:

$$\begin{aligned} d_{min}(S^{4,6}, S^{15,6}) &= 585.23 \\ d_{min}(S^{4,6}, S^{27,6}) &= 549.13 \\ d_{min}(S^{27,6}, S^{15,6}) &= 172.87, \end{aligned} \tag{6.4}$$

which indicates that volunteers 15 and 27 possess higher similarities than the other two volunteer combinations. Thus, for this example, 2 out of 3 examined volunteers present a similar physiological pattern, behaviour or dynamics regarding this specific stimulus.

Within this pattern analysis context, three different pattern clustering use cases were tackled based upon the separation or combination of the different batches. For all the use cases, the different analysed GSR segments were normalised (Z-score) and compared against each others. Note that in this case, each segment is referred as to the extracted GSR signal for every complete emotion-related stimulus. Previously to studying the individual segment-to-segment pattern analysis, we generated averaged matrix visualisations and aggregated DTW plots as shown in Figures 6-9 and 6-10, respectively. The former gives an insight regarding the averaged pattern similarity for all volunteers and the entire experiment, i.e. the total distance for each

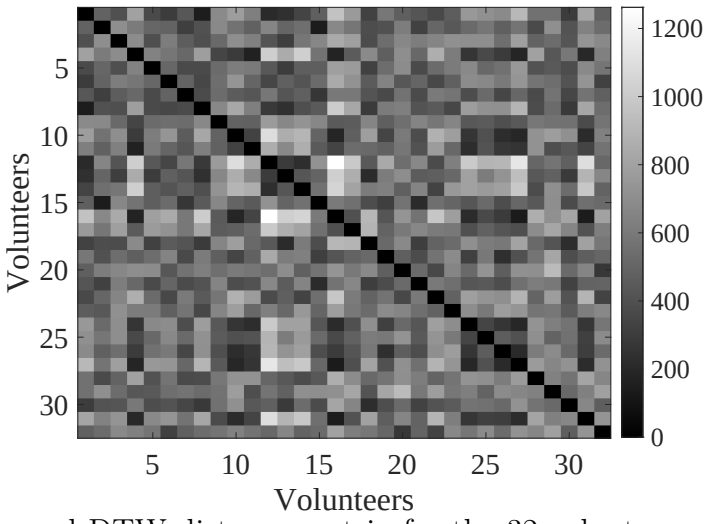


Figure 6-9: Averaged DTW distance matrix for the 32 volunteers visualising the 6 fear stimuli from the first batch of emotion-related stimuli.

volunteer-to-volunteer is calculated by averaging the set of distances obtained for each comparison that considered the whole experiment. It can be seen how the black spots on the matrix are the ones presenting the higher similarity (e.g., diagonal). This matrix can serve as a tool to assess an initial graphical perspective regarding pattern clustering. In this case, Figure 6-9 depicts the averaged matrix values for the 32 volunteers from the first batch and the 6 fear stimuli. Although some dark spots can be sighted within the matrix (e.g., volunteers 8-1, 15-2, 31-16, etc.), we cannot conclude that a pattern clustering formation exist. In case of aggregating all the volunteer-to-volunteer distances within the matrix and obtaining the mean and standard deviation (we omit the diagonal part), Figure 6-10 can be reported. This gives a macro perspective for each volunteer behaviour in comparison with the other volunteers. Thus, we can conclude that, on average, there are no extremely deviated volunteers. However, we cannot claim the existence of pattern formations either. Note that, for the sake of simplicity, only this matrix and plot are shown, but the rest of the use cases were also analysed and led to the same conclusion.

After performing the previous analysis and towards a quantification of segment clustering, subject-dependent and subject-independent leave-one-segment-out clustering studies were performed. First of all, the similarities (distances) for every GSR segment with respect to the rest of the segments from the same and different volunteers were extracted. Secondly, considering the set of gathered distances, the minimum was found. Finally, the current segment being processed was assigned to

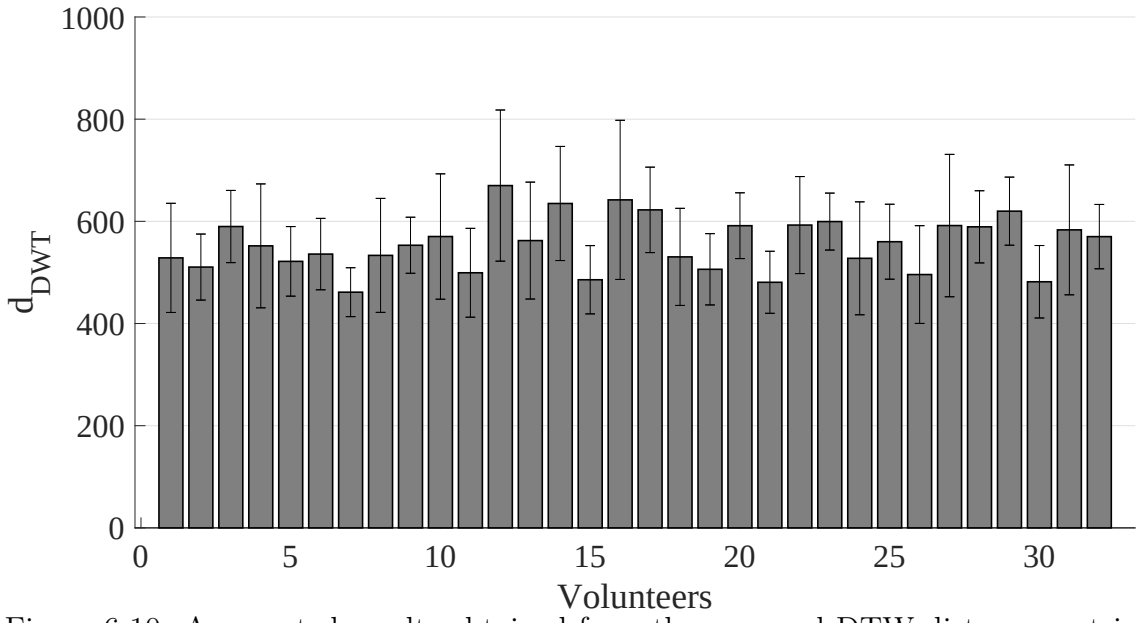


Figure 6-10: Aggregated results obtained from the averaged DTW distance matrix for the 32 volunteers from Figure 6-9.

the same label as the one of such minimum. Note that for this analysis, we considered the binarized ground truth labels as the ones that are expected (non-self-reported). This fact was based on assessing the feasibility of finding physiological pattern clustering without any subject-specific cognitive-biased information. Table 6.2 reports the results after carrying out the clustering studies for the different use cases. It can be observed that the highest results for the three compared metrics are obtained always when considering the subject-independent clustering approach. Note that this fact can also be affected by the data set size, as the amount of data for the subject-dependent clustering is considerably less than for the subject-independent. Overall, there are more similar physiological segments within the non-fear stimuli (specificity). In fact, the clustering of the fear stimuli segments does not surpass the 50.00% detection threshold (sensitivity). The highest difference between both batches is reported for the subject-independent clustering with up to 47.92% and 29.50% sensitivity for the first and second batch, respectively. Again, this fact can indicate a clear physiological response difference regarding the fear stimuli effect, however, it needs to be contextualised to the lesser amount of volunteers being evaluated for the second batch. Finally, the highest sensitivity is achieved when considering both batches together, although at the expense of obtaining the lowest specificity for the subject-independent clustering. Notwithstanding to the latter fact, the Gmean for such use case is one of the highest with up to 53.45%. Thus,

after this study we can conclude different key aspects regarding the physiological responses due to the non-fear and fear related stimuli as well as the overall physiological clustering:

- Analysing the GSR signal similarities between the different volunteers does not achieve a proper distinction between fear and non-fear related stimuli. Note that this analysis can be further extended to explore other signals as well as different pattern exploration techniques.
- The non-fear related stimuli clustering is better characterised or identified. This fact is given regardless of the considered batch.
- The observed sensitivity when considering both batches together is the highest one. This implies that there are fear related stimuli evoking similar physiological responses regardless of the batch and their specific audiovisual content. Thus, in this research we apply an agnostic-batch perspective by considering that both batches can be used together to further design a more efficient fear detection machine learning system.
- The low performance metrics for the different clustering use cases indicate that this information is not enough to unravel and distinguish the fear related physiological activation mechanisms. Thus, more signals and/or features might be exploited towards achieving such goal.

Clustering Type	Batch Number	Segment identification metrics		
		SPE	SEN	Gmean
Subject-dependent	1	54.29%	43.23%	48.45%
	2	48.57%	48.57%	48.57%
Subject-independent	1	61.33%	47.92%	**54.16%**
	2	**64.76%**	29.50%	43.70%
	1&2	57.34%	**49.83%**	53.45%

Table 6.2: Leave-one-segment-out clustering study for both subject-dependent and subject-independent. SPE: specificity, SEN: sensitivity, Gmean: geometric mean.

6.3.1.2 Recovery analysis

For the physiological recovery analysis, as previously explained in Section 6.1, an online stabilisation evaluation of the three different physiological signals being ac-

quired by Bindi's bracelet was performed during the experiments. This online process operated every ten seconds performing an online basic filtering, extracting the BPMs from the BVP signal, and finally assessing the BPMs, GSR, and SKT stabilisation for more than four consecutive processing windows. The latter process was done by hard-threshold adjustment following a 90% level confidence interval with respect to the level of the first window. To analyse the actual effect of these recovery stages, a posterior physiological study was performed. Specifically, features extracted from the GSR and the BVP signals were used to provide a detailed comparison between the recovery and the stimuli visualisation stages. On the one hand, the number of ERSCR or phasic peaks, amplitude and rise time are compared. On the other hand, different Poincaré-plots are elaborated to assess the sympathetic status within the recovery stage [284]. Note that this recovery analysis is done considering both bathes.

Figure 6-11 shows the averaged results for the detected peaks during the experiment. A distinction is done by dividing the fear and non-fear related physiological responses, which is also applied for their respective recovery stages. This process was performed considering the whole physiological signal acquired for the different stimuli and recovery stages, i.e. no segmentation was applied. This latter consideration is adopted as the main goal of this analysis is to evaluate the physiological responses. Thus, there is no need for real implementation constraints such as data segmentation. Note also that, for this analysis, cvxEDA algorithm was used, which is detailed in Section 2.5.2. As expected, the peaks detected for all the fear stimuli surpass the ones detected for the non-fear and the recovery stages. Specifically, during the fear stimuli visualisation, an average of 2.30 peaks were detected per stimulus with 0.81 standard deviation, while the non-fear stimuli produced 1.11 peaks with 0.52 standard deviation. One of the key aspects of these results is that the recovery stages are below such metrics for both types of stimuli. This is obtained for both the averaged and the standard deviation values: fear recovery presents a peak average of 0.99 (0.37), and non-fear recovery is up to 1.03 (0.48). To support the peak detection results, their amplitude and recovery time are also extracted and plotted in Figures 6-12 and 6-13. Note that the reported amplitude is obtained as the relative amplitude from the detected peak onset, as well as for the recovery time, see Figure

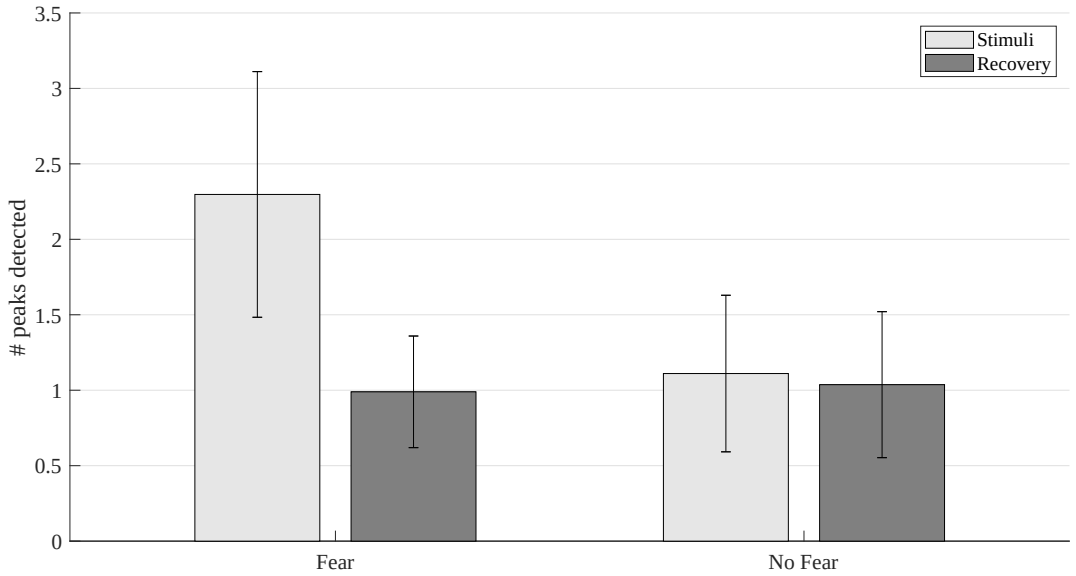


Figure 6-11: Averaged results comparison obtained from the GSR peak extraction process using cvxEDA algorithm for the 47 volunteers and both batches.

2-14. Overall, the same behaviour is observed for these metrics. However, the mean and standard deviation of the relative amplitude for the non-fear recovery exceed the non-fear stimuli metrics. Specifically, the non-fear stimuli reach an average relative amplitude of 0.01uS with a standard deviation of 0.006uS, and the non-fear recovery provides 0.02uS averaged relative amplitude with a standard deviation of 0.01uS. This physiological difference needs to be contextualised together with the recovery time of the extracted peaks, in which we observe exactly the same behaviour as for the detected peaks. Thus, observing this behaviour, we can conclude that, on average, the level of arousal is as expected for the recovery stages in comparison with both fear and non-fear stimuli together. Therefore, the application of the active recovery process implemented reduces the emotional bias between stimulus.

For the BVP analysis, we used a commonly applied tool to assess the sympathetic activation, which is known as Poincaré-plot. This is a recurrence plot in which the consecutive IBIs are transferred to a two-dimensional dispersion diagram to obtain a graphic image of the behaviour of the HRV for a given time interval, Figure 6-14. From this specific graph, different geometric metrics are obtained. Generally, the two most important are the standard deviation along and perpendicular to the line-of-identity, SD_2 and SD_1 respectively. It has been shown that these features can characterise the sympathetic and parasympathetic activation. For instance, having a narrow shape of the main cluster is an indication of dominance of the non-respiratory

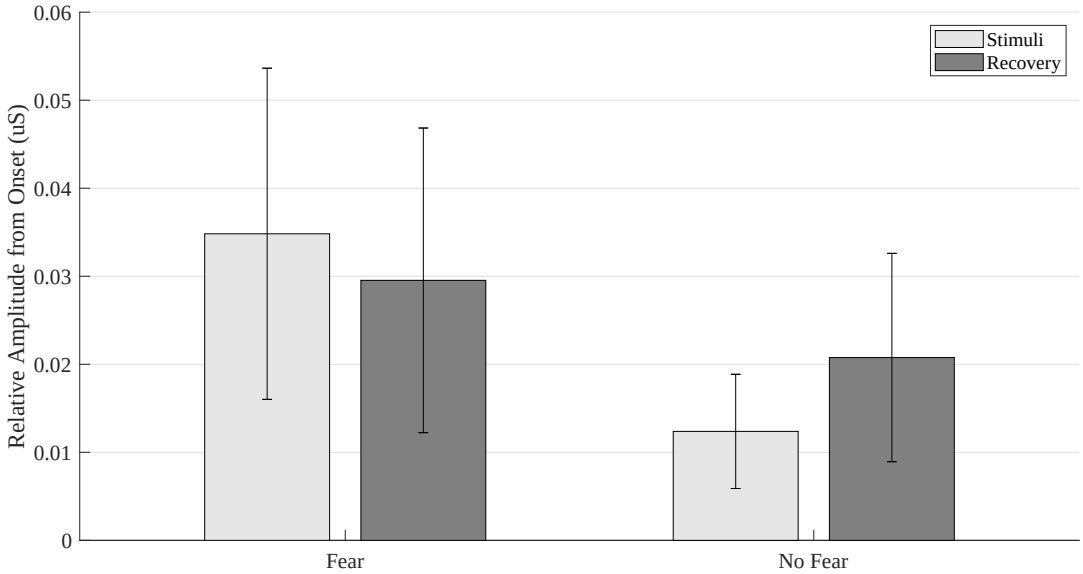


Figure 6-12: Averaged results comparison obtained from the GSR relative amplitude extraction process using cvxEDA algorithm for the 47 volunteers and both batches.

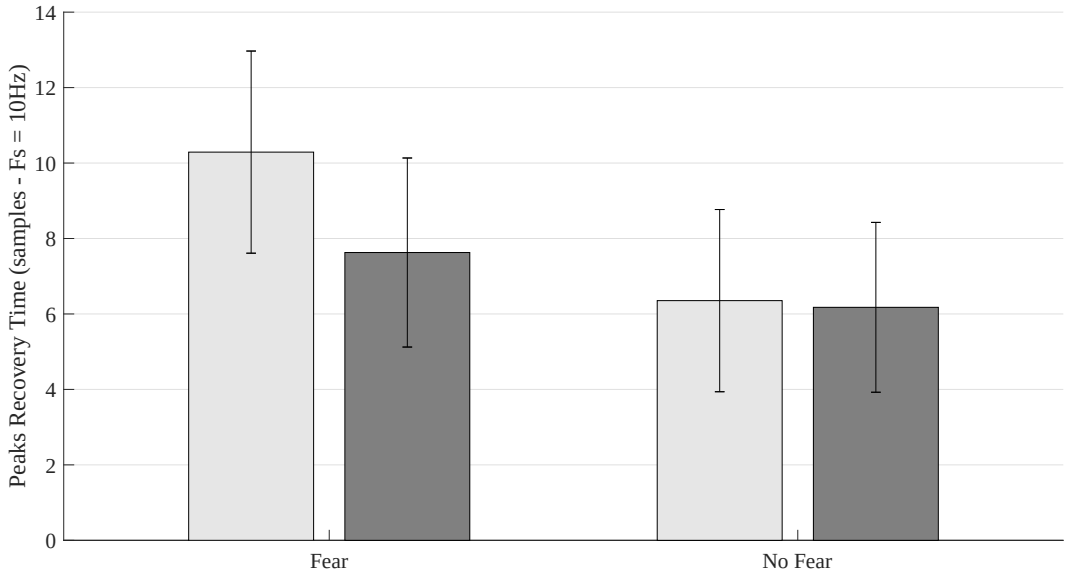


Figure 6-13: Averaged results comparison obtained from the GSR peak recovery time extraction process using cvxEDA algorithm for the 47 volunteers and both batches.

components regulating the heart rate, which is directly related to the sympathetic activation. Conversely, the wider the cluster, the more dominance of the respiratory components, which is related to the parasympathetic prevailing. Moreover, this type of plot allows studying the non-linearity of the cardiac information as well as being insensitive to trends in the IBIs [285–287]. Note that the IBI time series is explained and represented in Section 4.1.3.1 and equation 4.6. The calculation of both standard deviation features has been done following a simplification considering [288–290]. Thus, these are computed as the standard deviation of the time series obtained following:

$$\begin{aligned} SD_2(i) &= (\frac{\sqrt{2}}{2}) * (IBI_i + IBI_{i+1}), \\ SD_1(i) &= (\frac{\sqrt{2}}{2}) * (IBI_i - IBI_{i+1}). \end{aligned} \tag{6.5}$$

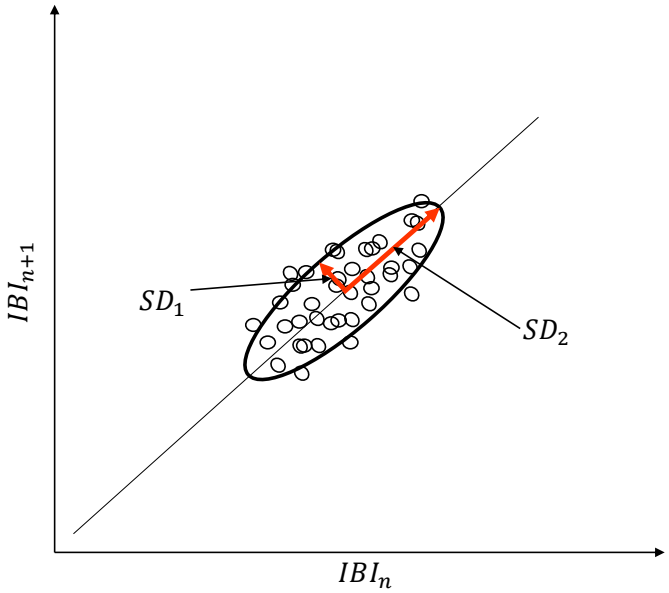


Figure 6-14: Exemplification of a recurrent Poincaré-plot and its standard deviation metrics along (SD_2) and perpendicular (SD_1) to the line-of-identity.

For our experiment, as done with the GSR signal, the whole BVP signal for the different stimuli and recovery stages was considered, i.e. no segmentation was applied. Figure 6-17 shows different perspectives for the three Poincaré-plots obtained: fear and non-fear stimuli, and recovery stages. Note that all the IBI time series for the 47 volunteers are contained within these Poincaré-plots. At first glance, we can observe that the fear points tend to be slightly closer to the bottom left corner. This fact is an indication of lower heart rate variability or higher cardiac rhythm. Moreover,

the recovery points are the ones presenting more dispersion or wider shape, which indicates a parasympathetic dominance.

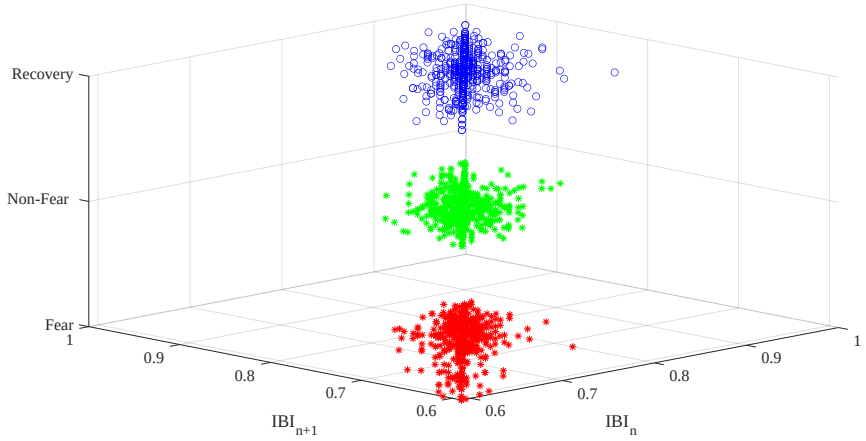


Figure 6-15: Different Poincaré-plots perspectives for all the 47 volunteers considering the fear stimuli (red-bottom), non-fear stimuli (green-middle), and recovery stages (blue-top). Frontal View.

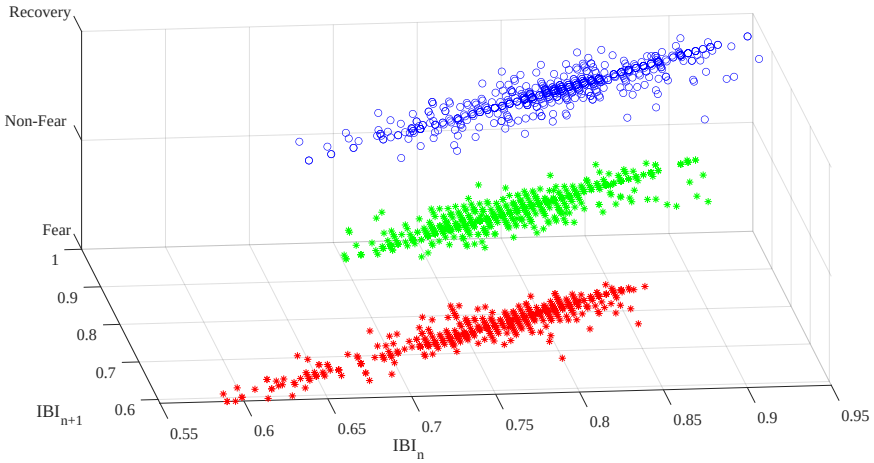


Figure 6-16: Different Poincaré-plots perspectives for all the 47 volunteers considering the fear stimuli (red-bottom), non-fear stimuli (green-middle), and recovery stages (blue-top). Longitudinal View.

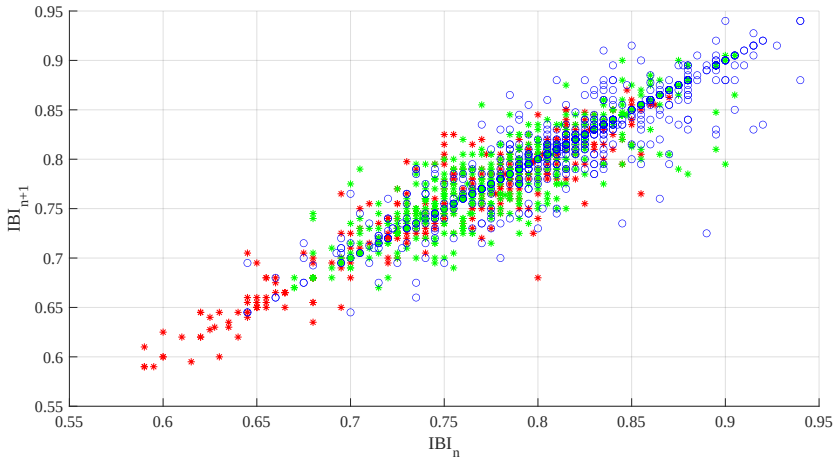


Figure 6-17: Different Poincaré-plots perspectives for all the 47 volunteers considering the fear stimuli (red-bottom), non-fear stimuli (green-middle), and recovery stages (blue-top). 2D View.

The analysis of these type of plots by visual exploration is a challenging task. Thus, Table 6.3 reports the average and standard deviation for the different SD_2 and SD_1 values obtained from this presented Poincaré-plots. Note that in this case the recovery stages are divided as done for the GSR analysis.

Stimuli Type	SD_2 (ms) $\mu(\sigma)$	SD_1 (ms) $\mu(\sigma)$
Fear	62.78 (9.99)	14.28 (2.91)
Fear Recovery	**72.37 (13.11)**	**17.19 (3.47)**
Non-Fear	60.96 (11.19)	14.42 (3.05)
Non-Fear Recovery	**70.57 (11.84)**	**17.22 (3.34)**

Table 6.3: Poincaré-plot features evaluation for the fear and non-fear stimuli, and their respective recovery stages. These metrics are the averaged mean and standard deviation for all the 47 volunteers.

Although the BVP information is not directly related with one emotional dimension, as the GSR is with the arousal, and its analysis is more complex, the obtained results reaffirm some of the physiological effects stated in the previous GSR exploration. For instance, when comparing the obtained features for the fear and non-fear stimuli, we can observe how while the average value of SD_2 decreases from 62.78 ms up to 60.96 ms, the SD_1 increases from 14.28 ms up to 14.42 ms. For the recovery stages, SD_1 increases in both stimuli types, fear and non-fear recovery, which indicates a wider clustering formation and a parasympathetic activation. However, SD_2 also increases for both recoveries. This physiological fact implies that the recovery stages cluster is more disperse in both directions, which is not desired. Ideally, SD_2 should have the opposite behaviour for the recovery stages. Thus, from this analysis, we conclude that the recovery process is actually having an effect into the parasympathetic part of the ANS, but it is not lowering down the sympathetic contribution.

The implemented physiological recovery process and its physiological effects have been verified throughout these analysis. One of the main limitations of this online recovery process is that it is not producing an averaged close to zero physiological response, and in some cases it is not suppressing the sympathetic contribution. This fact is mainly due to two main factors:

- Achieving a flat (non-active), parasympathetic prevailing, physiological response when being under a virtual reality emotion elicitation experiment is a challenging task. Considering that this experiment, for most of the volunteers, was the first virtual reality experience, the observed differences between the physiological dynamics into the non-fear stimuli and into the recovery stages are slightly noticeable for the GSR analysis.
- The hard-coded confidence interval implemented obviates the actual physiological trend or dynamic within every processing window. Moreover, the Poincaré-plot is insensitive to trends in the IBIs or heart-rate. For instance, it could happen that, given a set of different consecutive temporal processing windows and a hard-coded confidence intervals, the trend of the signal being evaluated is positive, which in case of the GSR signal would mean an arousal increment.

Up to my knowledge, no open public database considered an active biofeedback-based recovery monitoring within their experiments, which makes this part of our database, as well as the analysis presented in this section, a novel contribution. The stated limitations were used to keep researching into new online wearable recovery implementations. In fact, an improved version of the presented recovery process is being developed and implemented. For instance, one of the first stages of the new online recovery monitoring has been implemented in [280]. Specifically, this first stage used online feature extraction for the BVP signal and, by means of least squares linear regression, the different feature trends were analysed to ensure sympathetic activation reduction. Although this new recovery process is still under development, it has been already tested for a small sample size of volunteers and proved to outperform the initial recovery monitoring process.

6.3.2 Physiological uni-modal results

As stated in the previous Section, the reported results in this Section has been obtained with the signals acquired by the BioSignalPlux® research toolkit system. Thus, these results are based on an offline machine learning system implementation. Even though such system has not been embedded, these results represent the first baseline fear detection results of our dataset. Note that the design and implementation of the fear detection system based on our dataset is mainly motivated by the

limitations found in the previous fear detection systems proposed in Chapter 4. In this case, and targeting the improvement and deepening of the subject-independent models, the presented system is focused on such approach. Regarding the labels, both discrete and dimensional have been used following the same fear-binary approach as stated in Chapter 2 and applied in Chapter 4. However, due to the labelling inconsistencies observed for some of the volunteers in Section 6.2, we decided to exclude volunteers number 5, 6, 15, 33, and 40 for the discrete case and volunteers number 3, 5, 6, 13, 20, 21, 22, 40, and 42 for the dimensional use case from the evaluation since they had only around 25% of the positive class. This is considered here as the used labels are the self-reported ratings, unlike the previous Section that used the validated or target labels. Further research might be realised towards analysing and quantifying the effect of severely imbalanced subjects within subject-independent machine learning systems. The latter is not within the scope of this PhD thesis. It should be noted that during the development of the presented fear detection system, two supervised bachelor thesis [291, 292] and one supervised Master thesis [293] provided support into the design space exploration.

The implemented physiological data processing architecture is shown in Figure 6-18. The initial stages are based on the previous proof-of-concept systems presented in Chapter 4. First of all, the applied filtering stages follow the same processes as detailed in Section 6.3. For the data segmentation and overlapping, the 20-second and 50% overlapping strategy is used, as done for the MANHOB system in Section 4.2.2. The feature extraction process includes additional features in comparison with the previously presented fear detection systems. Specifically, 57 features are extracted: 31 from BVP, 20 from GSR, and 6 from SKT. These features are normalised following a Z-score technique and later fed into the feature selection stage. Note that, before the feature selection process, the train-test split is done in a personalised manner, by using a hybrid CV technique, LASO, as detailed in Section 3.1.7.3. This technique takes into account both the intra and inter variability of the volunteers, unlike LOSO and LOTO. Specifically, the LASO partition is done by leaving out half of every volunteer, i.e. the first seven audiovisual stimuli responses are used for training and the other seven are used for testing. The test set is further employed to execute a full blind testing, which will be used to assess the final

system performance. Note that this train-test partition configuration is one initial approximation and it can be further improved and/or performed differently. For the training process, a training and validation partition is done with a 5-kFold CV. This partition is also used during the hyperparameter optimisation done through SMBO, as detailed in Section 4.2.4. This architecture is applied, validated and tested based on the same three classifiers as for the fear detection system presented in Section 4.2.4: SVM, KNN, and ENS.

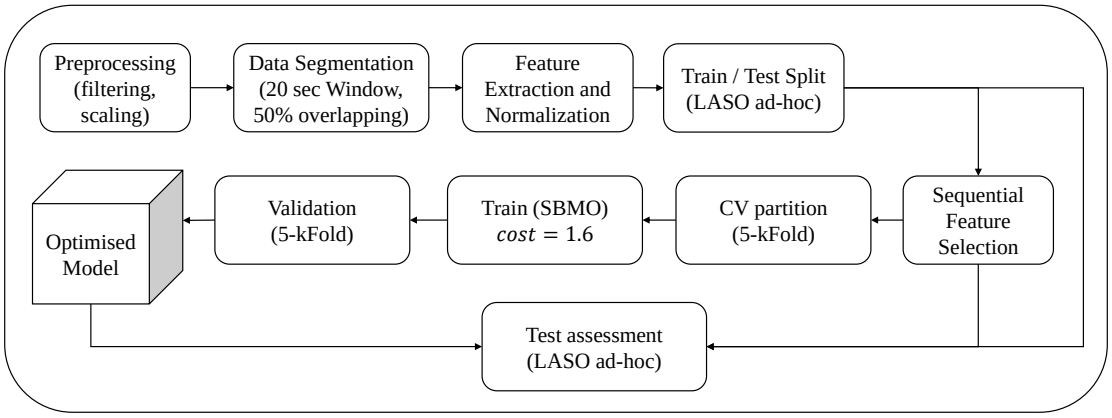


Figure 6-18: Physiological data processing architecture for training and testing the generated machine learning models using our own dataset.

6.3.2.1 Feature Extraction

Within the feature extraction procedure, the different delineation processes are also performed. In this case, the BVP signal is subjected to a stacking or an approximate computing approach using three different algorithms. The first algorithm is the one applied in Section 4.1.3. The second algorithm is given by [294] and it is based on an adaptive threshold method for PPG peak detection. The third algorithm is provided by [295] and it is based on a moving average of valley- peak differences along with local threshold filters to identify the systolic peaks of the PPG signal. These three algorithms can be grouped within the local maximum and minimum slope-based methods, which are the less computational demanding BVP delineation algorithms. Note that they obtained a peak detection accuracy over 90.00% for their respective validation. Unlike using just one delineation algorithm as done for the previous presented fear detection systems, this triple approach is considered in this case to deal as much as possible with any PPG morphological deformation. Note that in our dataset, the audiovisual stimuli are VR-based, which introduces more body movements (motion artefacts) than the rest of the public open databases based

on solely 2D stimuli. The delineation of the GSR signal has been performed as in Section 4.1.3. Thus, we assumed a linear combination of both GSR components, SCL and SCR, followed by equation 2.7.

Regarding specifically the number of new features added, Tables 6.4 and 6.5 detail each of the features being considered in this case. Note that the extracted features for the SKT signal did not change and they are the same as the ones specified in Table 4.14. For the BVP, six new features were included. In the time domain, the mean value of the signal and the root mean square of the IBI (HRV-RMSSD) were considered. The former is used to account for any respiration residual left that was strong enough after having filtered the signal. The latter is an additional metric derived from the extracted IBIs and it provides another indicator for the cardiac vagal control, i.e. the higher the metric the more parasympathetic activation. In the frequency domain, the different frequency bands were rearranged and more information regarding the energy contribution of the different bands, relative and normalised, was obtained. Note that such band definition rearrangement has been assumed due to recent publications [180]. Moreover, a major change was implemented for the PSD estimation considering the unevenly acquired IBIs. Thus, rather than interpolating and using the Welch's overlapping segment averaging estimator, the Lomb-Scargle periodogram has been employed to estimate the PSDs for the specified spectral bands [296]. This technique allows us to relax the frequency resolution considerations. After obtaining the PSD contribution for each band, their energy spectral density is computed, which was later normalised to obtain the energy ratio between the LF and the HF bands. Finally, the non-linear features has been expanded with up to seven derived Poincaré-plot metrics. These were mainly based upon [289, 290] and their computation use SD_2 and SD_1, which are detailed in the previous Section and given by equation 6.5. Thus, the longitudinal and the transverse length of the plot is computed following as:

$$\begin{aligned} L_{SD_2} &= 4 * SD_2, \\ T_{SD_1} &= 4 * SD_1. \end{aligned} \tag{6.6}$$

Note that these two recurrence-based features are directly related to both standard deviations SD_2 and SD_1. Thus, they follow the same physiological rationale but

on an enhanced manner due to the multiplication factor. Additionally, the Cardiac Sympathetic Index (CSI), Modified CSI (MCSI) and Cardiac Vagal Index (CVI) were also calculated. They are computed as follows

$$\begin{aligned} CSI &= L_{SD_2}/T_{SD_1} = SD_2/SD_1, \\ MCSI &= L_{SD_2}^2/T_{SD_1}, \\ CVI &= log_{10}(L_{SD_2} * T_{SD_1}). \end{aligned} \tag{6.7}$$

It can be observed that these features are strongly related to sympathetic and parasympathetic activation given a Poincaré-plot. For instance, the MCSI improves the longitudinal length to emphasise the sympathetic response, which can lead to distinguishing weaker sympathetic activations. One of the major changes within the BVP non-linear features is that the MSE is not considered in this system. This fact is due to observed limitations when extracting this feature for short temporal windows [297, 298]. For the GSR, two more features were included. These are the average relative recovery time and the average area under the detected peaks. The former has been already used in the previous Section. The area under the peaks is computed by means of trapezoidal approximation. Note that the latter can be further improved using Simpson's rule at the expense of increasing the computational time.

6.3.2.2 Feature Selection

To reduce the dimensionality of the problem, SFS is employed. This technique allows selecting the most relevant features and, thus, reducing the training and inference time complexity and storage requirements. Thus, we run the SFS for each of the three classifiers considering every generated training set of volunteers. In the case of the SVM, a RBF kernel is used with with $\gamma = 1$ and $C = 1$. The KNN is set to Euclidean distance with 10-K nearest neighbours. Lastly, the ENS uses an AdaBoost algorithm with boosted decision trees as weak learners and maximum number of splits up to 10. The cost function for each iteration of the SFS was given by $1 - MCC$. Note that this feature selection process is performed after the train/test split to avoid information leakage from the test set. In the next points, we provide the list of features that were selected at least once for all generated models. When using the binarized discrete labels, the following best features are obtained:

Table 6.4: Features extracted for the BVP signal and the proposed fear binary emotion recognition using our dataset.

Sensor	Domain	Features
BVP (31)	Time-domain: (4)	Filtered data mean value
		Mean of IBI
		HRV-SDNN
		HRV-RMSSD
	Frequency-domain: (12)	Normalised IBI PSD contribution (summation) for:
		Low frequency (LF) (0.01–0.15 Hz)
		High frequency (HF) (0.15–0.40 Hz)
		Ultra-High frequency (UHF) (0.40–1.00 Hz)
		Energy contribution of those IBI PSD bands
		Relative energy of those IBI PSD bands
		Normalised energy ratio between LF and HF
		Normalised energy ratio for LF and HF
	Non-linear: (15)	From Poincaré-plot: SD2, SD1, LSD2, TSD1, CSI, MCSI, CVI
		Detrended fluctuation for the filtered signal
		Recurrence rate
		Determinism
		Laminarity
		Longest RP diagonal line
		Diagonal lines entropy
		Trapping time
		Correlation dimension

- For the SVM-based system (15 features total selected):
  - BVP (9): filtered data mean value, HRV-RMSSD, energy contribution of HF and UHF, ratio LF/HF, SD_1, T_{SD1}, detrended fluctuation analysis for the filtered signal, laminarity and diagonal lines entropy.
  - GSR (5): filtered data mean value, and its standard deviation, area under the detected ERSCRs, first and third quartile distribution.
  - SKT (1): filtered mean value.
- For the KNN-based system (11 features total selected):
  - BVP (2): filtered data mean value, and laminarity.
  - GSR (8): filtered data mean value, average number of ERSCR peaks, average relative amplitude, rise time and recovery time of ERSCR peaks, area under the ERSCRs, first and third quartile distribution.
  - SKT (1): filtered mean value.
- For the ENS-based system (13 features total selected):
  - BVP (3): filtered data mean value, diagonal lines entropy and trapping time.
  - GSR (8): filtered data mean value, average number of ERSCR peaks,

Table 6.5: Features extracted for the GSR signal and the proposed fear binary emotion recognition using our dataset.

Sensor	Domain	Features
GSR (20)	Time-domain: (9)	Filtered data mean value ERSCR including number of peaks ERSCR amplitude and rise time ERSCR recovery time and area under the peak Filtered data Standard deviation First quartile Third quartile
	Frequency-domain: (3)	Power spectral density of two bands for SCL and SCR components (0–0.05 Hz, 0.05–1.5 Hz) Spectral density ratio for 0–0.05 Hz
	Non-linear: (8)	Detrended fluctuation for filtered data Recurrence rate Determinism Laminarity Longest RP diagonal line Diagonal lines entropy Trapping time Correlation dimension

average relative amplitude and rise time of ERSCR peaks, area under the ERSCRs, first and third quartile distribution, and laminarity.

- SKT (2): filtered mean value, and power spectral density of the lowest band (0–0.1 Hz).

When using the binarized dimensional labels, the following best features are obtained:

- For the SVM-based system (11 features total selected):
  - BVP (4): filtered data mean value, mean of IBI, recurrence rate and diagonal lines entropy.
  - GSR (5): filtered data mean value and its standard deviation, first and third quartile distribution, and trapping time.
  - SKT (2): filtered mean value and power spectral density of the lowest band (0–0.1 Hz).
- For the KNN-based system (8 features total selected):
  - BVP (2): filtered data mean value, and laminarity.
  - GSR (5): filtered data mean value, average number and average relative amplitude of ERSCR peaks, first and third quartile distribution.
  - SKT (1): filtered mean value.

- For the ENS-based system (8 features total selected):
  - BVP (3): filtered data mean value, diagonal lines entropy and trapping time.
  - GSR (4): filtered data mean value, average number of ERSCR peaks, average relative amplitude of ERSCR peaks, and first quartile distribution.
  - SKT (1): filtered mean value.

Overall, the amount of features selected per each model ranges between 15 and 20. Thus, the complexity of the problem is reduced to a relatively low amount of features for the different classifiers and for both labelling methodologies. Note that this fact drastically affects to different stages of the physiological architecture, such as the training and testing (inference). Moreover, regarding the specific number and nature of the most important listed features for each labelling use case, approximately 50% are temporal and morphological, 20% are frequency-based, and 30% are non-linear. Although each classifier did not select exactly the same features, such selection determines a first approximation to obtain the ones providing the most valuable information. For the binarized discrete labelling case, it should be highlighted that the three classifiers agreed on considering the filtered data mean value for the three sensors, features related to the ERSCR peaks, and some non-linear features directly related to the non-periodic characterisation of the system (laminarity, trapping time, diagol lines entropy). For the binarized dimensional labelling case, the same behaviour with respect to the filtered data mean value consideration is repeated, which is also accompanied with features related to the ERSCR peaks by two out of the three classifiers and the same non-linear features. However, the SFS applied for the binarized dimensional use case considers less features agreed among models. This fact can be a consequence of the stronger self-reported disagreement observed in such labelling.

6.3.2.3 Validation and testing results

Also, having applied the feature selection step, we decided to employ a cost-sensitive learning approach to deal with the imbalance labelling situation. This was done by tuning a miss-classification cost parameter, as performed with previous proposed fear detection systems in Section 4.1.4. For this specific case, a miss-classification cost of 1.6 was applied over the positive class (fear), which was fixed

Class	Discrete Labels	Dimensional Labels
Fear	1496	1335
Non-Fear	2107	1942
Balance (Fear/Non-Fear)	42/58%	40/60%

Table 6.6: Total number of instances for our dataset based on both binarized discrete and dimensional self-reported labels.

by an experimental parameter sweep after the feature selection stage. Note that this design consideration makes the system less prompt to omit a dangerous situation for the use case being addressed [184], i.e. it increases sensitivity.

The physiological machine learning system output is a binary label every 10 s, as denoted in Section 4.2.2. Thus, for this first approximation, we assume that the ground truth of a specific stimulus is the binarized self-reported label assigned to it, regardless of the total amount of instances generated, i.e. all the generated instances within the same stimulus have the same label. For instance, there are audiovisual stimulus within the same class that generates more instances than others. Such approximation can be critical for short-length stimulus, as stimulus number eighth from the first batch, whose duration is 23 s. Such duration implies one generated instance, which can seriously damage the balance of the system or even be insufficient to properly characterised the target emotion of that stimulus. This limitation is tackled by considering the complete number of instances for fear and non-fear classes without relying onto the amount of information provided by each stimulus. Table 6.6 reports the total number of instances based on both binarized discrete and dimensional self-reported labels considering a 20 second processing window and 50% overlapping. Note that these values are obtained for all the volunteers independently of imbalanced labels. It can be observed that the balance is close to that reported in Section 6.2.

The validation and testing results for the different classifiers are detailed in Table 6.7. Note that the average and mean absolute deviation values are shown for the different classifiers, partitions and labelling approaches. These results come from the 42 and 38 models considered for the binarized discrete and dimensional model, respectively. Overall, the obtained results are inline with the ones obtained in Section 4.2.4.2, in which the SVM showed the worst performance, followed by the KNN, and

Table 6.7: Validation and testing results for the different physiological machine learning systems using the first release of WEMAC. Results for both approaches binarized discrete (Disc) and dimensional (Dim) are shown.

Classifier	Partition Type	SEN (MAD)	SPE (MAD)	Gmean (MAD)	ACC (MAD)	AUC (MAD)	F1 (MAD)
SVM	Val-Disc	83.02(1.19)	78.79(0.99)	80.87(0.79)	80.72(0.79)	86.72(0.79)	78.16(0.16)
	Test-Disc	64.36(17.39)	67.88(13.77)	65.29(12.48)	65.70(11.44)	65.33(16.08)	62.84(13.79)
	Val-Dim	86.45(1.05)	75.36(2.91)	80.63(1.54)	82.26(1.07)	87.33(1.00)	82.23(0.99)
	Test-Dim	**72.78(13.21)**	53.86(11.24)	61.60(9.23)	62.51(9.03)	62.14(11.24)	**65.01(9.64)**
KNN	Val-Disc	81.15(4.52)	75.55(4.92)	78.28(4.71)	73.18(5.85)	86.84(4.26)	75.32(5.04)
	Test-Disc	65.53(14.63)	**69.00(13.49)**	**66.08(10.67)**	**66.87(9.64)**	**66.45(14.30)**	64.72(10.17)
	Val-Dim	84.27(4.49)	84.65(3.92)	84.46(4.17)	84.45(4.20)	92.34(3.44)	84.60(4.19)
	Test-Dim	61.08(14.32)	65.01(14.45)	61.37(7.68)	61.78(6.84)	61.43(10.46)	60.00(9.76)
ENS	Val-Disc	81.82(4.31)	75.40(5.29)	78.53(4.82)	68.17(3.70)	75.52(4.01)	64.19(3.74)
	Test-Disc	**68.55(12.10)**	61.61(16.81)	63.51(10.48)	64.50(9.54)	64.57(14.25)	**65.11(8.31)**
	Val-Dim	94.02(0.48)	93.18(0.53)	93.60(0.44)	93.71(0.44)	98.40(0.22)	93.72(0.44)
	Test-Dim	65.91(15.71)	**64.98(14.69)**	**63.75(9.28)**	**64.23(8.15)**	**66.62(12.20)**	63.37(10.63)

being the ENS the best one. In fact, when analysing the provided averaged metrics and their dispersion values together for both discrete and dimensional labelling, the AdaBoost (ENS) classifier is the one outperforming the other two. One of the key differences among them is that, while the SVM and the KNN are losing specificity for the dimensional use case, the ENS keeps the balance between sensitivity and specificity leading up to a very similar Gmean for both cases.

On the one hand, specifically for the discrete labelling, the best averaged results are obtained by the KNN classifier for both validation and testing partitions. Moreover, the highest specificity, Gmean, accuracy, and AUC are achieved in this case for this classifier with up to 69.00%, 66.08%, 66.87%, and 66.45%, respectively. On the other hand, when dealing with the dimensional labelling, the classifiers do not follow exactly the same behaviour as for the discrete. In fact, the best classifier in such case is the ENS reaching the highest specificity, Gmean, accuracy, and AUC with up to 64.98%, 63.75%, 64.23%, and 66.62%, respectively. Note that these results are obtained using a reduced set of features, as detailed in the previous Section. This fact, accompanied by the challenges when dealing with a subject-independent approach, confers high value and great potential to these first initial baseline WEMAC results.

To contextualise the behaviour of the classifiers for the different considered models, Figures 6-19 and 6-20 show the MCC performance metric over their test partitions. Note that this metric uses all the information from the confusion matrix and provides a correlation-like value considering the ground truth and predicted confusion

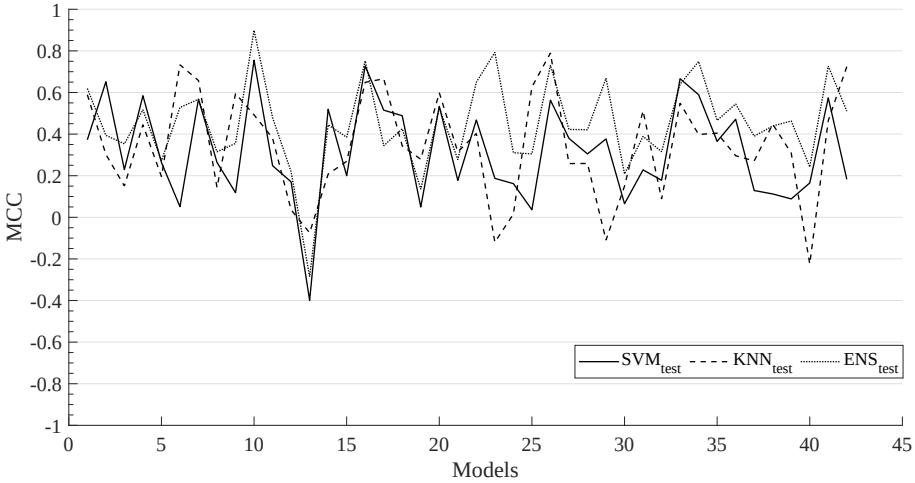


Figure 6-19: MCC test-metric evaluation for all the 42 models considered within the binarized discrete fear detection use case.

matrices. For both labelling use cases, we observe that the different models follow a similar behaviour for some of the considered volunteer. For instance, the correlation of the MCC metrics is up to 0.48(0.17) and 0.44(0.13) for the discrete and dimensional labelling, respectively. This fact is remarkable since it is an indicator that demonstrates the independence of the specific behaviour of each classifier with respect to the dataset used. In line with the reported results of Table 6.7, the discrete labelling approach outperforms the dimensional one. However, these plots also show a very subject-dependent results in some cases. This is reflected into a very high variability, which is also reported by the MAD of the different performance metrics. Specifically for these plots, the MCC metrics are approximately contained within -0.4 and 0.9 and within -0.4 and 0.6 for the discrete and the dimensional cases, respectively. Such variability and distributions are plotted in Figure 6-21, which shows the aggregated distribution for the different classifiers and labelling use cases. Note that the median value is the red line or horizontal line within the boxes. As already previously stated, the training using the binarized discrete labelling achieves better models (higher medians) than the binarized dimensional labelling. Moreover, the interquartile range dispersion is always smaller when applying the ENS classifier.

Up to my knowledge, the generated systems are the first fear detection systems using a reduced set of physiological signals and virtual reality stimuli. For instance, these results constitutes the physiological baseline for the WEMAC dataset. It has been demonstrated that, overall, the systems trained with the binarized discrete labelling obtain better models than the ones trained with the binarized dimensional

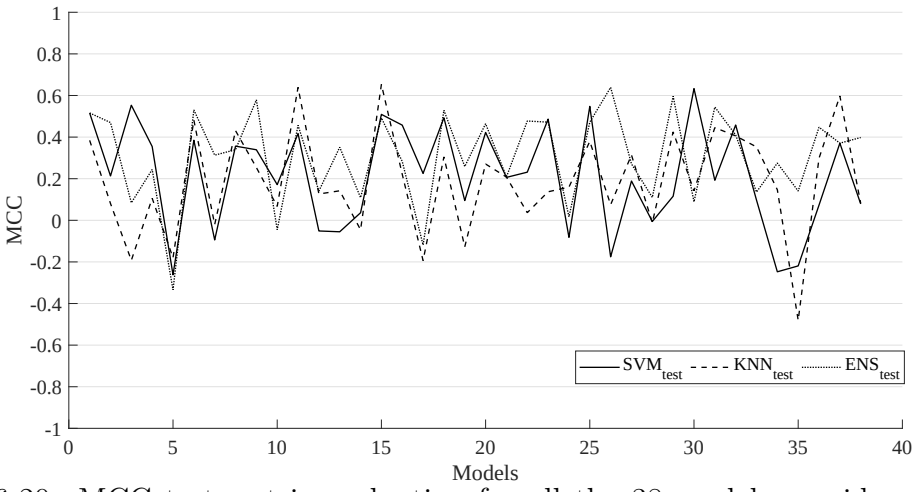


Figure 6-20: MCC test-metric evaluation for all the 38 models considered within the binarized dimensional fear detection use case.

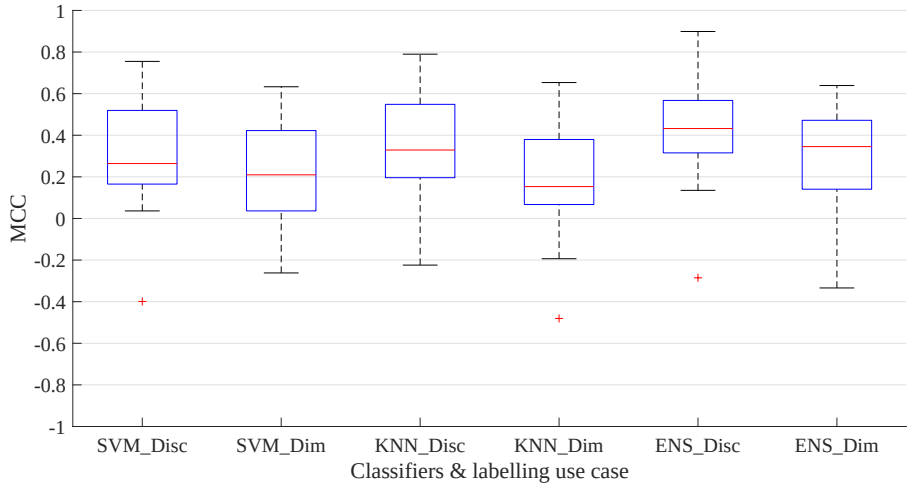


Figure 6-21: MCC test-metric box plot distribution for all the 42 and 38 models considered within the binarized discrete and dimensional fear detection use cases.

labelling. This fact is in line with the self-reported labelling response exploration conclusions stated in Section 6.2.

The main limitation of the generated machine learning models is the observed high variability and dispersion. This fact could be due to an over-fitting problem, which could be causing the generation of low bias and high variance models. However, the obtained validation metrics do not suggest that. In fact, the classifier providing the best validation performance metrics, that is AdaBoost with over 90.00%, is the one being the less susceptible to this type of problem, see Section 3.1.7. Thus, different aspects need to be studied and analysed to provide a proper explanation to this problem:

- The applied CV LASO technique leaves a small amount of samples into the testing partition. Although this technique is intended to deal with subject-

personalisation, the small testing dataset size can result into an unrepresentative dataset. Thus, other partition techniques might be exploited towards considering a higher testing dataset and/or including more inter and intra-variability. This fact is also affected by the limited amount of available data for each volunteer when dealing with laboratory-based databases.

- In view of such variability among the different volunteers, some processes within the proposed training architecture can be personalised based on the specific dataset, i.e. based on the specific combination of volunteers or the classes distribution. For instance, the cost-sensitive approach can be tuned for every different training set.
- The partition techniques for the generation of the train-validation datasets could be changed to the same CV technique as applied for the first partition (LASO).
- Further physiological analysis can be done to find clusters and extreme differences among the different volunteers. As a first approximation, such clusters can be based on simple physiological filters such as: level of GSR activation (hypo-activity vs hyperactivity), SKT ranges, and level of residual noise of the PPG signal after filtering.
- Different types of normalisation and scaling techniques can be also employed to evaluate their effect. This fact is directly related to the need of discovering the best way to model the problem, i.e. subject-independent fear detection.

Amongst these stated considerations, from my point of view, the one that is currently affecting the most is the unrepresentative test dataset risk. Specifically in this case, the applied LASO left approximately up to 1.3% of the total data for the test set. This is an average of 48 samples over a total of 3600 available instances. For the sake of providing a starting point regarding this specific discussion, Table 6.8 reports the results for the same physiological architecture when training the KNN system with the binarized discrete labelling, but using a LOSO partition for the train-test split. We can observe that the LOSO case achieves smaller average results, which can be affected by the intra-variability of the unseen volunteer that is not considering for the training. However, the most important fact is spotted at the dispersion difference. The LOSO system reports less variability, which can be

an indication of a more representative test set. Note that more studies and analysis need to be done to properly characterised this fact.

Table 6.8: Validation and testing results for the KNN machine learning systems using the binarized discrete labelling and a LOSO CV technique for the train-test partition.

Partition Type	SEN (MAD)	SPE (MAD)	Gmean (MAD)	ACC (MAD)	AUC (MAD)	F1 (MAD)
Test-Disc-LASO	65.53(14.63)	69.00(13.49)	66.08(10.67)	66.87(9.64)	66.45(14.30)	64.72(10.17)
Test-Disc-LOSO	64.05(10.69)	60.93(9.26)	61.74(6.98)	61.90(6.93)	62.25(9.54)	58.48(7.18)

Within this variability casuistry and besides the recommended analysis to be performed towards the improvement of the models, it should be also considered that the emotional latency and physiological dynamics of every volunteer for each stimulus is affecting the binary class separation. For instance, Figure 6-22 shows the LF/HF Ratio extracted from the IBI signal of volunteer number three. Specifically, each bar represents the extracted feature for a 20-second window with 18-seconds overlapping. Note that this overlapping is applied in this case to reduce the temporal resolution and enhance the feature dynamic visualisation along each stimulus. We can observe how the evolution of this specific feature varies within every stimulus. From this information, we can perform an analysis as the one provided in Section 6.3.1.1, in which the GSR patterns where studied. However, the problem to be highlighted in this case is that, regardless of the dynamical evolution of the features within the stimuli, the same label is being assigned to all generated instances. Thus, different machine learning techniques should be exploited towards considering temporal feature evolution or contextualisation together with a different labelling approach. The latter is referred to the possibility of applying semi-supervised machine learning to treat the current hard-labels as soft ones. This can even be thought as models in which the labels as learnable parameters. Moreover, the combination of both current labelling methodologies, discrete and dimensional, should be exploited towards taking advantage of each.

6.4 Multi-Modal data fusion framework

After having presented and explained the design of the physiological uni-modal machine learning system, the other main focus along this Chapter is the multi-

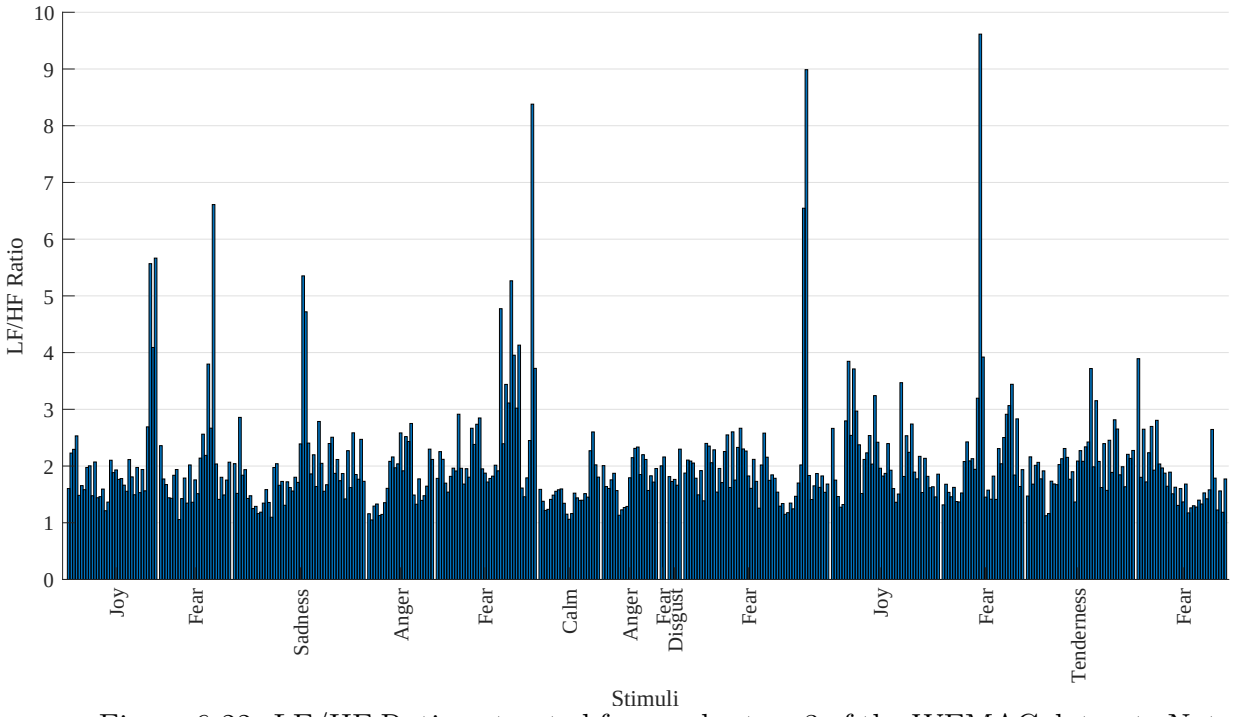


Figure 6-22: LF/HF Ratio extracted from volunteer 3 of the WEMAC dataset. Note that in the abscissa are represented the targeted emotions for the first Batch.

modal data fusion capabilities that the architecture of the Bindi system can offer[4]. In fact, emotion recognition systems based on multi-modal information are overtaking over uni-modals in the affective computing community [155, 174]. Most of the presented multi-modal systems in the literature used to be based on audio and visual data [299], speech and face gestures [300], EEG and facial expressions [301]. There are datasets gathering multi-modal information, which consider physiological and speech information [302, 303]. Little attention is paid to the multi-modal design exploration using physiological and speech information. Solely one work has been found in the literature using these two types of information for short term observations [304]. They employed a hybrid fusion by means of feature and decision level data fusion, which yield up to 55.00% accuracy for a subject-independent approach and a binary arousal-valence classification. Thus, the proposed multi-modal framework and methodologies serve as an initial approach towards working with real elicited fear in women and its proper processing considering both physiological and speech information.

Before going into details for the different designed systems and obtained results

[4]The research presented in this Section is based on a multidisciplinary work with the members of UC3M4Safety experts in Signal Theory and Communications

using the information collected during the development of the WEMAC dataset, a proper contextualisation regarding the multi-modal casuistry and capabilities within Bindi might be properly explained and detailed. On this basis, different arrangements of the system components were proposed to explore the possibilities of such multi-modal design space. This provided the outline of a design space exploration for different system architectures (physiological and speech/audio information data fusion coming from the bracelet and the pendant, respectively). Figure 6-23 shows such outline and depicts the potential relationship to be encountered for the different use cases presented. The proposed use cases are detailed as follows:

- Case 1: Uni-modal. This arrangement is the fear detection capacity baseline for each of the uni-modal systems, the physiological and the speech models.
- Case 2: Multi-modal with pre-alarm. In this case, the physiological information is continuously evaluated based on the uni-modal physiological system. When it detects that the user is experiencing fear, it triggers a pre-alarm to the Bindi APP. Note that this is done following a computing-on-the-edge approach, as it is the bracelet itself running a lightweight machine learning engine. The fear detection causes the Pendant to start recording audio for a brief period, resulting in a low-energy consumption strategy for the microphone. The audio signal is then sent to the for layer of the system, i.e. the Bindi APP, to perform fear detection using also a separate speech-based uni-modal intelligence engine.
- Case 3: Multi-modal with periodic audio sampling. This case only differs from the previous one in that there is not pre-alarm, but the speech/audio is sampled on a periodic basis.
- Case 4: Multi-modal with pre-alarm and periodic audio sampling. This set-up is based on the conjunction of the previous two use cases. Thus, it represents a middle-stage between having continuous multi-modal information and the previous cases.
- Case 5: Continuous Multi-modal. This is the last proposed arrangement of the system and it requires the highest amount of resources as both uni-modal systems, physiological and speech/audio, are always active, being the data fusion performed in a continuous manner.

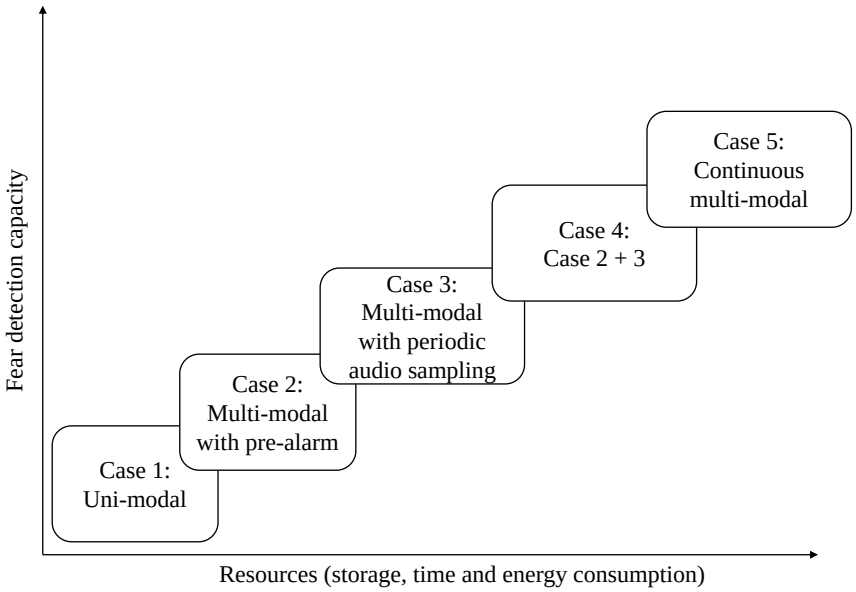


Figure 6-23: Design space exploration outline for the different modality arrangements to be performed with the architecture of Bindi.

It should be noted that these detailed use cases define the system architecture in terms of information availability, rather than specifying the applied data fusion methodology. The latter is referred as to the techniques used to perform the physiological and audio information data fusion. They can be actually done in different ways within the same use case. Note that the different typical data fusion techniques are described and explained in Chapter 3. In fact, this research work, besides exploring the physiological uni-modal fear detection system, focuses on analysing and comparing three proposed data fusion techniques by using different system architecture arrangements mainly related with Case 2 and 5. The rest of the use cases are not within the scope of this research. Thus, the analysis of other arrangements as well as the rest of the use cases are left for the subsequent datasets to be released within the UC3M4Safety database.

Within this multi-modal framework, I have performed the integration of the physiological system detailed in the previous Section. Specifically, the KNN classifier was employed. The speech uni-modal system has been designed and implemented by the components of the UC3M4Safety team with expertise in audio signal processing. This system includes the following fundamental modules: Voice Activity Detection (VAD), Spectral Subtraction (SS), feature extraction, and a neural network-based classifier [9]. Note that the binarized discrete labelling use case and the LASO CV were applied for both uni-modal systems, and the speech system also excluded the

high imbalanced volunteers specified in the previous Section. Finally, the physiological and speech uni-modal subsystems provide a binary label every 10 and 1 s, respectively.

Regarding the multi-modal design space exploration, Case 2 was the first implemented; it is explained in Chapter 5 and implemented in the first version of Bindi or Bindi 1.0 [221], which is based on a hierarchical data fusion strategy. In this version, physiological information is continuously collected by the Bracelet, which runs a lightweight uni-modal physiological fear detection intelligence engine. When it detects that the user is experiencing such emotion, it triggers a pre-alarm to the Bindi APP. This action causes the Pendant to start recording audio for a brief period, resulting in a low-energy consumption strategy for the microphone. The audio signal is then sent to the Bindi APP to perform fear detection using a speech-based uni-modal intelligence engine. Finally, if the latter system confirms the detection, the Bindi APP starts a safety procedure to help the user, triggering an alarm to the respective responders. The second system arrangement analysed, Bindi 2.0a , is also related with Case 2 and it is based on the same two uni-modal data processing pipelines in Bindi 1.0 but applying, at the final decision stage, a late fusion technique rather than a hierarchical agreement or confirmatory strategy [222], Figure 6-24. It inherits the pre-alarm functionality and casuistry from Bindi 1.0 to have a low-energy consumption for the microphone. Finally, the last system arrangement, Bindi 2.0b, is related with Case 5. Such system is a variation of Bindi 2.0a but based on a continuous physical and physiological data acquisition. There is not pre-alarm involved and this arrangment follows the late fusion scheme introduced in Bindi 2.0a.

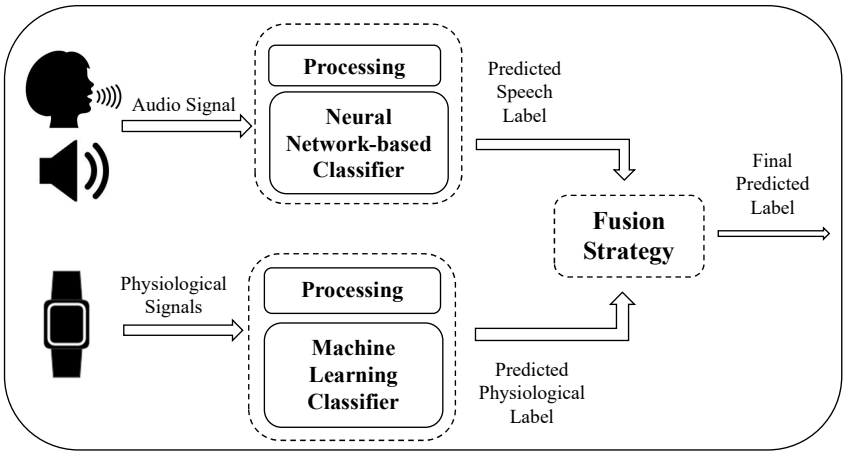


Figure 6-24: Data fusion block diagram for Bindi 2.0a and Bindi 2.0b.

The following subsections detail the different data fusion strategies considered and evaluated with the three system arrangements. The experimental results are an account of the validation process performed offline to evaluate the functionality of the data processing pipelines. This is done to later embedding such modules in the architecture, balancing the trade-offs observed.

6.4.1 Multi-modal data fusion methods

Before going into details for the proposed multi-modal framework, some points should be considered for the assessment of the different multi-modal system arrangements. First, according to the WEMAC dataset design, it should be noted that physiological data are gathered during the stimulus elicitation, whereas speech recording is registered during the subsequent audio annotation. That means that physiological and speech data are not aligned in time. However, both data are required to be aligned for Bindi 2.0b, unlike for Bindi 1.0 and Bindi 2.0a. Since during the labelling the volunteers are requested to revive the emotions felt during the stimulus elicitation, it is assumed that the correspondence is solid enough between both instants. However, this assumption will need further validation when the rest of the subsets in the UC3M4Safety Database became available.

As already detailed in the previous Section, the physiological and speech uni-modal subsystems estimate a binary label, $y_k^m \in \{0, 1\}$, for every time window k, where $m \in \{\text{phy}, \text{sp}\}$ are the two modalities, with phy and sp referring to the physiological and speech subsystems, respectively. However, each of the modalities uses a different time window length, T_m, in seconds. Moreover, the system is intended to output a response per time period n (each of the time periods is of length L), in seconds. Thus, an estimation of fear probability p_n^m for the n-th time period and the m-th modality is computed, given by

$$p_n^m = \frac{\sum_{k=1}^{K_m} y_{K_m \cdot n+k}^m}{K_m}, \tag{6.8}$$

where $K_m = \lfloor \frac{L}{T_m} \rfloor$, i.e., the number of time windows that we consider for each modality for the estimation of probabilities.

Thereafter, a single binary label, Y_n^m, corresponding to probability p_n^m can be

calculated as

$$Y_n^m = \begin{cases} 0 & \text{for } p_n^m < \text{th}_m \\ 1 & \text{otherwise} \end{cases}, \tag{6.9}$$

i.e., it will result in "1" (*fear*) if p_n^m is higher than the modality-related predefined threshold, $\text{th}_m \in \{0, 1\}$, or "0" (*no-fear*) otherwise. Note that the th_{phy} and th_{sp} values are discussed in Section 6.4.2.

As a metric to represent how confident each uni-modal system is for the class label predicted in a given period, entropy h_n^m for the n-th time period and m-th modality is calculated as

$$h_n^m = -[p_n^m \cdot \log(p_n^m) + (1 - p_n^m) \cdot \log(1 - p_n^m)]. \tag{6.10}$$

On this basis, three late fusion strategies are studied to produce fused system response Y_n^{f} for the n-th time period:

- Case 1, Lowest Entropy: The system's response corresponds to the binary label produced by the uni-modal system with the smallest entropy, i.e., the most confident one. To this end, fused fear probability p_n^f for the n-th time period is calculated as

$$p_n^{\text{f}} = \begin{cases} p_n^{\text{phy}} & \text{if } \; h_n^{\text{phy}} < h_n^{\text{sp}} \\ p_n^{\text{sp}} & \text{otherwise} \end{cases}. \tag{6.11}$$

  Next, applying the same rationale as in Equation (6.9), a fused binary label is obtained as

$$Y_n^{\text{f}} = \begin{cases} 0 & \text{for } p_n^{\text{f}} < \text{th}_{\text{f}} \\ 1 & \text{otherwise} \end{cases}, \tag{6.12}$$

  where, for now, th_{f} is the conventional 0.5.

- Case 2, Inverse Entropy Weighted Combination: Fused fear probability p_n^f for the n-th time period is computed as a weighted sum of probabilities, as given by

$$p_n^{\text{f}} = \sum_m w_n^m \cdot p_n^m, \tag{6.13}$$

where

$$w_n^m = \frac{1/h_n^m}{\sum_m 1/h_n^m}. \tag{6.14}$$

Next, a fused binary label is obtained according to Equation (6.12).

- Case 3, Logical OR: The system response corresponds to the logical OR computation over the binary labels for each uni-modal system. That is,

$$Y_n^{\mathrm{f}} = Y_n^{\mathrm{phy}} \vee \ Y_n^{\mathrm{sp}}. \tag{6.15}$$

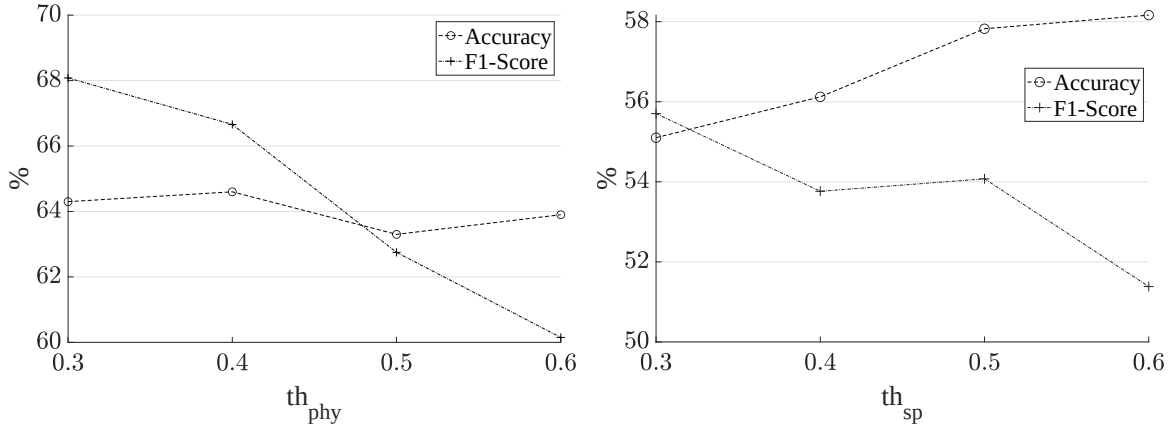
The three fusion strategies are based on the literature (e.g., [305]) and are proposed as a trade-off between low computational complexity and robustness considering the confidence of the system in the predictions. When comparing the three fusion strategies theoretically, the logical OR facilitates obtaining a fear class prediction without checking the subsystem confidence, which could lead to false detection. However, the lowest entropy strategy trusts the most confident model without considering the differences in the probabilities. Finally, the inverse entropy weighted combination establishes a trade-off between the probabilities and entropies for each uni-modal subsystem. Thus, the confidence of this last strategy might be higher than that of the others.

To sum up, regarding the testing procedure, the uni-modal subsystem's outputs are arrays of binary labels. Specifically, for the WEMAC the length of the arrays is equal to dividing the duration of each emotion-related stimulus by the respective uni-modal response window, i.e., 10 and 1 s for the physiological and speech subsystems, respectively. Afterwards, those collected arrays are processed by calculating the soft probabilities and its corresponding hard labels by applying the physiological (th_{phy}) and speech (th_{sp}) thresholds. The data fusion strategies proposed will also generate their corresponding hard labels as discussed before. The evaluation metrics selected, which are accuracy and F1-score, feed on the final hard labels obtained. Accuracy can fairly represent the prediction rates since class imbalance is low. F1-score is considered to deal with the slight unbalance observed. Although F1-score should be a good metric for a detection problem such as the one addressed, in which the number of positives should be relatively low in comparison with the negatives, the experimental setting considering here is almost balanced, and, therefore, this metric

is not as significant as expected to be when testing with data captured in real-life conditions.

6.4.2 Multi-modal data fusion results

The first analysis that needs to be done is the performance of the physiological and speech subsystems working independently in a continuous setting, that is taking into account all of the samples. This experiment is essential to determine the thresholds, th_{phy} and th_{sp}, that convert the binary labels obtained for each period into a single hard label. This step is relevant because it determines if the architecture is more or less prone to false alarms, independently of the Bindi version or multi-modal system arrangement considered. Thus, each parameter was swept in the range $[0.3, 0.6]$ with steps of 0.1 while generating the corresponding 42 uni-modal subsystems following the LASO approach and considering each video length as the different applicable periods. In this regard, Figs. 6-25a and 6-25b show th_{phy} and th_{sp} values versus accuracy and F1-score average metrics for the 42 testing groups in the physiological and speech subsystems, respectively.



(a) (b)

Figure 6-25: Parameter sweep for a) th_{phy} and b) th_{sp} in the physiological and speech uni-modal subsystems, respectively.

Analyzing Fig. 6-25a, it can be observed how F1-score decreases as th_{phy} grows, whereas accuracy remains rather stable. Note that F1-score depends to a great extent on the number of True Positives (TP) predicted but mostly disregards True Negatives (TN). Thus, if TP increases and the sum of False Positive (FP) and False Negative (FN) rates decrease, then F1-score increases. This trade-off causes the behavior observed, where the lower th_{phy} is set, the higher F1-score become. According to this analysis, th_{phy} was fixed to 0.40, getting 66.66% and 64.60% for

F1-score and accuracy, respectively. Note that these values are higher than the ones reported in Table 6.7 due to the effect of considering a set of uni-modal outputs for a given period of time. The reason for choosing this specific threshold value is the good compromise observed between both metrics and the fact that missing a TP could be dramatic for the Gender-based Violence use case. Additionally, the combined multi-modal system should refrain from triggering false alarms to avoid overwhelming the institutions in charge of protecting them, and this is why the speech subsystem is chosen to be more conservative in this regard. By analysing Fig. 6-25b for the speech subsystem, it can be observed how F1 and accuracy begin to diverge from 0.50 onward. Therefore, th_{sp} was fixed to this value, getting 54.07% and 57.82% for F1-score and accuracy, respectively. Note that accuracy could even be increased by choosing a higher th_{sp}.

Once th_{phy} and th_{sp} were fixed, we studied the average performance predicting over the 42 testing groups for the different architecture configurations, as shown in Figures 6-27 and 6-26. These configurations are the physiological uni-modal subsystem, the speech uni-modal subsystem, Bindi 1.0, Bindi 2.0a with lowest entropy data fusion, Bindi 2.0a with inverse entropy weighting data fusion, Bindi 2.0b with lowest entropy data fusion, Bindi 2.0b with inverse entropy weighting data fusion, and Bindi 2.0b with logical OR data fusion. Note that Bindi 2.0a was not combined with logical OR data fusion because it is equivalent to Bindi 1.0.

Analysing Figure 6-27, the physiological uni-modal subsystem achieves the highest accuracy providing up to 64.63% and surpassing even the fusion schemes. For the F1 metric, this subsystem also provides the second highest-rate with up to 66.67%. This behaviour can be related to the bias introduced toward detecting the positive class, first, with the cost-sensitive learning approach and second, with the parameter sweep of th_{phy}. In Figure 6-26, The speech uni-modal subsystem provides significant lower metrics than the physiological subsystem. This fact could be related to the limited number of samples to train the neural network and the possible limited quality of the samples due to the action of reliving the emotion felt in the dataset generation. This situation causes that Bindi 1.0 provides the worst metrics in this analysis due to the final system response falls on the speech subsystem. Bindi 2.0a and Bindi 2.0b both provide a similar accuracy close to the physiological subsystem in most

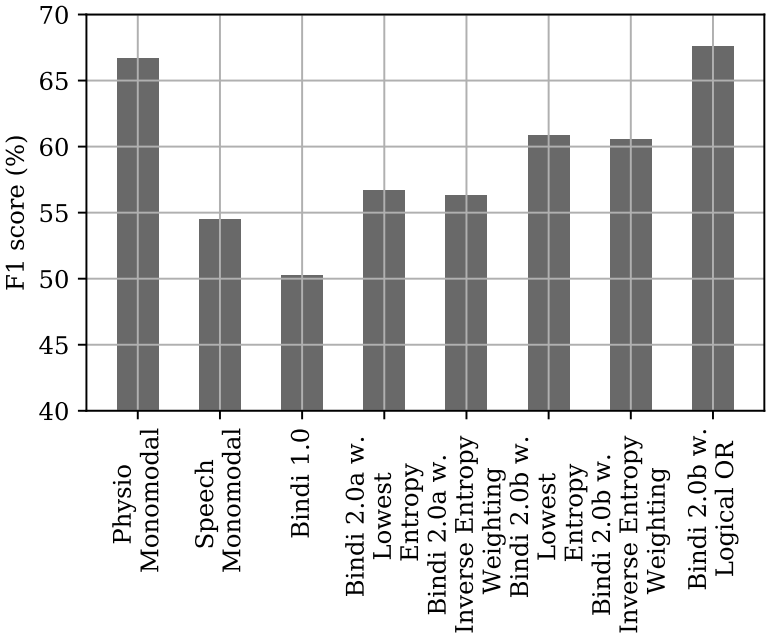


Figure 6-26: Average F1-score performance analysis predicting over the 42 testing volunteers for the different architecture configurations.

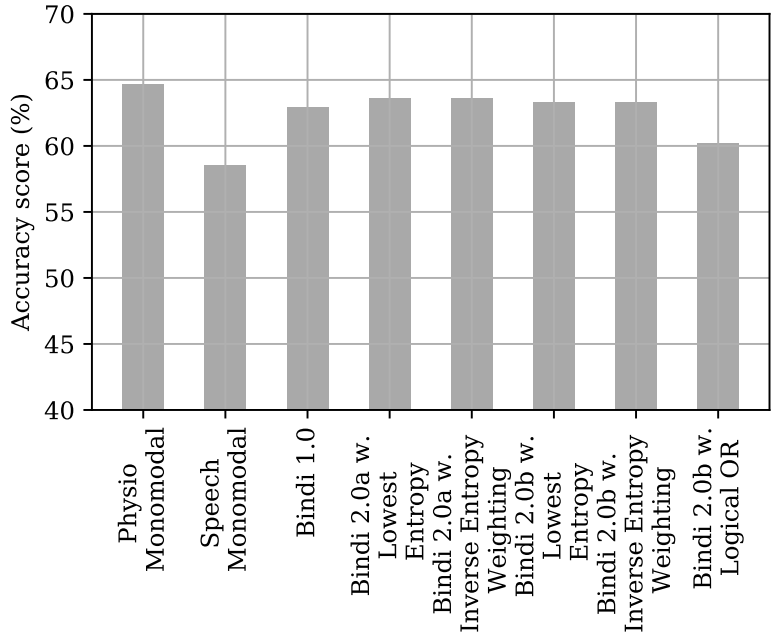


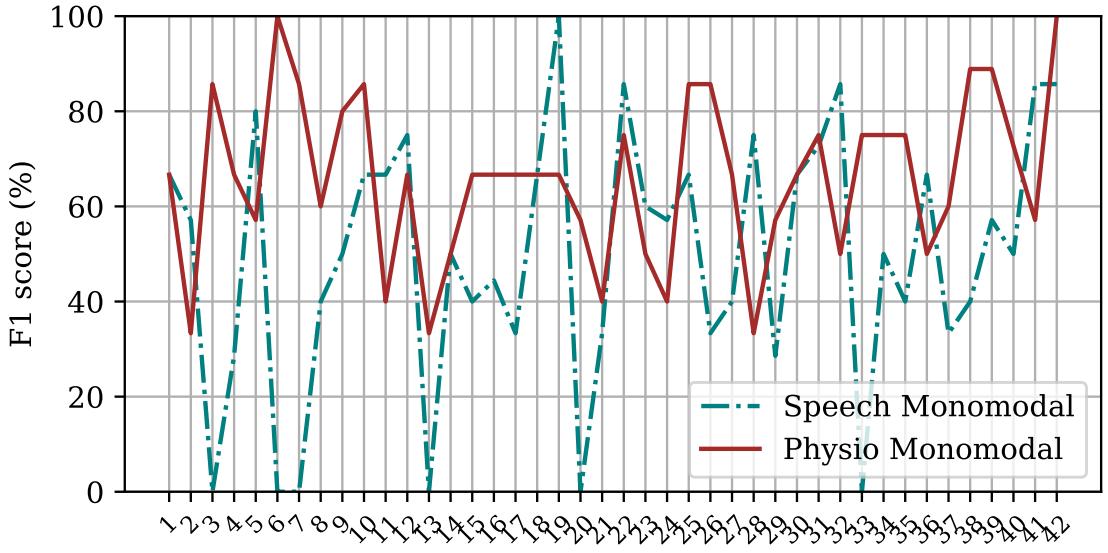
Figure 6-27: Average Accuracy score performance analysis predicting over the 42 testing volunteers for the different architecture configurations.

cases. However, Bindi 2.0b achieves the highest F1-score in all cases, especially for Bindi 2.0b with the logical OR data fusion. This latter strategy provides the highest F1-score, 67.59%, although accuracy is limited. This performance in F1-score could be related to the positive bias contributed by the physiological subsystem due to the lower threshold chosen, th_{phy}, which introduces a conservative bias towards not missing TP at the cost of increasing FP. However, as for the other architectures with fusion strategies, the speech subsystem may be slightly deteriorating the system performance in terms of F1-score and accuracy but is preventing Bindi 2.0a and Bindi 2.0b to produce too many FP. A short preview of this analysis and discussion of the confusion matrices obtained for each configuration can be found in [9].

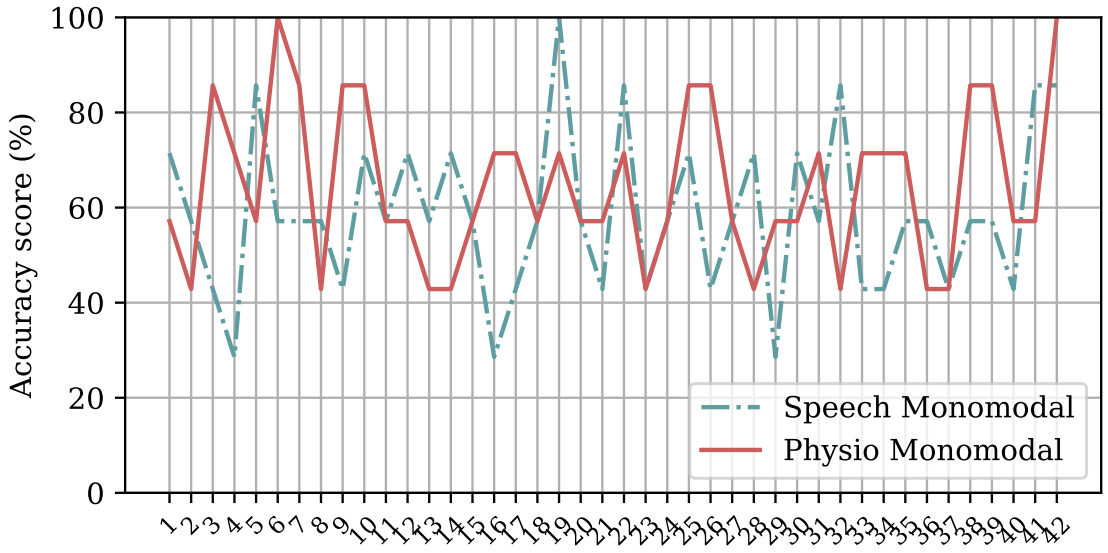
To elaborate the results shown in Figures 6-27 and 6-26, Table 6.9 presents detailed results for the different configurations, including the average standard deviation per volunteer tested. Low standard deviation rates are good indicators of a better generalization ability as long as the results are comparable. Note for example that, although Bindi 1.0 presents the lowest standard deviation, which can be seen as a good generalization, its scores are surpassed by most of the configurations, as previously stated. Moreover, it can be observed that the standard deviation values obtained were relatively high, especially for the F1-score. The cause is shown in Fig. 6-28, where F1-score and accuracy are provided for each of the 42 tests and uni-modal subsystem. It is observed that some volunteers have an F1-score of zero for the speech subsystem in this figure. This fact occurs because F1-score depends on the TP detected, and there were no positive predictions for some users.

		Physiological uni-modal	**Speech uni-modal**	**BINDI 1.0**	**Bindi 2.0a Lowest Entropy**	**Bindi 2.0a Inverse Entropy Weighting**	**Bindi 2.0b Lowest Entropy**	**Bindi 2.0b Inverse Entropy Weighting**	**Bindi 2.0b Logical OR**
F1-score	Mean	66.67	54.48	50.23	56.68	56.33	60.87	60.58	67.59
	Std	17.31	26.73	27.64	23.91	24.05	26.63	26.98	14.27
Accuracy	Mean	64.63	58.5	62.93	63.61	63.61	63.27	63.27	60.2
	Std	16.56	16.73	14.30	14.35	14.35	17.94	18.21	15.75

Table 6.9: Average performance analysis predicting over the 42 testing volunteers. Mean and standard deviations (Std).



(a)



(b)

Figure 6-28: Individual performance analysis for the two uni-modal subsystems.

6.5 Conclusion and discussion

This Chapter presented different essential contributions of this research. First of all, the WEMAC dataset is detailed and explained. Secondly, a physiological-based fear detection system is built upon the data collected in such dataset. Finally, a multi-modal framework contextualised on the Bindi technology is provided. Specifically, these contributions can be detailed as following:

- The generation of a novel emotion detection dataset that deals with all the limitations encountered within public available databases. This dataset belongs to the UC3M4Safety Database.
- The design and integration of an active physiological recovery process within the dataset experiments.
- The design, implementation, and evaluation of a physiological-based fear detection system using the WEMAC dataset.
- A novel multi-modal data fusion framework design using physiological and speech information.
- The application of a LASO model considering fear recognition, multi-sensorial signal fusion, and virtual reality stimuli for the first time.

For the physiological-based fear detection system, the best result is obtained using a KNN classifier and an AdaBoost (ENS) classifier for the fear-binarized discrete and dimensional labelling, respectively. The former reaches up to 66.87% and 66.45%, while the latter achieves up to 64.23% and 66.62% for the averaged ACC and AUC. The obtained results are in line with all the Leave-One-Out (subject or trial) systems presented in the literature, see Table 4.19. However, provided recommendations at the end of Section 6.3.2.3 might be exploited and investigated towards improving these baseline results.

Regarding specifically the proposed multi-modal framework, the best balanced fusion result is obtained for the Bindi 2.0b arrangement by applying a logical OR data fusion strategy. This method reports up to 60.20% and 67.59% for ACC and F1-score, respectively. These values represent a competitive result in comparison with the state-of-the-art that deal with similar multi-modal use cases [301,304,306]. Moreover, it is worth highlighting that the configurations described in this Chapter for fear detection through physiological, and speech data are just a possible way to

characterise the situations and contexts in which Bindi users can be involved. These are meant as initial baselines for further developments that have allowed the identification of important challenges. First, finding a suitable trade-off between TP-TN and FP-FN is crucial since the cost of missing a true need for help is appalling, but we also need to avoid interfering with the everyday life of Gender-based Violence Victims and the saturation of the protection services with false alarms. Thus, we have tried to reduce FNs as much as possible while FPs are maintained at an adequate rate. To this end, we considered strategies based on miss-classification costs and threshold parameters fixing. Specifically, we have fixed the th_{phy} parameter in the physiological subsystem to get a higher outcome of positive predictions with this system so that in a later stage, the speech (in Bindi 1.0) and data fusion strategies (in Bindi 2.0a and Bindi 2.0b) helped in correcting the bias while trying to maintain TP prediction. During this experimentation, the current speech uni-modal system provided lower performance rates than expected, which could be caused by the temporal misalignment of the physiological and speech data in WEMAC. The vanishing of the emotion elicited by the time the voice sample gets collected could be behind this decrease in performance.

In general, we conclude that the obtained uni-modal and multi-modal fear classification systems employing the WEMAC dataset report competitive results in comparison with the state-of-the-art. However, more research is needed to improve these systems towards their applicability into real life. Thus, the main goal of the proposed multi-modal framework and generated WEMAC is to ignite the community interest in this very challenging problem regarding Gender-based Violence and to start tackling the gender perspective into artificial intelligence.

As future work, the UC3M4Safety team plan to foster and develop a series of key items and future lines of action that have been identified as limitations along with the realisation of this work:

- To study other fusion alternatives and combination modes for the uni-modal subsystems.
- To increase the number of volunteers and available sensor data acquired with the Bindi edge devices.
- To include into the database the Gender-based Violence Victims data to better

understand their activation mechanisms under fear-related situations.

- To Embed the complete physiological uni-modal system architecture and data processing into the Bindi bracelet and to test its efficiency in real-life environments and situations in into-the-wild experiments.
- To evaluate the use of alternative score metrics, such as mutual information and area under the curve, to continue finding a proper balance between false alarms and miss probability.
- To develop and test subject-adaptation techniques to both the uni-modal and fusion models.

In the design of fear detection systems for preventing and combating Gender-based Violence situations, several problems may arise when the goal of a system is to work with real-life data. First, the difficulty of finding realistic data, and second, the low confidence on the architectures developed if the data used is acted or synthetic. This situation leads to the need to generate databases with real elicited emotions, which is, indeed, highly challenging and time-consuming. Above all, working with strong negative emotion elicitation, such as the evoked in WEMAC for fear detection in women in a laboratory environment, can lead to ethical issues. Thus, many resources must be devoted to safeguarding the welfare of the volunteers participating. This particular problem is magnified when the target group of volunteers are women who have suffered Gender-based Violence. This is because the failures of the protection system or service have critical consequences for them. For this reason, the second release of the WEMAC dataset currently being collected within UC3M4Safety Database comprises only Gender-based Violence Victims volunteers.

Part IV

Conclusion

Chapter 7

Conclusion

In this final Chapter, we will summarise the contributions of this PhD research based on the proposed goals. We will also provide some suggestions on possible topics to study in the future. These ideas come from the last year of investigation and may suppose the starting point of new research projects.

This PhD started with the creation of UC3M4Safety, a multidisciplinary team created when facing the Gender-based Violence problem and claiming that a multidisciplinary approach were needed to foster new and more innovative solutions to prevent and combat it. Driven by this motivation, we aimed to provide new tools to prevent and combat Gender-based Violence risky situations and, even, aggressions, from a technological perspective, but without leaving aside the different sociological considerations related to the problem. Within this context, and considering the technological potential of affective computing through physiological information to generate those new tools, we performed a detailed analysis regarding the disentangle of the relationship between physiological signals and fear-related emotions. This study provided us with the knowledge to propose a new approach to detect fear-related emotions making use of the different emotional theories and physiological affective indicators. This study was also accompanied by a comprehensive investigation regarding emotion-provoking tools, emotion assessment reports, emotion classification databases, affective computing systems design, and related methodologies and tools that allowed us to build a solid technological knowledge base to fulfil the challenges of this PhD.

Then, the fear binary classification approach has been included in different affec-

tive computing systems constructed onto publicly available datasets. Specifically, different specialised fear detection systems using time, frequency and non-linear domain features have been designed. The added value of the proposed architectures is the consideration of digital processing constraints to further properly embed such system into a wearable edge-device platform for allowing protection of vulnerable people. During the design of these systems, different limitations were spotted within the open available databases we were working with. For instance, there were no use of emotional immersive technology, the labelling methodology was not considering the gender perspective, a properly balanced stimuli distribution regarding the target emotions was not always assured, and the integration of a recovery processes based on the physiological signals of the volunteers to quantify and isolate the emotional activation between stimuli were not implemented. However, the proposed systems were successfully compared against the state-of-the-art.

Together with the design and validation of the different fear classification systems, a new wearable hardware solution to deploy the fear-related detection system architectures was proposed. Thus, we designed Bindi, an autonomous multimodal system towards the detection of risky situations under Gender-based Violence contexts. The edge-computing part of the system is a smart cyberphysical network. Specifically, this is accomplished by means of physiological and physical (audio and/or speech) smart sensors continuously monitoring the user. The fog-based layer of the system resides into a multimodal data fusion within an ad-hoc smartphone application. Moreover, the information is sent to specific computing servers in the cloud, which are responsible to store the collected data for further legal actions. The design of such a system can boost the generation of new mechanisms for the prevention and fight against Gender-based Violence.

Finally, after having identified the need for generating a new database and created a new technological tool, we designed and carried out the WEMAC dataset. It consists of 104 women who never experienced Gender-based Violence that performed different emotion-related stimuli visualisations in a laboratory environment. The previous fear binary classification systems were improved and applied to this novel multimodal dataset, leading up to competitive results in comparison with the state-of-the-art.

7.1 Contributions

To be more precise, we will sort the contributions in function of the Chapter in which they are made.

The contributions on Chapter 4 are the following:

- The application and validation of a new fear binary classification proposal using open available datasets and a reduced set of physiological signals.
- The design and evaluation of a fear classification system employing the DEAP database and the PA model. It achieved an AUC of 81.60% and a Gmean of 81.55% on average for a subject-independent approach and only two physiological signals (PPG and GSR).
- The design and evaluation of a fear classification system employing the MAHNOB database and the PAD model. It achieved an AUC of 86.00% and a Gmean of 73.78% on average for a subject-independent approach and only three physiological signals (PPG, GSR, and SKT). Note that this system was tested using LOSO.

The contributions on Chapter 5 are the following:

- The design, hardware and software, of a new smart-wearable system based on a reduced set of physiological signals and targeting the generating of new technological mechanisms and tools to prevent and combat Gender-based Violence.
- A simplified Signal-Quality-Assessment low-complexity fuzzy rule-base Mamdani inference model training design and implementation into the bracelet of Bindi. This is accompanied by a proposal definition and implementation of a novel online unsupervised fine-tuning based via scaled similarity between interval type II fuzzy sets for model self-adaptive updates. Results show that the system achieved overall accuracy of 93.72%. The proposed quality-aware system presents an energy consumption of up to 59.40 mJ, which directly impacts the overall energy consumption from 1.5% to 20.7% for transmission of noisy 12–60 seconds photoplethysmography signal.
- Different filtering strategies and feature extraction techniques were implemented into the bracelet of Bindi. This is accompanied by a successful measurement and results comparison with a specific research-grade toolkit.

The contributions on Chapter 6 are the following:

- The definition of the WEMAC dataset[1]. This is a collection of experiments captured in laboratory conditions with women volunteers. A set of audiovisual stimuli are employed to elicit realistic emotions using virtual reality and acquiring volunteers' physiological and speech information. Additionally, self-reported emotional annotations on dimensional and discrete emotional scales are also collected. The objectives and contributions of this novel multimodal dataset are multiple, as briefly shown below:
  1. The integration of immersive technology to elicit emotions. Virtual reality is employed as it offers the closest resemblance to real world scenarios, offering a high degree of correlation between the research conditions and the emotional phenomenon under study, i.e. with ecological validity.
  2. The consideration of a high number of volunteers. The first experiment accounted for a total of 104 non-Gender-based Violence women volunteers.
  3. The application of a properly balanced stimuli distribution regarding the target emotions. Prior to the generation of the dataset, a mixed methodology with expert judges and general public was applied to select the best audio-visual stimuli for provoking emotional reactions. A public pool was run with $1,332$ participants for labelling the pre-selected emotion-related stimuli.
  4. The modification of the labelling methodology to consider the gender perspective. This problem was addressed by changing the original Self-Assessment Manikins.
  5. The implementation of an active recovery process regarding the physiological stabilisation between stimuli. To the best of our knowledge, there is no public dataset that implemented an online stabilisation evaluation by means of physiological feedback assessment during the experiments.

  Amongst these contributions, I have been directly involved within into objectives $1, 2, 4$, and 5.
- The first experimental multimodal results with WEMAC. These show an aver-

[1]For the generation of this dataset, very hard team work has been required. For instance, a global amount of 7000 hours have been employed.

age accuracy of the fear recognition rate of up to 63.61% with the Leave-hAlf-Subject-Out (LASO) method. To the best of my knowledge, this is the first time a LASO model considering fear recognition, multisensorial signal fusion, and virtual reality stimuli has been presented.

7.2 Future work

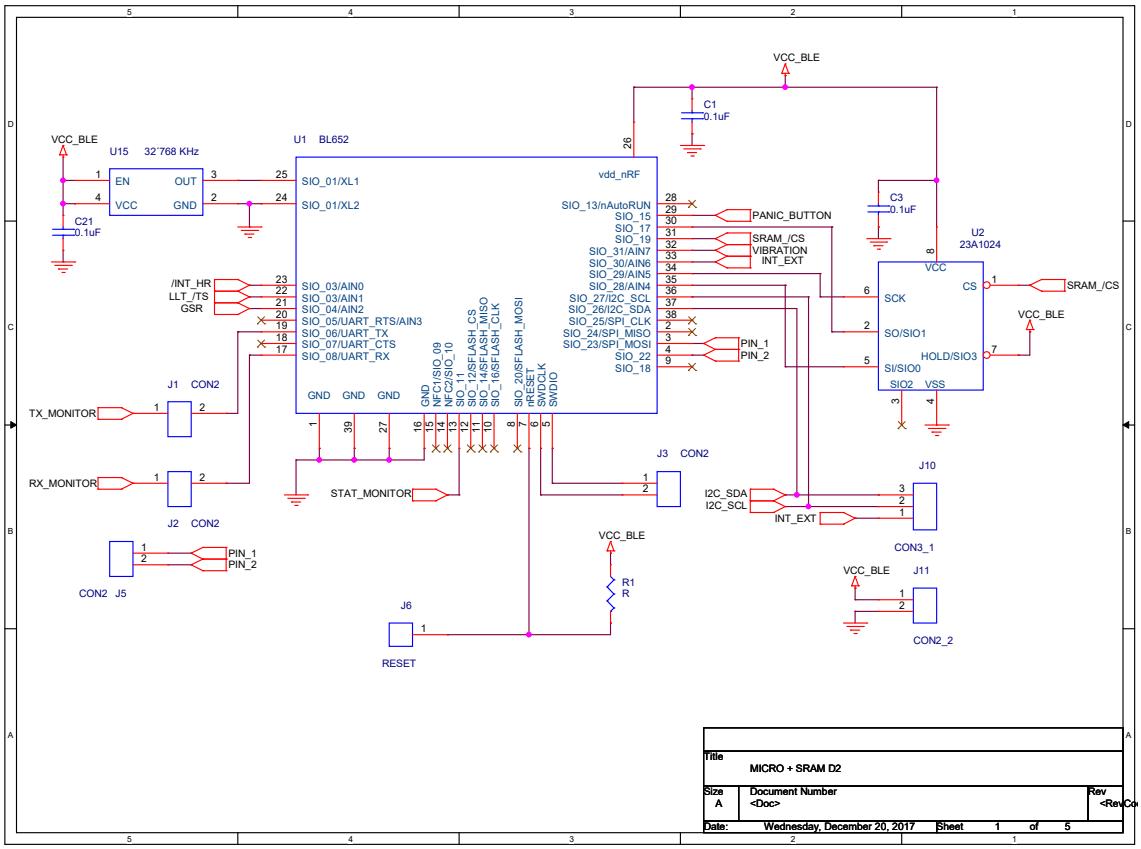
The author would like to provide some suggestions for future research:

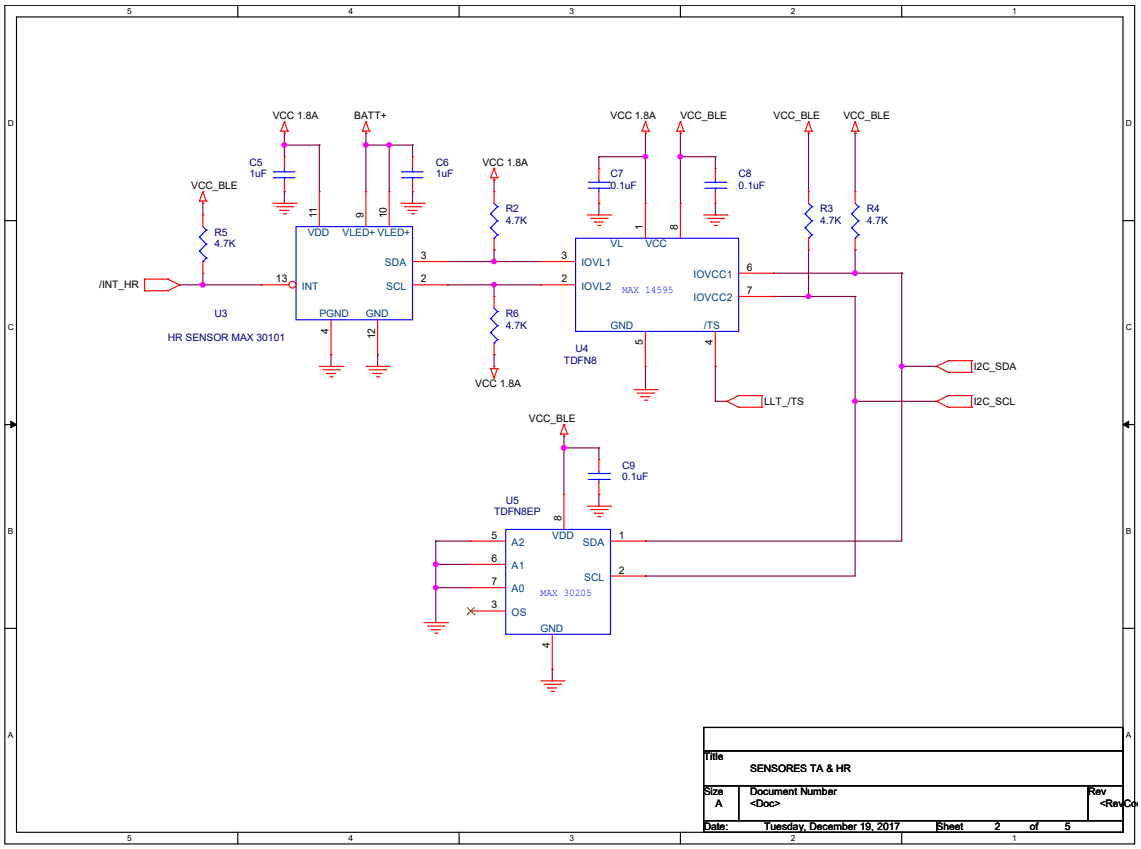
- The consideration of more physiological signals, or even biological ones, to expand this research scope and improve the fear classification results.
- Multivariate analysis considering the initial questionnaires collected at the beginning of the WEMAC experiments together with the physiological and auditory information.
- In case of collecting discrete and dimensional self-reports from volunteers, both label methodologies could be fused together by means of a lineal or non-linear combination. This is in line with the fact that both discrete and dimensional labels exist, but are intended to different purposes or characterise different aspects of the emotions.
- The research and implementation of compressed sensing techniques to reduce the power consumption of the bracelet. Work on this topic has been already initiated.
- The research and implementation of energy harvesting techniques within the bracelet would be interesting to observe the power consumption effect. Work on this topic has been already initiated.
- The research and integration of semi-supervised classification systems intended to deal with the emotion dynamics and/or weakly-supervised Learning for fine-grained emotion recognition using physiological signals. Work on this topic has been already initiated.
- The integration of neuromorphic computing into Bindi, such as Akida Neural Processor SoC.
- The implementation of novel motion artefact removal algorithms. For instance, synchrosqueezing techniques together with end-to-end edge-computing friendly neural networks have potential. Work on this topic has been already initiated.

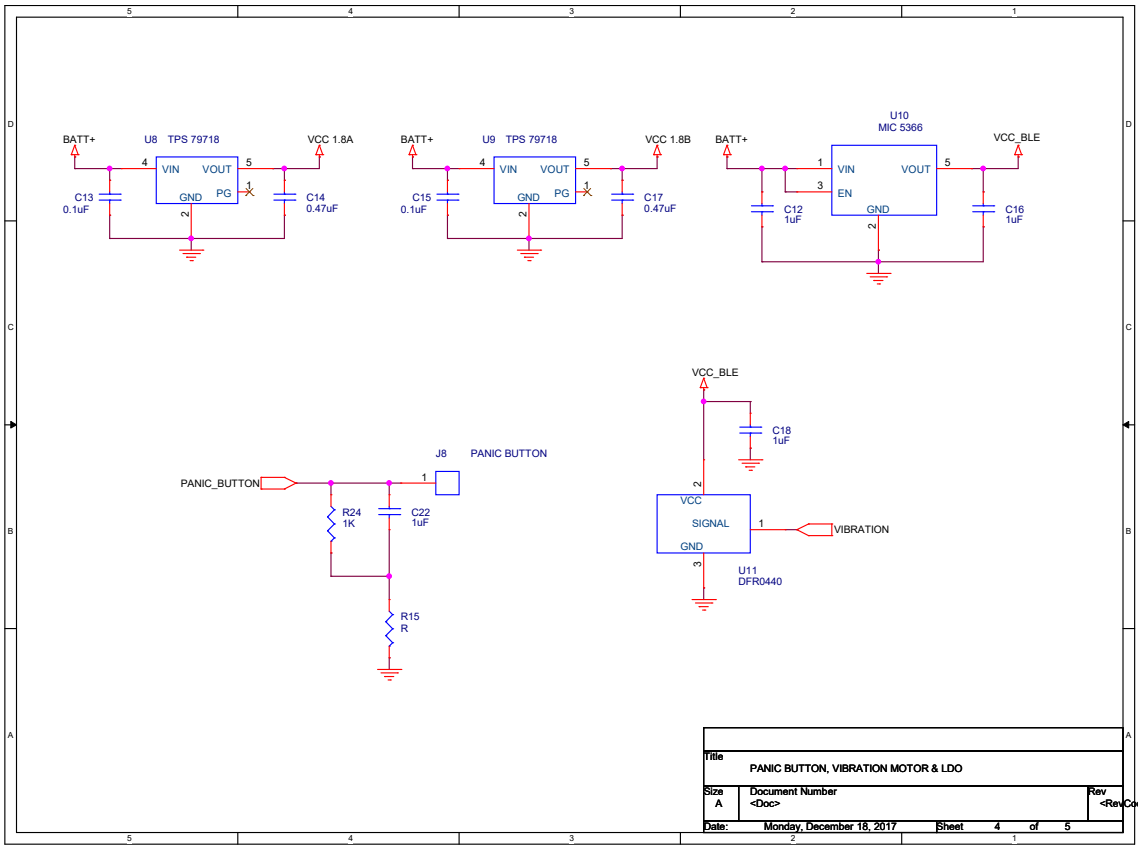
- The research of novel machine learning personalisation techniques and methods would boost the deployment possibilities of Bindi. Work on this topic has been already initiated.
- The design of new wearable form factors, rather than a bracelet and a pendant.
- The design of an expert system to be running in the cloud and operating in a multivariate basis. The goal of such system would be to correct or modify the edge-computing machine learning.

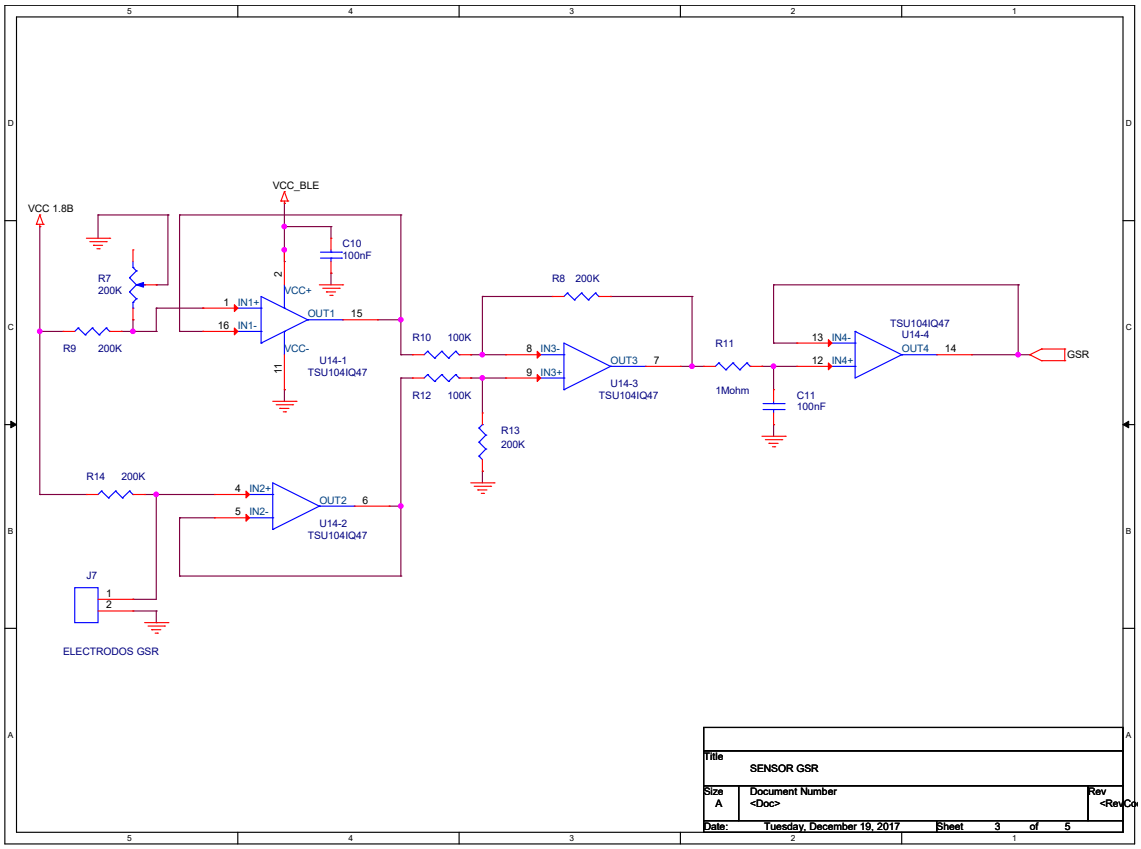
Appendix A

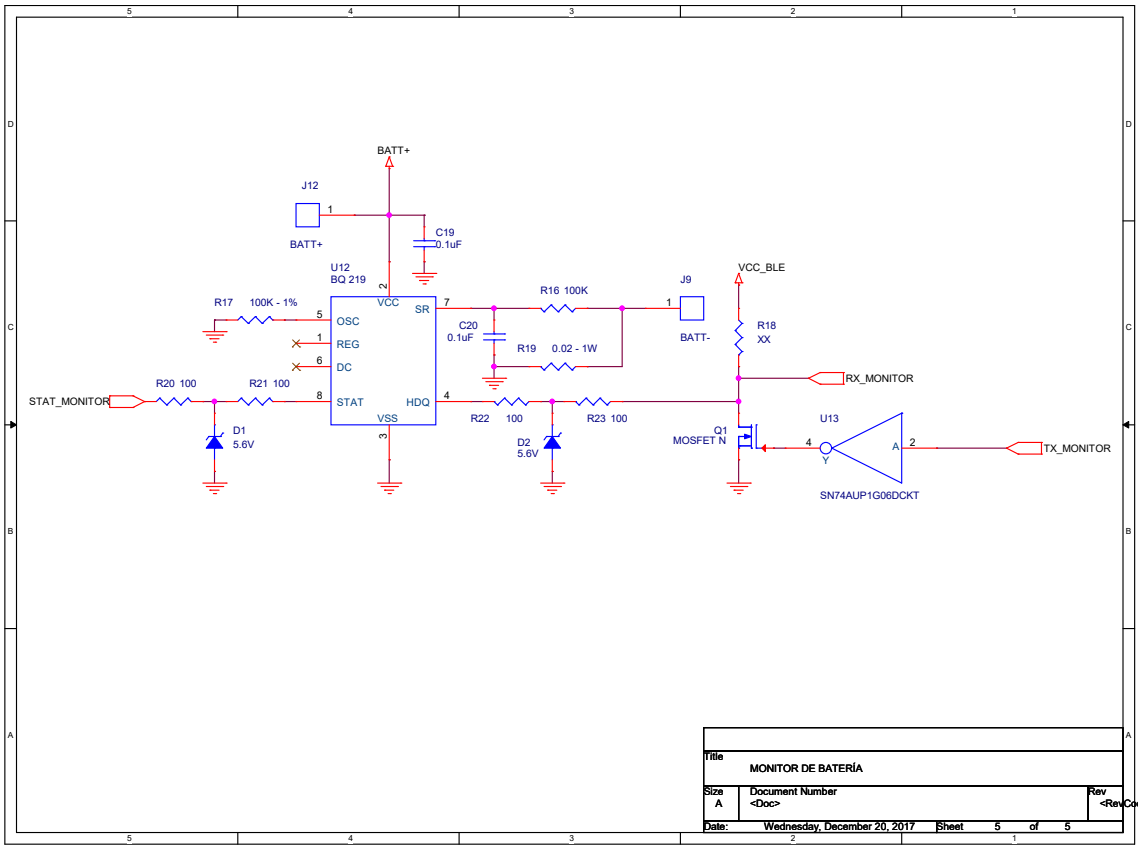
Bracelet schematics




VCC 1.8A
BATT+
VCC_BLE
C5 1uF
C6 1uF
R5 4.7K
R2 4.7K
R6 4.7K
/INT_HR
U3
HR SENSOR MAX 30101
VDD
VLED+
SDA
SCL
INT
PGND
GND
C7 0.1uF
C8 0.1uF
R3 4.7K
R4 4.7K
VL
VCC
IOVL1
IOVL2
MAX 14595
IOVCC1
IOVCC2
/TS
U4
TDFN8
LLT_/TS
I2C_SDA
I2C_SCL
C9 0.1uF
U5
TDFN8EP
A2
A1
A0
OS
MAX 30205
Title
SENSORES TA & HR
Size
A
Document Number
<Doc>
Rev
Date: Tuesday, December 19, 2017
Sheet 2 of 5


BATT+
U8 TPS 79718
VCC 1.8A
VIN
VOUT
GND
PG
C13 0.1uF
C14 0.47uF
U9 TPS 79718
VCC 1.8B
C15 0.1uF
C17 0.47uF
U10
MIC 5366
EN
C12 1uF
C16 1uF
VCC_BLE
J8
PANIC BUTTON
PANIC_BUTTON
R24 1K
C22 1uF
R15 R
C18 1uF
VCC
SIGNAL
GND
VIBRATION
U11
DFR0440
Title
PANIC BUTTON, VIBRATION MOTOR & LDO
Size
A
Document Number
<Doc>
Rev
Date: Monday, December 18, 2017
Sheet 4 of 5


VCC_BLE
VCC 1.8B
C10
100nF
R7
200K
R9
200K
U14-1
TSU104IQ47
R8 200K
R10 100K
R12 100K
R13
200K
U14-3
TSU104IQ47
R11
1Mohm
C11
100nF
U14-4
TSU104IQ47
GSR
R14 200K
U14-2
TSU104IQ47
J7
ELECTRODOS GSR
Title
SENSOR GSR
Size
A
Document Number
<Doc>
Rev
Date: Tuesday, December 19, 2017
Sheet 3 of 5


BATT+
J12
BATT+
C19
0.1uF
U12
BQ 219
R17 100K - 1%
R16 100K
J9
BATT-
VCC_BLE
R18
XX
C20
0.1uF
R19 0.02 - 1W
RX_MONITOR
R20 100
R21 100
STAT_MONITOR
D1
5.6V
R22 100
R23 100
D2
5.6V
Q1
MOSFET N
U13
SN74AUP1G06DCKT
TX_MONITOR
Title
MONITOR DE BATERÍA
Size
A
Document Number
<Doc>
Rev
Date: Wednesday, December 20, 2017
Sheet 5 of 5

Bibliography

[1] S. G. D. against Gender Violence. (2020) Fact-sheets on gender violence victims killed since 2003. [Online]. Available: violenciagenero.igualdad.gob.es/violenciaEnCifras/victimasMortales/fichaMujeres/home.htm

[2] I. Bakker, T. van der Voordt, P. Vink, and J. de Boon, "Pleasure, arousal, dominance: Mehrabian and russell revisited," *Current Psychology*, vol. 33, no. 3, pp. 405–421, Sep 2014.

[3] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, 2007. [Online]. Available: http://www.jstor.org/stable/40064702

[4] M. M. Bradley and P. J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[5] A. Ferrari, D. Micucci, M. Mobilio, and P. Napoletano, "On the personalization of classification models for human activity recognition," *IEEE Access*, vol. 8, pp. 32 066–32 079, 2020.

[6] M. Malik, J. T. Bigger, A. J. Camm, R. E. Kleiger, A. Malliani, A. J. Moss, and P. J. Schwartz, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use," *European Heart Journal*, vol. 17, no. 3, pp. 354–381, 03 1996.

[7] D. of the Spanish Government against Gender Violence, "Dispositivos de control telemático de medidas y penas de alejamiento," https://violenciagenero.igualdad.gob.es/informacionUtil/recursos/dispositivosControlTelematico/home.htm, (Accessed on 04/04/2021).

[8] *High-Sensitivity Pulse Oximeter and Heart-Rate Sensor for Wearable Health*, Maxim Integrated, 2020, pPG. [Online]. Available: https://www.mouser.es/datasheet/2/744/Seeed_105020003-1217653.pdf

[9] J. A. Miranda, E. Rituerto-González, M. F. Canabal, A. R. Bárcenas, J. M. Lanza-Gutiérrez, C. Pelaez-Moreno, and C. López-Ongil, "Bindi: Affective internet of things to combat gender-based violence," *IEEE Internet of Things*, 2022, manuscript submitted for publication.

[10] J. Lichtenauer and M. Soleymani, "Mahnob-hci-tagging database," 2011.

[11] U. Nations, "Declaration on the elimination of violence against women," 1993.

[12] E. Commission. (2020) Gender violence definition and forms. [Online]. Available: https://ec.europa.eu/info/policies/justice-and-fundamental-rights/gender-equality/gender-based-violence/what-gender-based-violence

[13] L. Sardinha *et al.*, "Global, regional, and national prevalence estimates

of physical or sexual, or both, intimate partner violence against women in 2018," *The Lancet*, 2 2022. [Online]. Available: https://doi.org/10.1016/S0140-6736(21)02664-7

[14] H. A. T. DIRECTIVE, "Council directive 75/117/eec of 10 february 1975 on the approximation of the laws of the member states relating to the application of the principle of equal pay for men and women," *Official Journal L*, vol. 45, no. 19/02, pp. 0019–0020, 1975.

[15] G. Maganza, "The lisbon treaty: A brief outline," *Fordham Int'l LJ*, vol. 31, p. 1603, 2007.

[16] J. Ulla and S. Rosamund, "The istanbul convention: A tool to tackle violence against women and girls," *European Parliamentary Research Service*, 2020.

[17] E. Commission. (2020) Actions against women violence. [Online]. Available: https://ec.europa.eu/justice/saynostopvaw/eu-actions.html

[18] M. Segrave and L. Vitis, Eds., *Gender, Technology and Violence*, 1st ed., ser. Routledge Studies in Crime and Society. United Kingdom: Routledge, 2017.

[19] T. Martínez, "Un recorrido por el sistema institucional en el ámbito de la violencia de género," *Revista de Estudios Socioeducativos. ReSed*, no. 7, pp. 256–257, 2019.

[20] J. J. López-Ossorio, J. L. González-Álvarez, and A. Andrés-Pueyo, "Predictive effectiveness of the police risk assessment in intimate partner violence," *Psychosocial Intervention*, vol. 25, pp. 1 – 7, 04 2016.

[21] M. de Sanidad. (2020) Atenpro. spanish social services. [Online]. Available: https://www.mscbs.gob.es/en/ssi/violenciaGenero/Recursos/ATENPRO/home.htm

[22] L. Arenas García, "The efficacy of electronic monitoring in gender violence: criminological analysis," *International e-Journal of Criminal Sciences*, no. 10, 2016.

[23] R. S. Recio, E. G. Alberola, C. I. F. Guarné *et al.*, "Prevention of violence against women: policies and actions on gender violence," *Informació Psicològica*, no. 111, pp. 35–50, 2016.

[24] P. Schmidt, A. Reiss, R. Dürichen, and K. V. Laerhoven, "Wearable-based affect recognition—a review," *Sensors*, vol. 19, no. 19, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/19/4079

[25] C. Morgan, *Introduction to Psychology*. McGraw-Hill, 1961.

[26] C. Darwin, P. Ekman, and P. Prodger, *The Expression of the Emotions in Man and Animals*. Oxford University Press, 1998.

[27] A. Colman, *A Dictionary of Psychology*, ser. Oxford Dictionary of Psychology. Oxford University Press, 2009.

[28] W. B. Cannon, "The james-lange theory of emotions: A critical examination and an alternative theory," *The American Journal of Psychology*, vol. 39, no. 1/4, pp. 106–124, 1927.

[29] R. S. Lazarus, "Progress on a cognitive-motivational-relational theory of emotion." *American Psychologist*, vol. 46, no. 8, pp. 819–834, 1991.

[30] A. Ellis, *Rational-emotive theory: Albert Ellis.*, ser. Operational theories of personality. Oxford, England: Brunner/Mazel, 1974, pp. x, 421–x, 421.

[31] D. G. MacKay, M. A. Shafto, J. K. Taylor, D. E. Marian, L. Abrams, and J. R. Dyer, "Relations between emotion, memory, and attention: Evidence from taboo stroop, lexical decision, and immediate memory tasks," *Memory*

& *Cognition*, vol. 32, pp. 474–488, 2004.
[32] P. Ekman, "What scientists who study emotion agree about," *Perspectives on Psychological Science*, vol. 11, no. 1, pp. 31–34, 2016, pMID: 26817724.
[33] A. Konar and A. Chakraborty, "Introduction to emotion recognition," in *Emotion Recognition: A Pattern Analysis Approach*, 2015.
[34] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.
[35] P. Ekman and W. V. Friesen, "Measuring facial movement," *Environmental psychology and nonverbal behavior*, vol. 1, no. 1, pp. 56–75, 1976.
[36] C. E. Izard, *Theories of Emotion and Emotion-Behavior Relationships*. Springer US, 1977, pp. 19–42.
[37] R. Plutchik, "Emotions: A general psychoevolutionary theory," *Approaches to emotion*, vol. 1984, pp. 197–219, 1984.
[38] N. H. Frijda, *The Emotions*. Cambridge University Press, 1986.
[39] K. Oatley and P. N. Johnson-laird, "Towards a cognitive theory of emotions," *Cognition and Emotion*, vol. 1, no. 1, pp. 29–50, 1987.
[40] P. Ekman and D. Cordaro, "What is meant by calling emotions basic," *Emotion review*, vol. 3, no. 4, pp. 364–370, 2011.
[41] B. Mesquita, "The legacy of nico h.frijda (1927–2015)," *Cognition and Emotion*, vol. 30, no. 4, pp. 603–608, 2016, pMID: 26943647.
[42] P. Ekman, "What scientists who study emotion agree about," *Perspectives on Psychological Science*, vol. 11, no. 1, pp. 31–34, 2016, pMID: 26817724.
[43] W. Wundt, "Vorselung über die menschen – und tierseele," *Voss Verlag: Leipzig, Germany*, pp. 145–172, 1863.
[44] C. A. Clark, "Book reviews : The measurement of meaning by charles e. osgood, george j. suci, and percy h. tannenbaum. urbana, illinois: University of illinois press, 1957. 342 p. $7.50," *Educational and Psychological Measurement*, vol. 18, no. 4, pp. 884–886, 1958.
[45] J. Russell, "A circumplex model of affect," *Journal of personality and social psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.
[46] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology*, vol. 14, no. 4, pp. 261–292, Dec 1996.
[47] H. A. Demaree, D. E. Everhart, E. A. Youngstrom, and D. W. Harrison, "Brain lateralization of emotional processing: Historical roots and a future incorporating "dominance"," *Behavioral and Cognitive Neuroscience Reviews*, vol. 4, no. 1, pp. 3–20, 2005, pMID: 15886400.
[48] E. Harmon-Jones, C. Harmon-Jones, and E. Summerell, "On the importance of both dimensional and discrete models of emotion," *Behavioral Sciences*, vol. 7, no. 4, 2017. [Online]. Available: https://www.mdpi.com/2076-328X/7/4/66
[49] T. Thanapattheerakul, K. Mao, J. Amoranto, and J. H. Chan, "Emotion in a century: A review of emotion recognition," in *Proceedings of the 10th International Conference on Advances in Information Technology*, 2018, pp. 1–8.
[50] C. Castelfranchi, *Affective Appraisal versus Cognitive Evaluation in Social Emotions and Interactions*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2000, pp. 76–106.
[51] E. Komulainen, K. Meskanen, J. Lipsanen, J. M. Lahti, P. Jylha, T. Melartin,

M. Wichers, E. Isometsa, and J. Ekelund, "The effect of personality on daily life emotional processes," *PLOS ONE*, vol. 9, pp. 1–9, 10 2014. [Online]. Available: https://doi.org/10.1371/journal.pone.0110907

[52] M. A. A. Mengual, *Trastorno de estrés postraumático: Daño cerebral secundario a la violencia (mobbing, violencia de género, acoso escolar).* Ediciones Díaz de Santos, 2007.

[53] M. Bianchin and A. Angrilli, "Gender differences in emotional responses: A psychophysiological study," *Physiology and Behavior*, vol. 105, no. 4, pp. 925–932, 2012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031938411005221

[54] L. Lambrecht, B. Kreifelts, and D. Wildgruber, "Gender differences in emotion recognition: Impact of sensory modality and emotional category," *Cognition & emotion*, vol. 28, no. 3, pp. 452–469, 2014.

[55] X. Chen, H. Yuan, T. Zheng, Y. Chang, and Y. Luo, "Females are more sensitive to opponent's emotional feedback: Evidence from event-related potentials," *Frontiers in Human Neuroscience*, vol. 12, p. 275, 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnhum.2018.00275

[56] S. Dobrišek, R. Gajšek, F. Mihelič, N. Pavešić, and V. Štruc, "Towards efficient multi-modal emotion recognition," *International Journal of Advanced Robotic Systems*, vol. 10, no. 1, p. 53, 2013.

[57] M. Á. Blanco Ruiz, L. Gutiérrez Martín, J. Á. Miranda Calero, M. F. Canabal Benito, E. Rituerto González, C. Luis Mingueza, J. C. Robredo García, B. Morán González, A. Páez Montoro, A. Ramírez Bárcenas *et al.*, "Uc3m4safety database description," http://hdl.handle.net/10016/32481, 2021.

[58] P. J. Lang, "International affective picture system (iaps): Affective ratings of pictures and instruction manual," *Technical report*, 2005.

[59] A. Marchewka, Ł. Żurawski, K. Jednoróg, and A. Grabowska, "The nencki affective picture system (naps): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database," *Behavior research methods*, vol. 46, no. 2, pp. 596–610, 2014.

[60] B. Kurdi, S. Lozano, and M. R. Banaji, "Introducing the open affective standardized image set (oasis)," *Behavior research methods*, vol. 49, no. 2, pp. 457–470, 2017.

[61] E. S. Dan-Glauser and K. R. Scherer, "The geneva affective picture database (gaped): a new 730-picture database focusing on valence and normative significance," *Behavior research methods*, vol. 43, no. 2, pp. 468–477, 2011.

[62] M. Wessa, P. Kanske, P. Neumeister, K. Bode, J. Heissler, S. Schönfelder *et al.*, "Emopics: Subjektive und psychophysiologische evaluation neuen bildmaterials für die klinisch-bio-psychologische forschung," *Zeitschrift für Klinische Psychologie und Psychotherapie*, vol. 39, no. Suppl. 1/11, p. 77, 2010.

[63] A. Haberkamp, J. A. Glombiewski, F. Schmidt, and A. Barke, "The disgust-related-images (dirti) database: Validation of a novel standardized set of disgust pictures," *Behaviour research and therapy*, vol. 89, pp. 86–94, 2017.

[64] A. C. Samson, S. D. Kreibig, B. Soderstrom, A. A. Wade, and J. J. Gross, "Eliciting positive, negative and mixed emotional states: A film library for affective scientists," *Cognition and emotion*, vol. 30, no. 5, pp. 827–856, 2016.

[65] A. Di Crosta, P. La Malva, C. Manna, A. Marin, R. Palumbo, M. C. Verroc-

chio, M. Cortini, N. Mammarella, and A. Di Domenico, "The chieti affective action videos database, a resource for the study of emotions in psychology," *Scientific data*, vol. 7, no. 1, pp. 1–6, 2020.

[66] T. L. Gilman, R. Shaheen, K. M. Nylocks, D. Halachoff, J. Chapman, J. J. Flynn, L. M. Matt, and K. G. Coifman, "A film set for the elicitation of emotion in research: A comprehensive catalog derived from four decades of investigation," *Behavior research methods*, vol. 49, no. 6, pp. 2061–2082, 2017.

[67] S. Carvalho, J. Leite, S. Galdo-Álvarez, and O. F. Gonçalves, "The emotional movie database (emdb): A self-report and psychophysiological study," *Applied psychophysiology and biofeedback*, vol. 37, no. 4, pp. 279–294, 2012.

[68] K. Umla-Runge, H. D. Zimmer, X. Fu, and L. Wang, "An action video clip database rated for familiarity in china and germany," *Behavior Research Methods*, vol. 44, no. 4, pp. 946–953, 2012.

[69] T. B. Alakus, M. Gonen, and I. Turkoglu, "Database for an emotion recognition system based on eeg signals and various computer games–gameemo," *Biomedical Signal Processing and Control*, vol. 60, p. 101951, 2020.

[70] M. Granato, D. Gadia, D. Maggiorini, and L. A. Ripamonti, "Software and hardware setup for emotion recognition during video game fruition," in *Proceedings of the 4th EAI International Conference on Smart Objects and Technologies for Social Good*, ser. Goodtechs '18. Association for Computing Machinery, 2018, p. 19–24.

[71] M. Granato, "Emotions recognition in video game players using physiological information," Ph.D. dissertation, Universita Degli Studi Di Milano, 2019.

[72] C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer, "The 'trier social stress test'–a tool for investigating psychobiological stress responses in a laboratory setting," *Neuropsychobiology*, vol. 28, no. 1-2, pp. 76–81, 1993.

[73] L. A. Mitchell, R. A. MacDonald, and E. E. Brodie, "Temperature and the cold pressor test," *The Journal of Pain*, vol. 5, no. 4, pp. 233–237, 2004.

[74] F. Scarpina and S. Tagini, "The stroop color and word test," *Frontiers in psychology*, vol. 8, p. 557, 2017.

[75] A. L. Shilton, R. Laycock, and S. G. Crewther, "The maastricht acute stress test (mast): Physiological and subjective responses in anticipation, and post-stress," *Frontiers in psychology*, vol. 8, p. 567, 2017.

[76] J. Wijsman, B. Grundlehner, H. Liu, J. Penders, and H. Hermens, "Wearable physiological sensors reflect mental stress state in office-like situations," in *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*. IEEE, 2013, pp. 600–605.

[77] M. Granato, D. Gadia, D. Maggiorini, and L. A. Ripamonti, "Software and hardware setup for emotion recognition during video game fruition," in *Proceedings of the 4th EAI International Conference on Smart Objects and Technologies for Social Good*, 2018, pp. 19–24.

[78] N. S. Suhaimi, C. T. B. Yuan, J. Teo, and J. Mountstephens, "Modeling the affective space of 360 virtual reality videos based on arousal and valence for wearable eeg-based vr emotion classification," in *2018 IEEE 14th International Colloquium on Signal Processing & Its Applications (CSPA)*. IEEE, 2018, pp. 167–172.

[79] 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 re-

ality: emotion recognition from brain and heartbeat dynamics using wearable sensors," *Scientific reports*, vol. 8, no. 1, pp. 1–15, 2018.

[80] J. Marín-Morales, C. Llinares, J. Guixeres, and M. Alcañiz, "Emotion recognition in immersive virtual reality: From statistics to affective computing," *Sensors*, vol. 20, no. 18, p. 5163, 2020.

[81] L. Constantine and H. Hajj, "A survey of ground-truth in emotion data annotation," in *2012 IEEE International Conference on Pervasive Computing and Communications Workshops*, 2012, pp. 697–702.

[82] M. Blanco-Ruiz, C. Sainz-de Baranda, L. Gutiérrez-Martín, E. Romero-Perales, and C. López-Ongil, "Emotion elicitation under audiovisual stimuli reception: Should artificial intelligence consider the gender perspective?" *International Journal of Environmental Research and Public Health*, vol. 17, no. 22, 2020. [Online]. Available: https://www.mdpi.com/1660-4601/17/22/8534

[83] J. L. Andreassi, *Psychophysiology: Human behavior and physiological response*. Psychology Press, 2010.

[84] O. P. Keifer Jr, R. C. Hurt, K. J. Ressler, and P. J. Marvar, "The physiology of fear: reconceptualizing the role of the central amygdala in fear learning," *Physiology*, vol. 30, no. 5, pp. 389–401, 2015.

[85] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," *Biological psychology*, vol. 84, no. 3, pp. 394–421, 2010.

[86] L. F. Barrett, "Are emotions natural kinds?" *Perspectives on Psychological Science*, vol. 1, no. 1, pp. 28–58, 2006.

[87] I. B. Mauss and M. D. Robinson, "Measures of emotion: A review," *Cognition and Emotion*, vol. 23, no. 2, pp. 209–237, 2009.

[88] G. Stemmler, "Physiological processes during emotion," in *The regulation of emotion*. Psychology Press, 2004, pp. 48–85.

[89] R. W. Levenson, "The autonomic nervous system and emotion," *Emotion Review*, vol. 6, no. 2, pp. 100–112, 2014.

[90] A. F. AX, "The physiological differentiation between fear and anger in humans," *Psychosomatic Medicine*, vol. 15, no. 5, 1953.

[91] O. Faust and M. G. Bairy, "Nonlinear analysis of physiological signals: a review," *Journal of Mechanics in Medicine and Biology*, vol. 12, no. 04, p. 1240015, 2012.

[92] S. Z. Spasić and S. Kesić, "Editorial: Nonlinearity in living systems: Theoretical and practical perspectives on metrics of physiological signal complexity," *Frontiers in Physiology*, vol. 10, p. 298, 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fphys.2019.00298

[93] Q. Yousef, M. Reaz, and M. A. M. Ali, "The analysis of ppg morphology: investigating the effects of aging on arterial compliance," *Measurement Science Review*, vol. 12, no. 6, p. 266, 2012.

[94] J. Fine, K. L. Branan, A. J. Rodriguez, T. Boonya-Ananta, J. C. Ramella-Roman, M. J. McShane, G. L. Coté *et al.*, "Sources of inaccuracy in photoplethysmography for continuous cardiovascular monitoring," *Biosensors*, vol. 11, no. 4, p. 126, 2021.

[95] G. Tusman, C. M. Acosta, S. Pulletz, S. H. Böhm, A. Scandurra, J. M. Arca, M. Madorno, and F. S. Sipmann, "Photoplethysmographic characterization of vascular tone mediated changes in arterial pressure: an observational study,"

Journal of clinical monitoring and computing, vol. 33, no. 5, pp. 815–824, 2019.

[96] N. Hayashi, N. Someya, T. Maruyama, Y. Hirooka, M. Y. Endo, and Y. Fukuba, "Vascular responses to fear-induced stress in humans," *Physiology & behavior*, vol. 98, no. 4, pp. 441–446, 2009.

[97] P. Shi, V. Azorin-Peris, A. S. Echiadis, J. Zheng, Y. Zhu, P.-Y. Cheang, and S. Hu, "Non-contact reflection photoplethysmography towards effective human physiological monitoring," *Journal of Medical and Biological Engineering*, 2010.

[98] A. Alzahrani, S. Hu, V. Azorin-Peris, L. Barrett, D. Esliger, M. Hayes, S. Akbare, J. Achart, and S. Kuoch, "A multi-channel opto-electronic sensor to accurately monitor heart rate against motion artefact during exercise," *Sensors*, vol. 15, no. 10, pp. 25 681–25 702, 2015.

[99] V. Rybynok and P. Kyriacou, "Beer-lambert law along non-linear mean light pathways for the rational analysis of photoplethysmography," in *Journal of Physics: Conference Series*, vol. 238, no. 1. IOP Publishing, 2010, p. 012061.

[100] T. Shimazaki, S. Hara, H. Okuhata, H. Nakamura, and T. Kawabata, "Cancellation of motion artifact induced by exercise for ppg-based heart rate sensing," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2014, pp. 3216–3219.

[101] H. Nogami, W. Iwasaki, N. Morita, and R. Takigawa, "Relationship between ac/dc ratio and light-blocking structure for reflective photoplethysmographic sensor," *Sensors and Materials*, vol. 30, no. 12, pp. 3021–3028, 2018.

[102] V. Hartmann, H. Liu, F. Chen, Q. Qiu, S. Hughes, and D. Zheng, "Quantitative comparison of photoplethysmographic waveform characteristics: Effect of measurement site," *Frontiers in Physiology*, vol. 10, p. 198, 2019.

[103] B. Bent, B. A. Goldstein, W. A. Kibbe, and J. P. Dunn, "Investigating sources of inaccuracy in wearable optical heart rate sensors," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–9, 2020.

[104] B. Baldaro, M. W. Battacchi, M. Codispoti, G. Tuozzi, G. Trombini, R. Bolzani, and D. Palomba, "Modifications of electrogastrographic activity during the viewing of brief film sequences," *Perceptual and motor skills*, vol. 82, no. 3_suppl, pp. 1243–1250, 1996.

[105] U. Dimberg, "Facial reactions to fear-relevant and fear-irrelevant stimuli," *Biological psychology*, vol. 23, no. 2, pp. 153–161, 1986.

[106] B. L. Fredrickson and R. W. Levenson, "Positive emotions speed recovery from the cardiovascular sequelae of negative emotions," *Cognition & emotion*, vol. 12, no. 2, pp. 191–220, 1998.

[107] R. Gilissen, M. J. Bakermans-Kranenburg, M. H. van IJzendoorn, and R. van der Veer, "Parent–child relationship, temperament, and physiological reactions to fear-inducing film clips: Further evidence for differential susceptibility," *Journal of Experimental Child Psychology*, vol. 99, no. 3, pp. 182–195, 2008.

[108] J. A. Etzel, E. L. Johnsen, J. Dickerson, D. Tranel, and R. Adolphs, "Cardiovascular and respiratory responses during musical mood induction," *International Journal of psychophysiology*, vol. 61, no. 1, pp. 57–69, 2006.

[109] Y. Wu, R. Gu, Q. Yang, and Y.-j. Luo, "How do amusement, anger and fear influence heart rate and heart rate variability?" *Frontiers in Neuroscience*,

vol. 13, p. 1131, 2019.

[110] E.-H. Jang, S. Byun, M.-S. Park, and J.-H. Sohn, "Predicting individuals' experienced fear from multimodal physiological responses to a fear-inducing stimulus," *Advances in cognitive psychology*, vol. 16, no. 4, p. 291, 2020.

[111] M. J. Christie, "Electrodermal activity in the 1980s: A review," *Journal of the Royal Society of Medicine*, vol. 74, no. 8, pp. 616–622, 1981, pMID: 7288800. [Online]. Available: https://doi.org/10.1177/014107688107400812

[112] W. Boucsein, *Electrodermal Activity*, ser. The Springer series in behavioral psychophysiology and medicine. Springer US, 2012. [Online]. Available: https://books.google.es/books?id=6N6rnOEZEEoC

[113] M. Schmelz, R. Schmidt, A. Bickel, H. Torebjork, and H. Handwerker, "Innervation territories of single sympathetic c fibers in human skin," *Journal of neurophysiology*, vol. 79, no. 4, pp. 1653–1660, 1998.

[114] P. Ellaway, A. Kuppuswamy, A. Nicotra, and C. Mathias, "Sweat production and the sympathetic skin response: Improving the clinical assessment of autonomic function," *Autonomic Neuroscience*, vol. 155, no. 1, pp. 109 – 114, 2010.

[115] R. Edelberg, "Electrodermal mechanisms: A critique of the two-effector hypothesis and a proposed replacement," in *Progress in electrodermal research*. Springer, 1993, pp. 7–29.

[116] H. F. Posada-Quintero and K. H. Chon, "Innovations in electrodermal activity data collection and signal processing: A systematic review," *Sensors*, vol. 20, no. 2, 2020. [Online]. Available: https://www.mdpi.com/1424-8220/20/2/479

[117] W. Boucsein, *Methods of Electrodermal Recording*. Boston, MA: Springer US, 2012, pp. 87–258.

[118] S. Grimnes, A. Jabbari, Ø. G. Martinsen, and C. Tronstad, "Electrodermal activity by dc potential and ac conductance measured simultaneously at the same skin site," *Skin Research and Technology*, vol. 17, no. 1, pp. 26–34, 2011. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1600-0846.2010.00459.x

[119] W. Boucsein, D. Fowles, S. Grimnes, G. Ben-Shakhar, W. Roth, M. Dawson, and D. Filion, "Publication recommendations for electrodermal measurements," *Psychophysiology*, vol. 49, pp. 1017–34, 08 2012.

[120] R. Edelberg, T. Greiner, and N. R. Burch, "Some membrane properties of the effector in the galvanic skin response," *Journal of Applied Physiology*, vol. 15, no. 4, pp. 691–696, 1960, pMID: 13819259. [Online]. Available: https://doi.org/10.1152/jappl.1960.15.4.691

[121] M.-Z. Poh, N. C. Swenson, and R. W. Picard, "A wearable sensor for unobtrusive, long-term assessment of electrodermal activity," *IEEE transactions on Biomedical engineering*, vol. 57, no. 5, pp. 1243–1252, 2010.

[122] J. Guerreiro, "A biosignal embedded system for physiological computing," Ph.D. dissertation, Instituto Superior de Engenharia de Lisboa, 2013.

[123] G. C. Pope and R. J. Halter, "Design and implementation of an ultra-low resource electrodermal activity sensor for wearable applications," *Sensors*, vol. 19, no. 11, p. 2450, 2019.

[124] R. Edelberg, "Electrical properties of the skin," *Methods in psychophysiology*, 1967.

[125] J. J. Braithwaite, D. G. Watson, R. Jones, and M. Rowe, "A guide for

analysing electrodermal activity (eda) and skin conductance responses (scrs) for psychological experiments," *Psychophysiology*, vol. 49, no. 1, pp. 1017–1034, 2013.

[126] M. Benedek and C. Kaernbach, "Decomposition of skin conductance data by means of nonnegative deconvolution," *psychophysiology*, vol. 47, no. 4, pp. 647–658, 2010.

[127] A. Greco, G. Valenza, A. Lanata, E. P. Scilingo, and L. Citi, "cvxeda: A convex optimization approach to electrodermal activity processing," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 4, pp. 797–804, 2015.

[128] F. Hernando-Gallego, D. Luengo, and A. Artés-Rodríguez, "Feature extraction of galvanic skin responses by nonnegative sparse deconvolution," *IEEE journal of biomedical and health informatics*, vol. 22, no. 5, pp. 1385–1394, 2017.

[129] D. M. Alexander, C. Trengove, P. Johnston, T. Cooper, J. August, and E. Gordon, "Separating individual skin conductance responses in a short interstimulus-interval paradigm," *Journal of neuroscience methods*, vol. 146, no. 1, pp. 116–123, 2005.

[130] C. A. Frantzidis, E. Konstantinidis, C. Pappas, and P. D. Bamidis, "An automated system for processing electrodermal activity." *Studies in health technology and informatics*, vol. 150, pp. 787–787, 2009.

[131] J. Choi, B. Ahmed, and R. Gutierrez-Osuna, "Development and evaluation of an ambulatory stress monitor based on wearable sensors," *IEEE transactions on information technology in biomedicine*, vol. 16, no. 2, pp. 279–286, 2011.

[132] M. P. Tarvainen, P. O. Ranta-Aho, and P. A. Karjalainen, "An advanced detrending method with application to hrv analysis," *IEEE transactions on biomedical engineering*, vol. 49, no. 2, pp. 172–175, 2002.

[133] A. Dzedzickis, A. Kaklauskas, and V. Bucinskas, "Human emotion recognition: Review of sensors and methods," *Sensors*, vol. 20, no. 3, p. 592, 2020.

[134] L. B. et. al., "Allostasis and allostatic load: Woman abuse and chronic illness - learning network - western university," http://www.vawlearningnetwork.ca/our-work/briefs/brief-13.html, (Accessed on 08/20/2021).

[135] V. Kosonogov, L. De Zorzi, J. Honoré, E. S. Martínez-Velázquez, J.-L. Nandrino, J. M. Martinez-Selva, and H. Sequeira, "Facial thermal variations: A new marker of emotional arousal," *PLOS ONE*, vol. 12, no. 9, pp. 1–15, 09 2017. [Online]. Available: https://doi.org/10.1371/journal.pone.0183592

[136] C. Goulart, C. Valadão, D. Delisle-Rodriguez, D. Tavares, E. Caldeira, and T. Bastos-Filho, "Emotional state analysis through infrared thermal imaging," in *XXVI Brazilian Congress on Biomedical Engineering*, R. Costa-Felix, J. C. Machado, and A. V. Alvarenga, Eds. Singapore: Springer Singapore, 2019, pp. 199–203.

[137] A. Kurz, "Physiology of thermoregulation," *Best Practice and Research Clinical Anaesthesiology*, vol. 22, no. 4, pp. 627–644, 2008, thermoregulation in Anesthesia and Intensive Care Medicine. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1521689608000554

[138] 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*, vol. 3, pp. 18–31, 2012.

[139] B. A. MacRae, S. Annaheim, C. M. Spengler, and R. M. Rossi, "Skin temper-

ature measurement using contact thermometry: a systematic review of setup variables and their effects on measured values," *Frontiers in physiology*, vol. 9, p. 29, 2018.

[140] G. Regalia, F. Onorati, M. Lai, C. Caborni, and R. W. Picard, "Multimodal wrist-worn devices for seizure detection and advancing research: focus on the empatica wristbands," *Epilepsy research*, vol. 153, pp. 79–82, 2019.

[141] P. Ekman, R. W. Levenson, and W. V. Friesen, "Autonomic nervous system activity distinguishes among emotions," *science*, vol. 221, no. 4616, pp. 1208–1210, 1983.

[142] R. W. Levenson, P. Ekman, and W. V. Friesen, "Voluntary facial action generates emotion-specific autonomic nervous system activity," *Psychophysiology*, vol. 27, no. 4, pp. 363–384, 1990.

[143] C. Collet, E. Vernet-Maury, G. Delhomme, and A. Dittmar, "Autonomic nervous system response patterns specificity to basic emotions," *Journal of the autonomic nervous system*, vol. 62, no. 1-2, pp. 45–57, 1997.

[144] J. Moltó, P. Segarra, R. López, À. Esteller, A. Fonfría, M. C. Pastor, and R. Poy, "Adaptación eapañola del" international affective picture system"(iaps). tercera parte." *Anales de Psicología/Annals of Psychology*, vol. 29, no. 3, pp. 965–984, 2013.

[145] M. Khomami 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*, vol. PP, p. 1, 01 2015.

[146] S. Vadrevu and M. S. Manikandan, "Real-time ppg signal quality assessment system for improving battery life and false alarms," *IEEE transactions on circuits and systems II: express briefs*, vol. 66, no. 11, pp. 1910–1914, 2019.

[147] D. Biswas, N. Simões-Capela, C. Van Hoof, and N. Van Helleputte, "Heart rate estimation from wrist-worn photoplethysmography: A review," *IEEE Sensors Journal*, vol. 19, no. 16, pp. 6560–6570, 2019.

[148] M. Soleymani, F. Villaro-Dixon, T. Pun, and G. Chanel, "Toolbox for emotional feature extraction from physiological signals (teap)," *Frontiers in ICT*, vol. 4, p. 1, 2017.

[149] P. van Gent, H. Farah, N. van Nes, and B. van Arem, "Heartpy: A novel heart rate algorithm for the analysis of noisy signals," *Transportation research part F: traffic psychology and behaviour*, vol. 66, pp. 368–378, 2019.

[150] S. A. H. Aqajari, E. K. Naeini, M. A. Mehrabadi, S. Labbaf, N. Dutt, and A. M. Rahmani, "pyeda: An open-source python toolkit for pre-processing and feature extraction of electrodermal activity," *Procedia Computer Science*, vol. 184, pp. 99–106, 2021.

[151] R. Parasuraman and Y. Jiang, "Individual differences in cognition, affect, and performance: behavioral, neuroimaging, and molecular genetic approaches," *NeuroImage*, vol. 59, no. 1, pp. 70–82, Jan 2012, 21569853[pmid].

[152] P. Verduyn, E. Delvaux, H. Coillie, F. Tuerlinckx, and I. Mechelen, "Predicting the duration of emotional experience: Two experience sampling studies," *Emotion*, vol. 9, pp. 83–91, 03 2009.

[153] M. M. Hassan, M. G. R. Alam, M. Z. Uddin, S. Huda, A. Almogren, and G. Fortino, "Human emotion recognition using deep belief network architecture," *Information Fusion*, vol. 51, pp. 10 – 18, 2019.

[154] 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*, vol. 18, p. 2074, 06 2018.

[155] J. Zhang, Z. Yin, P. Chen, and S. Nichele, "Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review," *Information Fusion*, vol. 59, pp. 103–126, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253519302532

[156] H. P. Martinez, Y. Bengio, and G. N. Yannakakis, "Learning deep physiological models of affect," *IEEE Computational Intelligence Magazine*, vol. 8, no. 2, pp. 20–33, 2013.

[157] 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 of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[158] A. Bakhshi and S. Chalup, "Multimodal emotion recognition based on speech and physiological signals using deep neural networks," in *Pattern Recognition. ICPR International Workshops and Challenges*, A. Del Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, and R. Vezzani, Eds. Cham: Springer International Publishing, 2021, pp. 289–300.

[159] J. Miranda, M. Canabal, L. Gutiérrez-Martín, J. M. Lanza-Gutiérrez, and C. López-Ongil, "A design space exploration for heart rate variability in a wearable smart device," in *2020 XXXV Conference on Design of Circuits and Integrated Systems (DCIS)*, 2020, pp. 1–6.

[160] J. Rubin, R. Abreu, S. Ahern, H. Eldardiry, and D. G. Bobrow, "Time, frequency & complexity analysis for recognizing panic states from physiologic time-series," in *Proceedings of the 10th EAI International Conference on Pervasive Computing Technologies for Healthcare*, ser. PervasiveHealth '16. Brussels, BEL: ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2016, p. 81–88.

[161] G. Valenza, A. Lanata, and E. P. Scilingo, "The role of nonlinear dynamics in affective valence and arousal recognition," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 237–249, 2012.

[162] E. Kanjo, E. M. Younis, and C. S. Ang, "Deep learning analysis of mobile physiological, environmental and location sensor data for emotion detection," *Information Fusion*, vol. 49, pp. 46–56, 2019.

[163] 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*, vol. 7, pp. 57–67, 2018.

[164] F. Li, L. Yang, H. Shi, and C. Liu, "Differences in photoplethysmography morphological features and feature time series between two opposite emotions: Happiness and sadness," *Artery Research*, vol. 18, pp. 7–13, 2017.

[165] S. Kotsiantis, "Feature selection for machine learning classification problems: a recent overview," *Artificial Intelligence Review*, vol. 42, no. 1, pp. 157–176, 2011.

[166] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, 2018.

[167] G. Valenza, A. Lanata, and E. P. Scilingo, "The role of nonlinear dynamics

in affective valence and arousal recognition," *IEEE transactions on affective computing*, vol. 3, no. 2, pp. 237–249, 2011.

[168] S. Vijayakumar, R. Flynn, and N. Murray, "A comparative study of machine learning techniques for emotion recognition from peripheral physiological signals," in *2020 31st Irish Signals and Systems Conference (ISSC)*. IEEE, 2020, pp. 1–6.

[169] L. Van Der Maaten, E. Postma, J. Van den Herik *et al.*, "Dimensionality reduction: a comparative," *J Mach Learn Res*, vol. 10, no. 66-71, p. 13, 2009.

[170] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Proceedings of the 24th International Conference on Neural Information Processing Systems*, ser. NIPS'11. Red Hook, NY, USA: Curran Associates Inc., 2011, p. 2546–2554.

[171] T. Yu and H. Zhu, "Hyper-parameter optimization: A review of algorithms and applications," *arXiv preprint arXiv:2003.05689*, 2020.

[172] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization." *Journal of machine learning research*, vol. 13, no. 2, 2012.

[173] R. Turner, D. Eriksson, M. McCourt, J. Kiili, E. Laaksonen, Z. Xu, and I. Guyon, "Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020," *arXiv preprint arXiv:2104.10201*, 2021.

[174] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[175] V. K. Chauhan, K. Dahiya, and A. Sharma, "Problem formulations and solvers in linear svm: a review," *Artificial Intelligence Review*, vol. 52, no. 2, pp. 803–855, 2019.

[176] P. C. Sen, M. Hajra, and M. Ghosh, "Supervised classification algorithms in machine learning: A survey and review," in *Emerging technology in modelling and graphics*. Springer, 2020, pp. 99–111.

[177] W. Wang and D. Sun, "The improved adaboost algorithms for imbalanced data classification," *Information Sciences*, vol. 563, pp. 358–374, 2021.

[178] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: analysis of affective physiological state," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

[179] 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*, vol. 9, no. 2, pp. 147–160, 2018.

[180] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, "Introducing wesad, a multimodal dataset for wearable stress and affect detection," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ser. ICMI '18. New York, NY, USA: Association for Computing Machinery, 2018, p. 400–408.

[181] J. A. Miranda, M. F. Canabal, J. M. Lanza-Gutiérrez, M. Portela-García, C. López-Ongil, and T. R. Alcaide, "Meaningful data treatment from multiple physiological sensors in a cyber-physical system," in *DCIS 2017: XXXII Conference on Design of Circuits and Integrated Systems*, 2017, pp. 100–104.

[182] J. A. Miranda, M. F. Canabal, J. M. Lanza-Gutiérrez, M. P. García, and

C. López-Ongil, "Toward fear detection using affect recognition," in *2019 XXXIV Conference on Design of Circuits and Integrated Systems (DCIS)*, 2019, pp. 1–4.

[183] T. Christy, L. I. Kuncheva, and K. W. Williams, "Selection of physiological input modalities for emotion recognition," *UK: Bangor University*, 2012.

[184] J. A. Miranda Calero, R. Marino, J. M. Lanza-Gutierrez, T. Riesgo, M. Garcia-Valderas, and C. Lopez-Ongil, "Embedded emotion recognition within cyber-physical systems using physiological signals," in *2018 Conference on Design of Circuits and Integrated Systems (DCIS)*, 2018, pp. 1–6.

[185] O. Bălan, G. Moise, A. Moldoveanu, M. Leordeanu, and F. Moldoveanu, "Fear level classification based on emotional dimensions and machine learning techniques," *Sensors*, vol. 19, no. 7, 2019.

[186] L. Petrescu, C. Petrescu, A. Oprea, O. Mitruț, G. Moise, A. Moldoveanu, and F. Moldoveanu, "Machine learning methods for fear classification based on physiological features," *Sensors*, vol. 21, no. 13, p. 4519, 2021.

[187] J. A. Miranda, M. F. Canabal, L. Gutiérrez-Martín, J. M. Lanza-Gutierrez, M. Portela-García, and C. López-Ongil, "Fear recognition for women using a reduced set of physiological signals," *Sensors*, vol. 21, no. 5, p. 1587, 2021.

[188] L. Sörnmo and P. Laguna, "Chapter 7 - ecg signal processing," in *Bioelectrical Signal Processing in Cardiac and Neurological Applications*, ser. Biomedical Engineering. Burlington: Academic Press, 2005, pp. 453 – 566. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780124375529500076

[189] S. Mittal, "A survey of techniques for approximate computing," *ACM Computing Surveys (CSUR)*, vol. 48, no. 4, p. 62, 2016.

[190] M. T. Valderas, J. Bolea, P. Laguna, R. Bailón, M. Vallverdú *et al.*, "Mutual information between heart rate variability and respiration for emotion characterization," *Physiological measurement*, vol. 40, no. 8, p. 084001, 2019.

[191] M. Costa, A. L. Goldberger, and C.-K. Peng, "Multiscale entropy analysis of complex physiologic time series," *Physical review letters*, vol. 89, no. 6, p. 068102, 2002.

[192] J. Kim and E. Andre, "Emotion-specific dichotomous classification and feature-level fusion of multichannel biosignals for automatic emotion recognition," in *2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, 2008, pp. 114–119.

[193] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[194] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[195] C. Godin, F. Prost-Boucle, A. Campagne, S. Charbonnier, S. Bonnet, and A. Vidal, "Selection of the most relevant physiological features for classifying emotion," *Emotion*, vol. 40, p. 20, 2015.

[196] A. Albraikan, D. P. Tobón, and A. El Saddik, "Toward user-independent emotion recognition using physiological signals," *IEEE Sensors Journal*, vol. 19, no. 19, pp. 8402–8412, 2018.

[197] J. Rottenberg, R. Ray, and J. Gross, "Emotion elicitation using films in: Coan ja, allen jjb, editors. the handbook of emotion elicitation and assessment," 2007.

[198] W.-H. Lin, D. Wu, C. Li, H. Zhang, and Y.-T. Zhang, "Comparison of heart rate variability from ppg with that from ecg," in *The international conference on health informatics*. Springer, 2014, pp. 213–215.

[199] G. Lu, F. Yang, J. Taylor, and J. Stein, "A comparison of photoplethysmography and ecg recording to analyse heart rate variability in healthy subjects," *Journal of medical engineering & technology*, vol. 33, no. 8, pp. 634–641, 2009.

[200] M. Bolanos, H. Nazeran, and E. Haltiwanger, "Comparison of heart rate variability signal features derived from electrocardiography and photoplethysmography in healthy individuals," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, 2006, pp. 4289–4294.

[201] M. Soleymani, J. Davis, and T. Pun, "A collaborative personalized affective video retrieval system," in *2009 3rd international conference on affective computing and intelligent interaction and workshops*. IEEE, 2009, pp. 1–2.

[202] M. E. Dawson, A. M. Schell, and D. L. Filion, *The Electrodermal System*, 4th ed., ser. Cambridge Handbooks in Psychology. Cambridge University Press, 2016, p. 217–243.

[203] V. Shusterman, K. P. Anderson, and O. Barnea, "Spontaneous skin temperature oscillations in normal human subjects," *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, vol. 273, no. 3, pp. R1173–R1181, 1997.

[204] J. Pan and W. J. Tompkins, "A real-time qrs detection algorithm," *IEEE Transactions on Biomedical Engineering*, vol. BME-32, no. 3, pp. 230–236, 1985.

[205] C.-K. Peng, S. Havlin, H. E. Stanley, and A. L. Goldberger, "Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series," *Chaos: an interdisciplinary journal of nonlinear science*, vol. 5, no. 1, pp. 82–87, 1995.

[206] C. M. van-den Bleek and J. C. Schouten, "Deterministic chaos: a new tool in fluidized bed design and operation," *The Chemical Engineering Journal and the Biochemical Engineering Journal*, vol. 53, no. 1, pp. 75 – 87, 1993.

[207] M. M. Carl Rhodes, "The false nearest neighbors algorithm: An overview," *Computers & Chemical Engineering*, vol. 21, pp. S1149 – S1154, 1997, supplement to Computers and Chemical Engineering.

[208] S. Schinkel, O. Dimigen, and N. Marwan, "Selection of recurrence threshold for signal detection," *The European Physical Journal Special Topics*, vol. 164, pp. 15–53, 10 2008.

[209] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: Recommendations for practitioners," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 3, p. 252–264, mar 1991. [Online]. Available: https://doi.org/10.1109/34.75512

[210] F. Nasoz, C. L. Lisetti, K. Alvarez, and N. Finkelstein, "Emotion recognition from physiological signals for user modeling of affect," in *Proceedings of the 3rd Workshop on Affective and Attitude User Modelling (Pittsburgh, PA, USA*, 2003.

[211] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun, "Emotion assessment

from physiological signals for adaptation of game difficulty," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 41, no. 6, pp. 1052–1063, 2011.

[212] P. Rathod, K. George, and N. Shinde, "Bio-signal based emotion detection device," in *2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE, 2016, pp. 105–108.

[213] B. Zhao, Z. Wang, Z. Yu, and B. Guo, "Emotionsense: Emotion recognition based on wearable wristband," in *2018 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computing, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, 2018, pp. 346–355.

[214] J. A. Miranda, A. P. Montoro, C. López-Ongil, and J. Andreu-Pérez, "Ft2f-sqa: Few-shot type-2 fuzzy-based subject-invariant ppg quality assessment for extreme edge physiological monitoring," *IEEE TIM*, 2022, manuscript submitted for publication.

[215] GSMA, "Connected women – the mobile gender gap report 2021," https://www.gsma.com/r/wp-content/uploads/2021/07/The-Mobile-Gender-Gap-Report-2021.pdf, June 2021, (Accessed on 02/15/2022).

[216] K. Eisenhut, E. Sauerborn, C. García-Moreno, and V. Wild, "Mobile applications addressing violence against women: a systematic review," *BMJ global health*, vol. 5, no. 4, p. e001954, 2020.

[217] A. J. Yugueros García, "Violencia de género, seguridad de las víctimas desde la perspectiva psicosocial," *REVISTA DE GÉNERO E IGUALDAD*, 2021.

[218] N. Karusala and N. Kumar, "Women's safety in public spaces: Examining the efficacy of panic buttons in new delhi," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17. New York, NY, USA: Association for Computing Machinery, 2017, p. 3340–3351.

[219] M. N. Islam, N. T. Promi, J. M. Shaila, M. A. Toma, M. A. Pushpo, F. B. Alam, S. N. Khaledur, T. T. Anannya, and M. F. Rabbi, "Safeband: A wearable device for the safety of women in bangladesh," in *Proceedings of the 16th International Conference on Advances in Mobile Computing and Multimedia*, 2018, pp. 76–83.

[220] Á. González-Prieto, A. Brú, J. C. Nuño, and J. L. González-Álvarez, "Machine learning for risk assessment in gender-based crime," *arXiv preprint arXiv:2106.11847*, 2021.

[221] J. A. Miranda, M. F. Canabal, M. Portela García, and C. Lopez-Ongil, "Embedded emotion recognition: Autonomous multimodal affective internet of things," in *Proceedings of the cyber-physical systems workshop*, vol. 2208, 2018, pp. 22–29.

[222] E. Rituerto-González, J. A. Miranda, M. F. Canabal, J. M. Lanza-Gutiérrez, C. Peláez-Moreno, and C. López-Ongil, "A hybrid data fusion architecture for bindi: A wearable solution to combat gender-based violence," in *Multimedia Communications, Services and Security*, A. Dziech, W. Mees, and A. Czyżewski, Eds. Cham: Springer International Publishing, 2020, pp. 223–237.

[223] M. Felipe Canabal, "iGlove: Plataforma para el desarrollo de investigación

sobre la detección de emociones," in *Master Thesis*. University Carlos III de Madrid, 2019.

[224] M. T. Quazi, "Human emotion recognition using smart sensors: a thesis submitted in fulfilment of the requirements for the degree of master of engineering in electronics and communication engineering, school of engineering and advanced technology, massey university, palmerston north, new zealand, february 2012," Ph.D. dissertation, Massey University, 2012.

[225] A. Ramirez Barcenas, "Configuración de monitor de batería y análisis de consumo en nodo inalámbrico para prevención de violencia sexual," in *Bachelor Thesis*. University Carlos III de Madrid, 2018.

[226] *nRF52832 Product Specification v1.8*, Nordic Semiconductors, 2021, rev. 1.8. [Online]. Available: https://infocenter.nordicsemi.com/pdf/nRF52832_PS_v1.8.pdf

[227] *Grove - Vibration Motor User Manual*, Seed Studio, 2015, buzzer. [Online]. Available: https://www.mouser.es/datasheet/2/744/Seeed_105020003-1217653.pdf

[228] *Advance Battery Monitor IC*, Texas Instruments, 2003, sLUS465E. [Online]. Available: https://www.ti.com/lit/ds/symlink/bq2019.pdf

[229] *Miniature Single Cell, Fully Integrated Li-Ion, Li-Polymer Charge Management Controller*, MicroChip, 2005, dS21984A. [Online]. Available: https://cdn-shop.adafruit.com/datasheets/MCP73831.pdf

[230] M. van Dooren, J. H. Janssen *et al.*, "Emotional sweating across the body: Comparing 16 different skin conductance measurement locations," *Physiology & behavior*, vol. 106, no. 2, pp. 298–304, 2012.

[231] D. Castaneda, A. Esparza, M. Ghamari, C. Soltanpur, and H. Nazeran, "A review on wearable photoplethysmography sensors and their potential future applications in health care," *International journal of biosensors & bioelectronics*, vol. 4, no. 4, p. 195, 2018.

[232] G. de Cuerva Camacho, "Eliminación de artefactos de movimiento en señales fotopletismográficas para sistemas portables orientados a la detección de emociones," in *Master Thesis*. University Carlos III de Madrid, 2019.

[233] S. Beres and L. Hejjel, "The minimal sampling frequency of the photoplethysmogram for accurate pulse rate variability parameters in healthy volunteers," *Biomedical Signal Processing and Control*, vol. 68, p. 102589, 2021.

[234] A. Paez Montoro, "Optimización de la medida óptica de pulso cardíaco para su integración en el sistema de detección de emociones Bindi," in *Bachelor Thesis*. University Carlos III de Madrid, 2019.

[235] A. Baba and M. Burke, "Measurement of the electrical properties of ungelled ecg electrodes," *International Journal of Biology and Biomedical Engineering*, vol. 2, 11 2007.

[236] M. F. Canabal, J. A. Miranda, J. M. Lanza-Gutiérrez, A. I. Pérez Garcilópez, and C. López-Ongil, "Electrodermal activity smart sensor integration in a wearable affective computing system," in *2020 XXXV Conference on Design of Circuits and Integrated Systems (DCIS)*, 2020, pp. 1–6.

[237] M. F. Canabal, J. A. Miranda, A. P. Montoro, I. P. Garcilópez, S. P. Álvarez, E. G. Ares, and C. López-Ongil, "Design and validation of an efficient and adjustable gsr sensor for emotion monitoring," *IEEE Sensors*, 2022, manuscript in progress.

[238] *Human Body Temperature Sensor*, Maxim Integrated, 2016, sKT. [Online]. Available: https://datasheets.maximintegrated.com/en/ds/MAX30205.pdf

[239] *±0.1°C Accurate, I2C Digital Temperature Sensor*, Maxim Integrated, 2020, sKT. [Online]. Available: https://datasheets.maximintegrated.com/en/ds/MAX30208.pdf

[240] *MAX30208 Evaluation System*, Maxim Integrated, 2019, sKT. [Online]. Available: https://datasheets.maximintegrated.com/en/ds/MAX30208EVSYS.pdf

[241] *S132 Softdevice Specification*, Nordic Semiconductors, 2019, s132. [Online]. Available: https://infocenter.nordicsemi.com/pdf/S132_SDS_v7.1.pdf

[242] A. Ramírez-Bárcenas, M. Portela-García, M. García-Valderas, and C. López-Ongil, "System dependability in edge computing wearable devices," in *2020 XXXV Conference on Design of Circuits and Integrated Systems (DCIS)*, 2020, pp. 1–6.

[243] C. Orphanidou, *Signal quality assessment in physiological monitoring: state of the art and practical considerations*. Springer, 2017.

[244] M. Elgendi, "Optimal signal quality index for photoplethysmogram signals," *Bioengineering*, vol. 3, no. 4, 2016.

[245] R. Krishnan *et al.*, "Two-stage approach for detection and reduction of motion artifacts in photoplethysmographic data," *IEEE Trans. on biomedical engineering*, vol. 57, no. 8, pp. 1867–1876, 2010.

[246] E. K. Naeini *et al.*, "A real-time ppg quality assessment approach for healthcare internet-of-things," *Procedia Computer Science*, vol. 151, pp. 551–558, 2019.

[247] S. Vadrevu *et al.*, "Real-time ppg signal quality assessment system for improving battery life and false alarms," *IEEE TCAS2*, vol. 66, no. 11, pp. 1910–1914, 2019.

[248] G. Narendra Kumar Reddy *et al.*, "On-device integrated ppg quality assessment and sensor disconnection/saturation detection system for iot health monitoring," *IEEE TIM*, vol. 69, no. 9, pp. 6351–6361, 2020.

[249] S. Alam *et al.*, "On-board signal quality assessment guided compression of photoplethysmogram for personal health monitoring," *IEEE TIM*, vol. 70, pp. 1–9, 2021.

[250] Z. Zhao *et al.*, "Sqi quality evaluation mechanism of single-lead ecg signal based on simple heuristic fusion and fuzzy comprehensive evaluation," *Frontiers in Physiology*, vol. 9, p. 727, 2018.

[251] J. M. Mendel, "Uncertain rule-based fuzzy systems," *Introduction and new directions*, p. 684, 2017.

[252] A. Mueen *et al.*, "The fastest similarity search algorithm for time series subsequences under euclidean distance," 2017.

[253] T. Nakamura *et al.*, "Merlin: Parameter-free discovery of arbitrary length anomalies in massive time series archives," in *2020 IEEE International Conference on Data Mining (ICDM)*, 2020, pp. 1190–1195.

[254] Y. Zhu *et al.*, "Matrix profile xi: Scrimp++: Time series motif discovery at interactive speeds," in *2018 IEEE International Conference on Data Mining (ICDM)*, 2018, pp. 837–846.

[255] S. Lloyd, "Least squares quantization in pcm," *IEEE Trans. on information theory*, vol. 28, no. 2, pp. 129–137, 1982.

[256] F. Liu *et al.*, "Encoding words into interval type-2 fuzzy sets using an interval

approach," *IEEE Trans. on Fuzzy Systems*, vol. 16, no. 6, pp. 1503–1521, 2008.

[257] J. Andreu-Perez *et al.*, "Explainable artificial intelligence based analysis for interpreting infant fnirs data in developmental cognitive neuroscience," *Communications biology*, vol. 4, no. 1, pp. 1–13, 2021.

[258] H. Hagras, "Toward human-understandable, explainable ai," *Computer*, vol. 51, no. 9, pp. 28–36, 2018.

[259] M. Antonelli *et al.*, "Multiobjective evolutionary optimization of type-2 fuzzy rule-based systems for financial data classification," *IEEE Trans. on Fuzzy Systems*, vol. 25, no. 2, pp. 249–264, 2017.

[260] H. Hagras *et al.*, "An incremental adaptive life long learning approach for type-2 fuzzy embedded agents in ambient intelligent environments," *IEEE Trans. on Fuzzy Systems*, vol. 15, no. 1, pp. 41–55, 2007.

[261] H. T. Nguyen *et al.*, "Computing degrees of subsethood and similarity for interval-valued fuzzy sets: Fast algorithms," in *9th International Conference on Intelligent Technologies*, 2008.

[262] W. Karlen *et al.*, "Capnobase: Signal database and tools to collect, share and annotate respiratory signals," in *2010 Annual Meeting of the Society for Technology in Anesthesia*, 2010, p. 27.

[263] W. Karlen, J. M. Ansermino, and G. Dumont, "Adaptive pulse segmentation and artifact detection in photoplethysmography for mobile applications," in *2012 Annual Conference of the IEEE EMBS*, 2012, pp. 3131–3134.

[264] W. Karlen, "Csl pulse oximetry artifact labels," 2021.

[265] F. Li, L. Yang, H. Shi, and C. Liu, "Differences in photoplethysmography morphological features and feature time series between two opposite emotions: Happiness and sadness," *Artery Research*, vol. 18, pp. 7–13, 2017. [Online]. Available: https://doi.org/10.1016/j.artres.2017.02.003

[266] Q. Yousef, M. B. I. Reaz, and M. Ali, "The analysis of ppg morphology: Investigating the effects of aging on arterial compliance," *Measurement Science Review*, vol. 12, pp. 266–271, 12 2012.

[267] G. McVeigh, C. Bratteli, D. Morgan, C. Alinder, S. Glasser, S. Finkelstein, and J. Cohn, "Age-related abnormalities in arterial compliance identified by pressure pulse contour analysis: Aging and arterial compliance," *Hypertension*, vol. 33, no. 6, pp. 1392–1398, 1999, cited By 286.

[268] F. Foroozan, M. Mohan, and J. S. Wu, "Robust beat-to-beat detection algorithm for pulse rate variability analysis from wrist photoplethysmography signals," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2136–2140.

[269] M. Soundararajan, S. Arunagiri, and S. Alagala, "An adaptive delineator for photoplethysmography waveforms," *Biomedical Engineering / Biomedizinische Technik*, vol. 61, pp. 645 – 655, 2016.

[270] H. S. Shin, C. Lee, and M. Lee, "Adaptive threshold method for the peak detection of photoplethysmographic waveform," *Computers in Biology and Medicine*, vol. 39, no. 12, pp. 1145–1152, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482509001826

[271] D. Morelli, A. Rossi, M. Cairo, and D. A. Clifton, "Analysis of the impact of interpolation methods of missing rr-intervals caused by motion artifacts on hrv features estimations," *Sensors*, vol. 19, no. 14, p. 3163, 2019.

[272] Z. Zhang *et al.*, "Troika: A general framework for heart rate monitoring using

wrist-type photoplethysmographic signals during intensive physical exercise," *IEEE Trans. on biomedical engineering*, vol. 62, no. 2, pp. 522–531, 2014.

[273] D. Biswas *et al.*, "Heart rate estimation from wrist-worn photoplethysmography: A review," *IEEE Sensors Journal*, vol. 19, no. 16, pp. 6560–6570, 2019.

[274] U. Satija *et al.*, "A review of signal processing techniques for electrocardiogram signal quality assessment," *IEEE Reviews in Biomedical Engineering*, vol. 11, pp. 36–52, 2018.

[275] J. T. VanderPlas, "Understanding the lomb–scargle periodogram," *The Astrophysical Journal Supplement Series*, vol. 236, no. 1, p. 16, may 2018. [Online]. Available: https://doi.org/10.3847/1538-4365/aab766

[276] M. Á. Blanco Ruiz, L. Gutiérrez Martín, J. Á. Miranda Calero, M. F. Canabal Benito, E. Romero Perales, C. Sainz de Baranda Andujar, R. San Segundo Manuel, D. Larrabeiti López, C. Peláez-Moreno, and C. López Ongil. (2021) UC3M4Safety Database - List of Audiovisual Stimuli Annotations. [Online]. Available: https://doi.org/10.21950/CXAAHR

[277] M. Á. Blanco Ruiz *et al.* (2021) UC3M4Safety Database - List of Audiovisual Stimuli (Video). [Online]. Available: https://doi.org/10.21950/LUO1IZ

[278] C. S. de Baranda Andújar, M. B. Ruiz, J. Á. M. Calero, L. G. Martín, M. F. C. Benito, R. San Segundo, and C. L. Ongil, "Perspectiva de género y social en las stem: La construcción de sistemas inteligentes para detección de emociones," *Sociología y tecnociencia: Revista digital de sociología del sistema tecnocientífico*, vol. 11, no. 1, pp. 83–115, 2021.

[279] L. Gutiérrez Martín, "Gender Perspective in Emotion Elicitation under Audiovisual Stimuli Reception: Fear and Panic Emotions are gender related?" in *Master Thesis*. University Carlos III de Madrid, 2020.

[280] L. Velasco Gonzalez, "Diseño e implementación de un sistema de recuperación fisiológica para experimentos de reconocimiento de emociones," in *Bachelor Thesis*. University Carlos III de Madrid, 2021.

[281] T. Kostoulas, G. Chanel, M. Muszynski, P. Lombardo, and T. Pun, "Dynamic time warping of multimodal signals for detecting highlights in movies," in *Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And infLuence*, 2015, pp. 35–40.

[282] S. Gashi, E. Di Lascio, and S. Santini, "Using students' physiological synchrony to quantify the classroom emotional climate," in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, 2018, pp. 698–701.

[283] K. R. Papakannu, "Examining user engagement via facial expressions in augmented reality with dynamic time warping," Ph.D. dissertation, Arizona State University, 2021.

[284] R. Satti, N.-U.-H. Abid, M. Bottaro, M. De Rui, M. Garrido, M. R. Raoufy, S. Montagnese, and A. R. Mani, "The application of the extended poincaré plot in the analysis of physiological variabilities," *Frontiers in Physiology*, vol. 10, 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fphys.2019.00116

[285] N. Bu, "Poincaré analysis based on short-term heart rate variability data for stress evaluation," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 1–4.

[286] P. Shi, S. Hu, and H. Yu, "Recovery of heart rate variability after treadmill exercise analyzed by lagged poincaré plot and spectral characteristics," *Medical & biological engineering & computing*, vol. 56, no. 2, pp. 221–231, 2018.

[287] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in public health*, p. 258, 2017.

[288] J. K. Kim and J. M. Ahn, "New marker for vascular health based on the poincare plot analysis using acceleration plethysmogram," *International Journal of Applied Engineering Research*, vol. 13, no. 21, pp. 15 417–15 423, 2018.

[289] M. Toichi, T. Sugiura, T. Murai, and A. Sengoku, "A new method of assessing cardiac autonomic function and its comparison with spectral analysis and coefficient of variation of r–r interval," *Journal of the autonomic nervous system*, vol. 62, no. 1-2, pp. 79–84, 1997.

[290] J. Jeppesen, S. Beniczky, P. Johansen, P. Sidenius, and A. Fuglsang-Frederiksen, "Using lorenz plot and cardiac sympathetic index of heart rate variability for detecting seizures for patients with epilepsy," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2014, pp. 4563–4566.

[291] A. Aranzana Sánchez, "Implementación de Técnicas de Extracción de Características para el Reconocimiento de Emociones Usando Sensores Fotopletismográficos," in *Bachelor Thesis*. University Carlos III de Madrid, 2020.

[292] W. Allay Bakhtaoui, "Implementación de Técnicas de Extracción de Características para el Reconocimiento de Emociones Usando Sensores para la Conductividad de la Piel," in *Bachelor Thesis*. University Carlos III de Madrid, 2020.

[293] F. Adrián Hernández Gant, "Diseño de modelos de aprendizaje para detección de miedo en Bindi," in *Master Thesis*. University Carlos III de Madrid, 2021.

[294] H. S. Shin, C. Lee, and M. Lee, "Adaptive threshold method for the peak detection of photoplethysmographic waveform," *Computers in biology and medicine*, vol. 39, no. 12, pp. 1145–1152, 2009.

[295] S. Kuntamalla and L. R. G. Reddy, "An efficient and automatic systolic peak detection algorithm for photoplethysmographic signals," *International Journal of Computer Applications*, vol. 97, no. 19, 2014.

[296] J. T. VanderPlas, "Understanding the lomb–scargle periodogram," *The Astrophysical Journal Supplement Series*, vol. 236, no. 1, p. 16, 2018.

[297] O. De Wel, M. Lavanga, A. C. Dorado, K. Jansen, A. Dereymaeker, G. Naulaers, and S. Van Huffel, "Complexity analysis of neonatal eeg using multiscale entropy: applications in brain maturation and sleep stage classification," *Entropy*, vol. 19, no. 10, p. 516, 2017.

[298] S. Arunachalam, S. Kapa, S. Mulpuru, P. Friedman, and E. Tolkacheva, "Improved multiscale entropy technique with nearest-neighbor moving-average kernel for nonlinear and nonstationary short-time biomedical signal analysis," *Journal of healthcare engineering*, vol. 2018, 2018.

[299] C. Guanghui and Z. Xiaoping, "Multi-modal emotion recognition by fusing correlation features of speech-visual," *IEEE Signal Processing Letters*, vol. 28, pp. 533–537, 2021.

[300] H. Pérez-Espinosa, R. Zatarain-Cabada, and M. L. Barrón-Estrada, "Emotion recognition: from speech and facial expressions," in *Biosignal Processing and Classification Using Computational Learning and Intelligence*. Elsevier, 2022,

pp. 307–326.

[301] Y. Huang, J. Yang, S. Liu, and J. Pan, "Combining facial expressions and electroencephalography to enhance emotion recognition," *Future Internet*, vol. 11, no. 5, p. 105, 2019.

[302] A. Muaremi, B. Arnrich, and G. Tröster, "Towards measuring stress with smartphones and wearable devices during workday and sleep," *BioNanoScience*, vol. 3, no. 2, pp. 172–183, 2013.

[303] A. Exler, A. Schankin, C. Klebsattel, and M. Beigl, "A wearable system for mood assessment considering smartphone features and data from mobile ecgs," in *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: Adjunct*, 2016, pp. 1153–1161.

[304] J. Kim and E. André, "Emotion recognition using physiological and speech signal in short-term observation," in *International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems*. Springer, 2006, pp. 53–64.

[305] Y. Huang, J. Yang, S. Liu, and J. Pan, "Combining facial expressions and electroencephalography to enhance emotion recognition," *Future Internet*, vol. 11, no. 5, p. 105, 2019.

[306] Y. Cimtay, E. Ekmekcioglu, and S. Caglar-Ozhan, "Cross-subject multimodal emotion recognition based on hybrid fusion," *IEEE Access*, vol. 8, pp. 168 865–168 878, 2020.